Research Article

Arena: Rapid and Accurate Reconstruction of Full Atomic RNA Structures From Coarse-grained Models

Zion R. Perry [1], Anna Marie Pyle [2,3,4,*] and Chengxin Zhang [2,3,5,*]

*1 - **Department of Molecular Biophysics and Biochemistry,** Yale University, New Haven, CT 06511, USA*
*2 - **Department of Molecular,** Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA*
*3 - **Howard Hughes Medical Institute,** Chevy Chase, MD 20815, USA*
*4 - **Department of Chemistry,** Yale University, New Haven, CT 06511, USA*
*5 - **Department of Computational Medicine and Bioinformatics,** University of Michigan, Ann Arbor, MI 48109, USA*

***Correspondence to Anna Marie Pyle and Chengxin Zhang:** anna.pyle@yale.edu (A.M. Pyle), zcx@umich.edu (C. Zhang) @zionrperry (Z.R. Perry), @pylelab (A.M. Pyle), @ChengxinZhang3 (C. Zhang)*
https://doi.org/10.1016/j.jmb.2023.168210
Edited by David Mathews

Abstract

RNA tertiary structures from experiments or computational predictions often contain missing atoms, which prevent analyses requiring full atomic structures. Current programs for RNA reconstruction can be slow, inaccurate, and/or require specific atoms to be present in the input. We present Arena (Atomic Reconstruction of RNA), which reconstructs a full atomic RNA structure from residues that can have as few as one atom. Arena first fills in missing atoms and then iteratively refines their placement to reduce non-ideal geometries. We benchmarked Arena on a dataset of 361 RNA structures, where Arena achieves high accuracy and speed compared to other structure reconstruction programs. For example, Arena was used to reconstruct full atomic structures from a single phosphorus atom per nucleotide to, on average, within 3.63 Å RMSD of the experimental structure, while virtually removing all clashes and running in <3 s, which is 353× and 46× faster than state-of-the-art programs PDBFixer and C2A, respectively. The Arena source code is available at https://github.com/pylelab/Arena and the webserver at https://zhang-group.org/Arena/.

© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

Elaborate three-dimensional (3D) structures enable RNA to participate in diverse cellular processes. Developments in high-resolution cryo-electron microscopy (cryo-EM) for RNA 3D structure determination make this an exciting time for studying RNA structure-based mechanisms.[1] Full atomic structures provide the most insight into structure–function relationships of RNA molecules, such as better understanding RNA folding and catalysis,[2] designing small molecule drugs,[3] and performing more accurate computational simulations.[4]

Experimental and computational studies, however, often generate partially or fully coarse-grained models, which lack full atomic information. In RNA crystallography and cryo-EM, often the phosphate and base atoms are the most well-resolved, but much of the structure lacks sufficient resolution to determine the positions of all the atoms.[5] Computational programs for RNA 3D structure prediction, such as NAST[6] and SimRNA,[7] improve simulation speed using coarse-grained

0022-2836/© 2023 The Authors. Published by Elsevier Ltd.This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

representations, which require other post-processing methods to reconstruct a full atomic model. The challenge of full atomic reconstruction from coarse-grained models has largely been solved for proteins with PULCHRA,[8] which reconstructs full atomic structures quickly and accurately. The programs available for RNA full atomic reconstruction have begun to address this challenge, but improvements in accuracy, speed, and ease of use are needed.

State-of-the-art programs that most directly solve this challenge for RNA are PDBFixer[9] and C2A.[10] PDBFixer adds missing atoms to a target structure using a standard template PDB for each residue and superimposes the template atoms onto the target atoms. PDBFixer then runs a set of short Langevin dynamics simulations to reduce clashes. C2A requires a coarse-grained target, a definition of RNA secondary structure segments, and a reference full atomic 3D structure (the default is PDB 1N32) from which full atomic structure fragments are derived. C2A chooses fragments in the reference structure that match the target on a coarse-grained level using a Monte Carlo simulation to reduce collisions and then builds the reference fragment's atomic detail into the target.

Two other programs reconstruct a subset of atoms. Rosetta rna_thread,[11] originally created for homology modeling, can reconstruct side chains given the sugar-phosphate backbone. RCrane,[5] originally created for RNA crystallography refinement, reconstructs the sugar-phosphate backbone given the base, phosphorus (P), and C1′ atoms. RCrane performs backbone reconstruction by matching a library of empirically determined backbone conformers[12] with the backbone pseudo-torsion angles derived from the base, P, and C1′ atom positions. The sugar conformation is further refined by downhill simplex minimization. We benchmarked RCrane.CLI, which is a fully automated version of RCrane that builds the structure from the given atom positions without the need to locate them in an electron density map.

Here, we introduce Arena (Atomic Reconstruction of RNA), which is highly accurate, fast, and easy to use. Notably, Arena generates a full atomic structure for any coarse-grained RNA which has at least one atom present per nucleotide, regardless of its identity. Arena first superimposes atoms from an ideal RNA helix[13] onto the target and iteratively refines the atom positions by correcting bond lengths, bond angles, base and base pair conformations, and steric clashes. Arena, written in C++, is fully automated, and does not require the user to install third-party packages or libraries nor provide additional inputs, such as sequences, structural templates, or experimental data. The program is freely available at https://github.com/pylelab/Arena and as a webserver at https://zhanggroup.org/Arena/.

Results

Overview of Arena

Arena (Atomic Reconstruction of RNA) takes an RNA structure with missing atoms and produces a full atomic, refined structure (Figure 1). As input, Arena takes an RNA in PDB format, which can be of any length and must contain at least one atom per nucleotide of any identity (Figure 1(A)). Arena first fills in all the missing atoms for each nucleotide by superimposing fragments from an ideal A-form RNA helix[13] onto the input structure (Figure 1(B)). Meanwhile, the secondary structure is assigned with CSSR[14] prior to refinement (Figure 1(C)). The structure then undergoes multiple steps of refinement to correct nonideal bond lengths and angles (Figure 1(D)), correct base and base pair conformations (Figure 1(E)), and remove clashing atom pairs (Figure 1(F)). Arena iterates through the refinement steps according to the length of the RNA (between 100 and 1,000 iterations) and updates the secondary structure assignment every 10th iteration (Figure 1(C)). This results in a refined, full atomic structure in <3.5 s on average (Figure 1(G)).

Benchmarking datasets

Arena and other programs were benchmarked on a dataset of 361 non-redundant RNA structures curated from the PDB according to previous studies.[14,15] Briefly, the RNAs, which include rRNAs, tRNAs, snRNAs, introns, riboswitches, etc., are single chains between 30 and 692 nucleotides, have $\geq$10 canonical base pairs, have resolution <4 Å, and share <80% sequence identity.

From this PDB dataset, we generated 5 datasets to evaluate the ability of Arena and other programs to rebuild RNAs from different starting structures and/or due to the input requirements of other programs. Each dataset contains a different subset of atoms, including 1) the backbone atoms (P, OP1, OP2, O5′, C5′, C4′, O4′, C3′, O3′, C2′, O2′, C1′), 2) the P, C1′, and side chain atoms (N9, C8, N7, C5, C6, N6, N1, C4, C2, O2, N3, N4, O6, N2, O4), 3) the P atoms, 4) the C3′ atoms, and 5) the N atoms involved in the N-glycosidic bond (N1 for nucleotides C and U; N9 for nucleotides A and G). Additionally, we generated 2 computationally predicted coarse-grained datasets from NAST (C3′ atoms) and SimRNA (P, C4′, N9, C2 and C6 for nucleotides G and A; P, C4′ N1, C2 and C4 for nucleotides C1 and U) (Text S1). All PDB files and source code for benchmarking are available at https://doi.org/10.5281/zenodo.7897374.

We evaluated each program's performance with four metrics, the heavy atom root-mean-square-deviation (RMSD) from the experimental structure, interaction network fidelity (INF), number of atom pairs that clash with each other, and

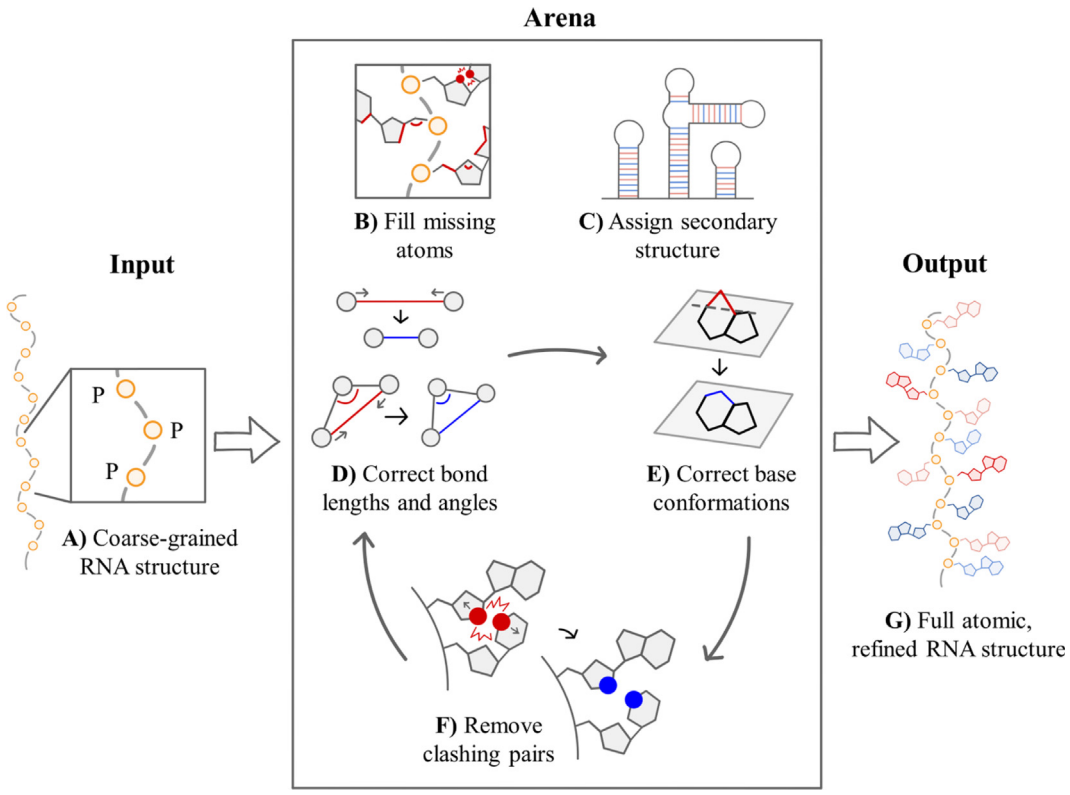

Figure 1. Computational workflow of Arena. Starting from a coarse-grained RNA structure, shown here containing only phosphorus (P) atoms **(A)**, Arena fills in any missing non-hydrogen atoms **(B)** and assigns secondary structure **(C)**. Arena then iterates through several stages of refinement to correct bond lengths and bond angles **(D)**, correct base and base pair conformations **(E)**, and remove clashing atom pairs **(F)** to generate a full atomic and refined RNA structure **(G)**.

computational runtime. INF quantifies the consistency of base–base interactions between two structures, where an INF score of 1 indicates perfect agreement (Text S2).[16] While Arena, PDBFixer, rna_thread, and RCrane keep the coordinates of the original input atoms unchanged in the output structure, C2A may shift the coordinates of the input atoms. Since this coordinate shift is caused by superimposition of local fragments onto the input structure, the C2A output is in the same frame as the input structure. Therefore, there is no need to perform global superimposition of the output structure onto the native structure before the RMSD calculation. To obtain consistent runtime statistics, all benchmarks were performed single-threaded on Intel Xeon E5-2660 v3 CPUs.

Optimization of Arena

We first optimized the number of iterations that Arena performs to correct bond lengths and angles, base and base pair conformations, and clashes (Figure 1(D-F)). We tested between 1 and 10,000 iterations and used these results to inform a linear function where the number of iterations is proportional to the length of the input RNA, with a minimum of 100 iterations (Figure S1). This function provides a good balance between accuracy (RMSD of 3.63 Å, 0.29 clashes) and runtime (2.80 s). Performing 10,000 iterations sharply increases the average runtime (38.09 s), most severely for longer sequences, while the accuracy is not significantly improved (RMSD of 3.63 Å, 0.21 clashes). Thus, between 100 and 1,000 iterations is satisfactory for all input RNAs tested.

Performance of Arena

We evaluated the performance of Arena on the datasets described above. First, we evaluated reconstruction of the side chains from structures containing the backbone atoms (Table 1, Figure 2(A-D)). Arena outperformed the other programs with an RMSD of 0.81 Å, INF of 0.83, 30.66 clashes, and runtime of 3.21 s. Rosetta rna_thread had the next best RMSD (0.98 Å), INF (0.81), and runtime (4.43 s) after Arena but had on average 12× more clashes (374.21 clashes).

Next, we evaluated reconstruction of the backbone from structures containing the P, C1′, and base atoms (Table S1, Figure S2(A-D)). While Arena is 70× faster and contains 77% fewer clashes than the next best program

Table 1 Performance of programs on reconstruction from backbone atoms.

	Arena	PDBFixer	C2A	rna_thread
RMSD (Å)	0.81 ± 0.01*	2.25 ± 0.02 (4.42E-221)	4.70 ± 0.16 (1.21E-82)	0.98 ± 0.01 (1.89E-67)
INF	0.83 ± 0.00*	0.09 ± 0.00 (0)	0.56 ± 0.01 (6.24E-193)	0.81 ± 0.00 (5.70E-21)
Clashes	30.66 ± 2.10*	221.93 ± 10.86 (3.65E-61)	586.81 ± 37.14 (6.34E-42)	374.21 ± 18.98 (1.68E-59)
Runtime (s)	3.21 ± 0.23*	485.11 ± 76.20 (6.81E-10)	222.22 ± 59.76 (2.75E-4)	4.43 ± 0.08 (1.15E-11)
MAE (Å)	0.17 ± 0.00*	0.59 ± 0.01 (1.26E-209)	1.66 ± 0.06 (2.87E-84)	0.19 ± 0.00 (1.70E-28)

Average values are reported with standard error of the mean.
*p-value between Arena and each program was calculated using a paired, two-tailed t-test.

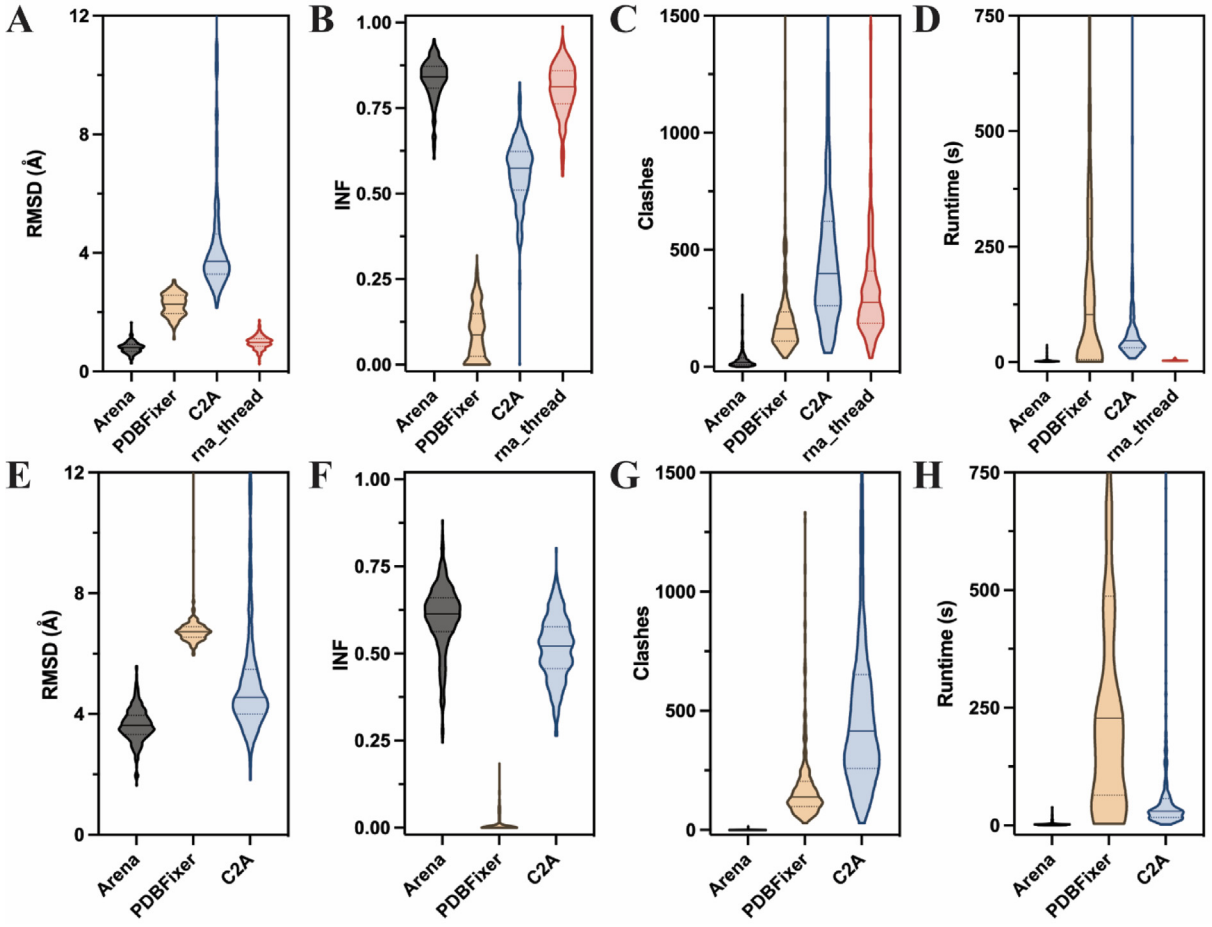

Figure 2. Performance of programs on reconstruction from backbone atoms and P atoms. Arena, PDBFixer, C2A, and rna_thread were used to reconstruct the side chains from input containing the backbone atoms, for which RMSD **(A)**, INF **(B)**, clashes **(C)**, and runtime **(D)** were calculated. Arena, PDBFixer, and C2A were then used to reconstruct the full structure from input only containing the P atoms, for which RMSD **(E)**, INF **(F)**, clashes **(G)**, and runtime **(H)** were calculated. The violin plots show the median (solid line) and quartiles (dotted lines) and were created in GraphPad Prism.

(RCrane), the RMSD of RCrane is on average 0.11 Å lower than that of Arena. Since all the base atoms were defined by the input, the INF score is 1.00 except for C2A, which moves the input atom positions. Whereas Arena does not consider backbone torsion angles, RCrane uses a pseudotorsional system that is aided by a library of empirically determined backbone conformers.[12] Although Arena allows flexibility of the phosphate and ribose groups, the final structure may be biased by the initial backbone conformation used by Arena, which is fixed by the ideal RNA helix template. Conversely, RCrane adjusts the initial conformation based on the coordinates of the P and base atoms, which explains why RCrane has a slightly better RMSD for backbone reconstruction than Arena.

We next tested the more challenging tasks of reconstruction from input models of only one atom per nucleotide. We started with models containing only the P atom, which is the most relevant usage case of Arena for experimental biologists, as ~5.6% of RNA chains in the PDB database only contain P atoms.[14] Arena significantly outperformed PDBFixer and C2A with an RMSD of 3.63 Å, INF of 0.60, 0.28 clashes, and runtime of 2.70 s (Table 2, Figure 2(E-H)). We then evaluated reconstruction from input containing only the C3′ atoms, which is a representation used by the NAST and NASTI programs for RNA structures[17] (Table S2, Figure S2(E-H)). Arena outperformed PDBFixer and C2A with an RMSD of 2.98 Å, INF of 0.69, 0.23 clashes, and runtime of 2.93 s. Lastly, we used structures with only the glycosidic N atoms to evaluate Arena's reconstruction from a single side chain atom and completely undefined backbone. Arena again outperformed PDBFixer with a 60% lower RMSD (2.74 Å), 19× higher INF (0.74), 99.6% fewer clashes (2.24 clashes), and a 275× faster runtime (3.48 s) (Table S3, Figure S2(I-L)).

Next, we tested each program's ability to rebuild structures from computationally predicted coarse-grained structures using NAST and SimRNA (Text S1). Both programs predict the RNA's 3D structure from sequence and generate a coarse-grained structure of C3′ atoms (NAST) or the P, C4′, glycosidic N, C2 and C4 atoms (SimRNA). We then compared Arena's performance to PDBFixer and the corresponding full atomic reconstruction program written for NAST (C2A) or SimRNA. Arena outperformed both PDBFixer and C2A for the NAST-generated inputs (Figure S3(A-D), Table S4) and outperformed or was on par with PDBFixer and SimRNA for the SimRNA-generated inputs (Figure S3(E-G), Table S5). However, the initial model accuracy will affect the final reconstruction. For example, reconstruction from the experimentally determined C3′ positions (Table S2) results in a lower RMSD, higher INF, and lower clashes compared to reconstruction from the computationally determined C3′ positions (Table S4). Moreover, for 11.1% of SimRNA structures, Arena-reconstructed models have at least one backbone discontinuity, which was mainly caused by unusual local conformations of the SimRNA coarse-grained models (See Figure S4 and Table S6 for an example on PDB 6Q9A, chain 4). To address this issue, we benchmarked Arena (extended) on the SimRNA dataset, which performs an additional simulation after the default run where all atoms are allowed to move. This fixes any discontinuous structures and removes 99.9% of clashes for a slight increase in runtime by approximately 0.2 s (Figure S3(E-H), Table S5). We also benchmarked Arena (extended) on the NAST dataset, which removes 92.5% of clashes for an approximately 0.9 s increase in runtime (Figure S3(A-D), Table S4). Overall, for the coarse-grained predictions from NAST and SimRNA, Arena can generate a full atomic structure at least as good as, if not better than, the full atomic reconstruction program developed specifically for each coarse-grained prediction program.

Arena can also be used to rebuild RNA models of NMR structures, and we have provided an additional script for this use case. We tested Arena on two NMR structures (2N3Q and 2KOC) using only the P atoms as input (Table S7) and find that Arena's performance is similar to that on single-model structures (Table 2). Finally, we compared Arena's performance on different classes of RNA, including tRNAs, rRNAs, snRNAs, ribozymes, and riboswitches (Figure S5, Table S8). Compared to the full benchmarking dataset, Arena performs slightly better on ribozymes and riboswitches and slightly worse on rRNAs and snRNAs, while there is no difference in performance on tRNAs (significance threshold of $p \leq 0.05$). While Arena can satisfactorily reconstruct any input RNA, it may perform better on shorter and/or more structurally defined RNAs, such as some ribozymes and riboswitches, which may have more accurate initial atom positions.

We report the Pearson correlation coefficients between performance metrics for the reconstruction from P atoms by Arena (Figure S6). As expected, the runtime increases linearly with the length ($r = 0.94$) (S6D). Although the absolute number of clashes increases with the length ($r = 0.45$) (S6C), larger structures do not necessarily have more clashes when clashes are normalized by the length ($r = 0.14$) (S6E). A lower RMSD generally tracks with a higher INF

Table 2 Performance of programs on reconstruction from P atoms.

	Arena	PDBFixer	C2A
RMSD (Å)	3.63 ± 0.03*	6.77 ± 0.03 (7.10E-228)	5.15 ± 0.11 (4.32E-37)
INF	0.60 ± 0.01*	0.01 ± 0.00 (2.93E-285)	0.52 ± 0.00 (7.21E-49)
Clashes	0.28 ± 0.06*	186.73 ± 9.27 (4.13E-61)	635.04 ± 42.50 (1.22E-39)
Runtime (s)	2.70 ± 0.23*	953.50 ± 145.46 (2.04E-10)	122.93 ± 24.33 (9.98E-7)
MAE (Å)	1.13 ± 0.01*	2.66 ± 0.01 (3.62E-283)	1.85 ± 0.05 (3.70E-49)

Average values are reported with standard error of the mean.
*p-value between Arena and each program was calculated using a paired, two-tailed t-test.

(r = −0.66) (S6F), although the two metrics are independently valuable.

There are several reasons why Arena is much faster than the other programs tested. Firstly, Arena does not perform any stochastic Monte Carlo or molecular dynamics simulations but instead uses an ultra-fast deterministic simulation that performs distance-based calculations. Arena thus has fewer degrees of freedom when searching through conformational space, so it more quickly generates a final structure. More importantly, since all movements in the deterministic simulation are based on ideal RNA geometries, the simulation effectively avoids sampling of suboptimal conformations, thereby producing more accurate conformations than stochastic simulations. Arena also performs a defined number of iterations regardless of performance. By contrast, PDBFixer performs clash removal using the OpenMM package,[9] whose simulation may become locally trapped if clashes cannot be readily removed. Finally, Arena is written in highly compact C++ code, which is much faster than the scripting languages used by the other programs, such as PDBFixer (Python 3), C2A (Python 2.5), and RCrane (Perl).

Case studies on two RNA structures

Notably, Arena is able to reduce or completely remove all clashing atom pairs. We demonstrate this with the reconstruction of an 88-nt 18S rRNA fragment (PDB 5OSG, chain 2) from an input containing only the P atoms (Figure 3, Table 3). Compared to PDBFixer and C2A, Arena achieves a lower-RMSD structure with INF of 0.54 in less than 6% of the runtime (Figure 3(A)). The PDBFixer-generated structure fails to generate any base–base interactions (Figure 3(B,E)), while the C2A-generated structure, although reproducing base–base interactions like Arena (INF of 0.55), has large structural shifts in the backbone compared to the experimental structure (Figure 3(C)). PDBFixer and C2A leave 209 and 592 clashes, respectively, while Arena removes all clashes in the final structure. The number of clashes for each structure is consistent with the average number of clashes from the full-scale benchmarking on P atom structures (averages of 0.28, 186.73, and 635.04 clashes for Arena, PDBFixer, and C2A, respectively). An example region with clashes is shown with residues 2–4 and 11–13 for Arena (Figure 3(D)), PDBFixer

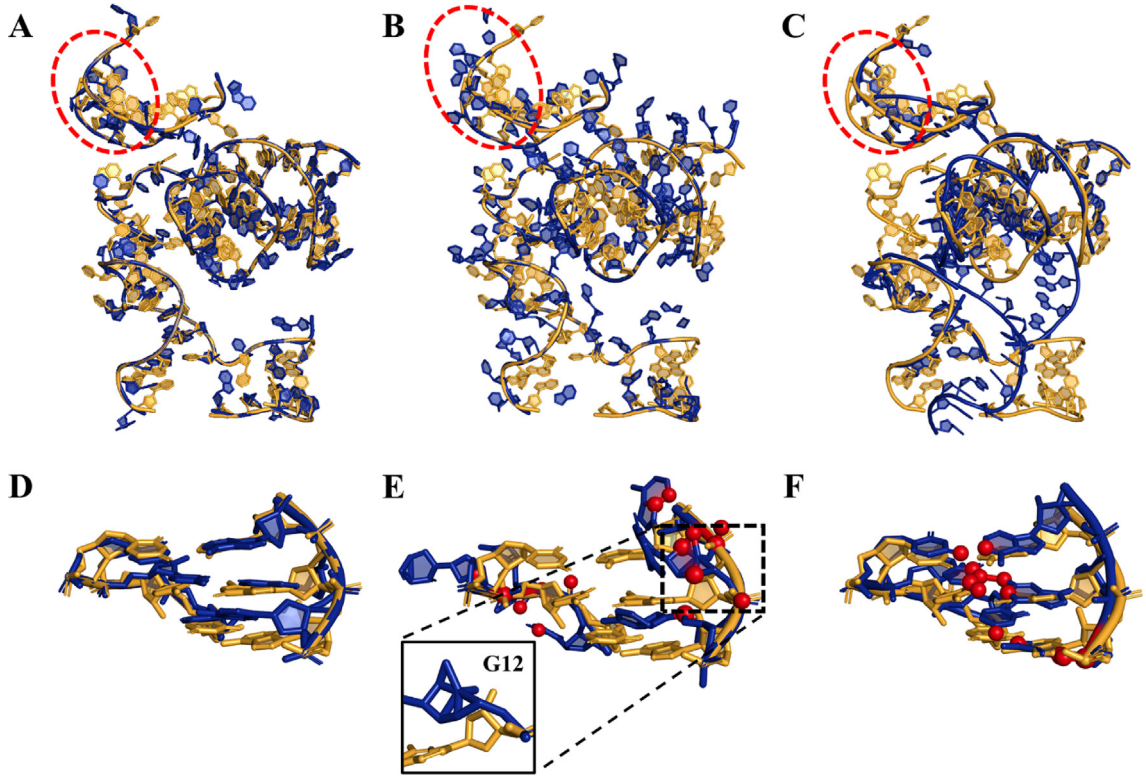

Figure 3. Arena generates a structure close to native with no clashes. The native structure (orange) of the 88-nt 18S rRNA fragment (PDB 5OSG, chain 2) is overlaid with the output structure (blue) generated from an input containing only the P atoms by the programs Arena **(A, D)**, PDBFixer **(B, E)**, and C2A **(C, F)**. The red oval approximately indicates residues shown in **D–F** (res 2–4, 11–13). Atoms that are part of one or more clashing atom pairs are shown as red spheres. The nonbiological ribose ring of residue G12 that was generated by PDBFixer (blue) is shown compared to the native ring (orange) **(E, inset)**. Structures were visualized in PyMOL.

Table 3 Performance of programs on reconstruction of 5OSG from P atoms.

	Arena	PDBFixer	C2A
RMSD (Å)	4.15	7.72	9.52
INF	0.54	–	0.55
Clashes	0	209	592
Runtime (s)	1.64	289.71	28.92
MAE (Å)	1.40	2.82	3.39

(Figure 3(E)), and C2A (Figure 3(F)). We also noticed that PDBFixer sometimes creates nonbiological bonds, such as in the ribose ring of residue G12 (Figure 3(E) inset).

A user may commonly encounter experimental structures that are partially or fully coarse-grained, which accounts for at least 5.6% of RNAs deposited in the PDB.[14] In such cases, users may still need the corresponding full atomic models, which are required for many downstream analyses such as detection of non-canonical base pairs or discovery of atomic contacts between ligands and RNAs. Therefore, we tested Arena's ability to reconstruct a partially coarse-grained structure outside the benchmarking dataset. We chose the cryo-EM structure of the 146-nt 9S rRNA fragment (PDB 6SG9, chain CA) (Figure 4, Table 4). Part of the structure is full atomic, but there are large stretches of residues missing side chains. We used each program to generate a full atomic structure, which we compared to the experimental electron density map. We calculated the average (Table 4) and per-residue (Figure S7) Q-scores between each output and the electron density map, wherea higher Q-score means a greater agreement between the atomic positions and the density map.[18] Arena results in the same Q-score as the original structure (0.34) (Figure 4(A)), compared to 0.33 for PDBFixer (Figure 4(B)), 0.04 for C2A due to large shifts in the backbone (Figure 4(C)), and 0.33 for rna_thread (Figure 4(D)). This shows that Arena successfully adds more atomic details without ruining the original structure.

Discussion

We developed the Arena program to build full atomic, refined structures of any coarse-grained RNA. On a benchmark of 361 experimental RNA structures, we found that Arena is fast (<40 s for all RNAs tested, with average runtimes <3.5 s), achieves low RMSDs (<5.6 Å for all RNAs tested, with average RMSDs < 3.7 Å), has high INF (average 0.60–1), and reduces or completely removes all clashes compared to other programs (removes on average ≥77% of clashing atom pairs). Due to its fast runtime, Arena does not require a high-performance computing cluster and can be run locally. Arena is amenable to a range of RNA sizes, and we tested RNAs between 30 and 692 nt in the benchmarking dataset. Arena can be used to reconstruct low-resolution experimental structures from cryo-EM, X-ray crystallography, and NMR or computationally predicted coarse-grained structures and is applicable for all types of RNAs.

A major reason for the high speed of Arena is that it performs superimposition followed by deterministic refinement based on ideal RNA geometries. This proves to be a more effective and accurate way to rebuild a full atomic structure compared to the other tested programs that perform stochastic simulations. At the same time, Arena has some limitations that could be addressed in future versions of the program. While the current runtime of Arena is sufficiently fast for most applications, the clash removal step is one of the slowest and could be improved by using a lattice coordinate system for the atoms.[19]

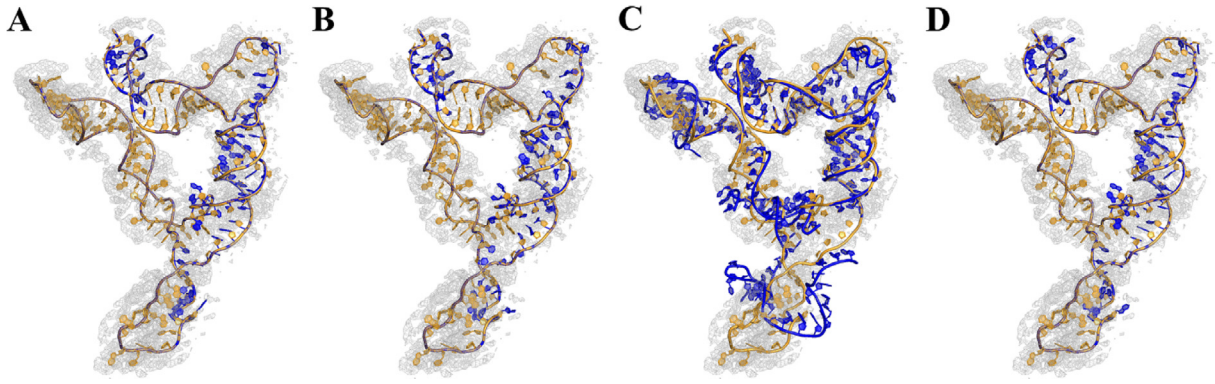

Figure 4. Arena generates a structure from a semi-coarse-grained model that fits its experimental electron density. The native structure (orange) of the 146-nt 9S rRNA fragment (PDB 6SG9, chain CA) is overlaid with the resulting structure (blue) from Arena **(A)**, PDBFixer **(B)**, C2A **(C)**, and rna_thread **(D)**. The outputs were generated from the original semi-coarse-grained structure in the PDB. The electron density map (gray) is shown with contour level 3.0. Structures were visualized in PyMOL.

Table 4 Performance of programs on reconstruction of 6SG9.

	Arena	PDBFixer	C2A	rna_thread	6SG9
Q-score	0.34 ± 0.02 (0.59)	0.33 ± 0.02 (0.03)	0.04 ± 0.01 (8.68E-26)	0.33 ± 0.02 (0.30)	0.34 ± 0.02*
RMSD (Å)	0.00	0.00	10.39	0.00	–
INF	0.76	0.88	0.77	0.29	–
Clashes	63	181	635	352	–
Runtime (s)	2.99	15.65	79.38	9.14	–
MAE (Å)	0.00	0.00	3.79	0.00	–

*p-value between the per-residue Q-scores for the original structure (6SG9) and the output generated by each respective program was calculated using a paired, two-tailed t-test.

To further increase the accuracy of the final structure, Arena could implement a check to fix improper ribose pucker conformations.

Materials and Methods

Arena first generates a full atomic structure from the coarse-grained input and then iteratively refines the placement of those atoms by correcting bond lengths and bond angles, base and base pair conformations, and pairwise atomic clashes. For the default Arena option, only the missing atoms added to the input structure are movable and fixed during the subsequent refinement steps. For the extended Arena option, the default Arena run is first performed; if the structure is discontinuous or atoms are still movable (because the performance metrics do not meet the specified thresholds), Arena performs an identical refinement but with all atoms allowed to move. Additional running options are described at https://github.com/pylelab/Arena#usage.

Step 1: Generate full atomic structure via superimposition. The program first fills in all missing atoms via superimposition of fragments from an ideal RNA helix onto the input atoms. The ideal RNA is from the A-form 3DNA fiber model[20]) for RNA with twist 32.7° and rise 2.548 Å.[13] Fragments chosen from the ideal RNA match the nucleotide identities of the input, such that inconsistent base atoms are automatically excluded (for example, N6 for nucleotide U would be excluded). The fragments are superimposed onto the input target, so that the pre-existing atoms in the target are not moved. The superimposition algorithm uses three-dimensional least squares minimization to find the rotation matrix and translation vector that minimize the RMSD between the target and ideal RNA atoms, which is an extension of the Kabsch algorithm.[21] Arena performs four levels of superimposition: residue-level, base pair-level, base pair stack-level, and local fragment-level. To perform superimposition, there must be at least three pre-existing atoms from the input structure in the level of consideration. Ideally, we want to include the minimum number of residues that together have a total of three atoms. For example, if every residue in the input has three atoms, then only residue-level superimposition is necessary, in which the corresponding 1-nt fragment from the ideal RNA is superimposed onto each residue in the target. For residues with $\leq$2 atoms, the program proceeds to base pair-level filling based on the assigned secondary structure from CSSR.[14] There must be at least three atoms between the bases in the pair for superimposition to occur. If each residue in a pair only has one atom, the program checks a base pair stack, which includes the two residues adjacent to the current residue. If both adjacent residues are base-paired, the program uses a 3-nt helix (base-paired) fragment for the superimposition; if only one is base paired, the program used a 2-nt helix fragment corresponding to the base-paired residues. Finally, if the current residue is not base-paired and has $\leq$2 atoms, the program uses local fragment-level superimposition based on the current residue and its two adjacent residues without considering base pairs. Residues without even a single atom are skipped.

Step 2: Refine bond lengths and bond angles. Arena checks every covalent bond connection (intra-residue and the inter-residue O3′-P bonds) and computes the bond length in 3D space. Arena implements a check to ensure residues are connected (in case of missing residues); sequentially adjacent residue pairs with at least one atom pair within 8 Å are considered connected. Arena calculates the fractional difference between the computed bond length and the ideal bond length taken from the Amber forcefield parameters in QRNAS.[22] If this is greater than the allowable 10% tolerance, then each atom is moved half of the difference between the calculated and ideal bond length. If one of the atoms is not movable because it was present in the input structure, then the movable atom is moved the full distance. Only bonds for which at least one atom is movable are checked. Refining bond angles happens by defining pseudobonds between the previous and subsequent atoms of the current atom. This distance is adjusted identically to bond lengths by comparing to an ideal pseudobond length c calculated as $c = \sqrt{a^2 + b^2 - 2ab * cos(\theta)}$, where a and b are the ideal bond lengths between the previous atom and subsequent atom to the current atom,

respectively, and θ is the ideal bond angle taken from the Amber forcefield parameters in QRNAS.[22] Inter-nucleotide pseudobonds are also considered (C3′-P, O3′-OP1, O3′-OP2, and O3′-O5′). Arena increases or decreases the pseudobond length to indirectly adjust the angle.

Step 3: Refine base and base pair conformations. First, each base's conformation is checked individually by computing the RMSD between the coordinates of the current residue's base atoms (including the C1′ atom) and an ideal 1-nt fragment from the ideal RNA. To compute the RMSD, the ideal base is superimposed onto the target by a least squares fit.[21] If the RMSD $\geq$0.1 Å, then the coordinates of the target are moved to match those from the superimposed ideal base. Second, each Watson-Crick base pair assigned from CSSR[14] is checked (A:U and G:C). The ideal base pair from the ideal RNA helix is superimposed onto the target, and the RMSD is calculated in the same way. If the RMSD $\geq$0.28 Å, the coordinates of the target base pair are moved to match those from the superimposed ideal base pair. The RMSD thresholds were chosen from those used in DSSR.[23] To reduce running time, the CSSR secondary structure assignment, which is required in base pair conformation correction, is updated every tenth iteration.

Step 4: Remove clashing atom pairs. Arena first iterates through every possible residue combination, calculating the 3D distance between the C1′ atoms. If the C1′-C1′ distance is >18.16 Å for the nucleotide pair, it is not physically possible for the constituent atoms to clash (Text S3). If this condition is not met, Arena proceeds to iterate through all inter-residue atom pairs and calculates their 3D distance. The distance threshold that defines a clash is calculated as the sum of the van der Waals radii of the two atoms minus 0.4 Å.[24] If the atom–atom distance is less than the threshold, they are considered to sterically clash, and the atoms are moved apart as described in *Step 2*.

The program iterates through *Steps 2–4* to generate a refined structure. The number of iterations increases with the length of the RNA according to a linear function, where the maximum number of iterations is *0.4*L + 100*, where *L* is the length of the RNA in nucleotides. The program finishes after reaching the maximum number of iterations or until no more atoms are moved, whichever comes first.

CRediT authorship contribution statement

Zion R. Perry: Software, Writing – original draft. **Anna Marie Pyle:** Supervision, Writing – review & editing. **Chengxin Zhang:** Conceptualization, Methodology, Software, Writing – review & editing.

DATA AVAILABILITY

We have shared all links to the program and data in the manuscript.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests that influenced the work reported in this paper. A.M.P. is an associate editor of *Journal of Molecular Biology*. This does not alter the editorial policy or peer review process for this manuscript.

Acknowledgements

This work used the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by the National Science Foundation [2138259, 2138286, 2138307, 2137603, and 2138296]. Z.R.P. is supported by the National Science Foundation Graduate Research Fellowship Program [DGE-2139841]. A.M.P. is supported as an Investigator of the Howard Hughes Medical Institute. Funding for open access charge: Howard Hughes Medical Institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.168210.

Received 9 February 2023;
Accepted 12 July 2023;
Available online 20 July 2023

Keywords:
RNA tertiary structure;
RNA structure prediction;
full atomic reconstruction;
structure refinement;
coarse-grained structure

Abbreviations:
RMSD, root-mean-square-deviation; INF, Interaction Network Fidelity; nt, nucleotide; PDB, Protein Data Bank

References

1. Ma, H., Jia, X., Zhang, K., Su, Z., (2022). Cryo-EM advances in RNA structure determination. *Signal Transduct. Target Ther.* **7**, 58.
2. Pyle, A.M., (2016). Group II Intron self-splicing. *Annu. Rev. Biophys.* **45**, 183–205.
3. Childs-Disney, J.L., Yang, X., Gibaut, Q.M.R., Tong, Y., Batey, R.T., Disney, M.D., (2022). Targeting RNA

structures with small molecules. *Nature Rev. Drug Discov.* **21**, 736–762.
4. Pyle, A.M., Schlick, T., (2016). Challenges in RNA structural modeling and design. *J. Mol. Biol.* **428**, 733–735.
5. Keating, K.S., Pyle, A.M., (2010). Semiautomated model building for RNA crystallography using a directed rotameric approach. *PNAS* **107**, 8177–8182.
6. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., et al., (2009). Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15**, 189–199.
7. Boniecki, M.J., Lach, G., Dawson, W.K., Tomala, K., Lukasz, P., Soltysinski, T., et al., (2016). SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63.
8. Rotkiewicz, P., Skolnick, J., (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465.
9. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., et al., (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659.
10. Jonikas, M.A., Radmer, R.J., Altman, R.B., (2009). Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics* **25**, 3259–3266.
11. Cheng, C.Y., Chou, F.C., Das, R., (2015). Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol.* **553**, 35–64.
12. Richardson, J.S., Schneider, B., Murray, L.W., Kapral, G. J., Immormino, R.M., Headd, J.J., et al., (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14**, 465–481.
13. Chandrasekaran, R., Arnott, S., (1989). The structures of DNA and RNA helices in oriented fibers. Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group VII/1b. *Nucleic Acids*, 31–170.
14. Zhang, C., Pyle, A.M., (2022). CSSR: assignment of secondary structure to coarse-grained RNA tertiary structures. *Acta Crystallogr. D Struct. Biol.* **78**, 466–471.
15. Zhang, C., Zhang, Y., Pyle, A.M., (2022). rMSA: a sequence search and alignment algorithm to improve RNA structure modeling. *J. Mol. Biol.* **167904**
16. Cruz, J.A., Blanchet, M.F., Boniecki, M., Bujnicki, J.M., Chen, S.J., Cao, S., et al., (2012). RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **18**, 610–625.
17. Petersen, N.P., (2020). Computational methods for RNA 3D structure prediction and comparison. Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.
18. Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M.F., Chiu, W., (2020). Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nature Methods* **17**, 328–334.
19. Zhang, Y., Kolinski, A., Skolnick, J., (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145–1164.
20. Li, S., Olson, W.K., Lu, X.-J., (2019). Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.* **47**, W26–W34.
21. Kabsch, W., (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **32**, 922–923.
22. Stasiewicz, J., Mukherjee, S., Nithin, C., Bujnicki, J.M., (2019). QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct. Biol.* **19**, 5.
23. Lu, X.-J., Bussemaker, H.J., Olson, W.K., (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142-e
24. Chen, V.B., Arendall III, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., et al., (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D* **66**, 12–21.