Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RNAflash: Fast and Sensitive Structure-aware RNA Homology Detection and Alignment

Xiaoyong Pan

2008xypan@sjtu.edu.cn

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
https://orcid.org/0000-0001-5010-464X

Bingtong Chen

Shanghai Jiao Tong University

Shengfan Wang

Shanghai Jiao Tong University

Weimin Zhu

Shanghai Jiao Tong University

Yiming Xue

Shanghai Jiao Tong University

Hong-Bin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
https://orcid.org/0000-0002-4029-3325

Article

Keywords: RNA homology search, remote homology detection, RNA language models, multiple sequence alignment, contrastive learning

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11489874/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

RNAflash: Fast and Sensitive Structure-aware RNA Homology Detection and Alignment

Bingtong Chen[1], Shengfan Wang[1], Weimin Zhu[1], Yiming Xue[1], Hong-Bin Shen[1], Xiaoyong Pan[1,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University,
and Key Laboratory of System Control and Information Processing,
Ministry of Education of China, Shanghai, 200240, China

[*]Corresponding author: 2008xypan@sjtu.edu.cn

Abstract

RNA homology search plays a crucial role in RNA structure prediction, facing a trade-off between sensitivity and speed. Here we present RNAflash, a two-stage framework that combines RNA language model-based retrieval with structure-informed reranking and alignment. A retrieval bi-encoder RNAflash-R trained through contrastive learning first retrieves candidate RNA sequences from a large pre-encoded database. Then, its cross-encoder RNAflash-A reranks and aligns the retrieved RNA candidates via jointly trained prediction heads, using predicted secondary structures to refine nucleotide correspondence scores. Across held-out Rfam families, RNAflash retrieved RNA homologs more sensitively than sequence-based methods, and produced alignments comparable to those of structure-based methods. In large-scale searches of RNAcentral, RNAflash generated multiple sequence alignments (MSAs) with stronger coupling and covariation signals than sequence-based and Infernal pipelines, while performing search and MSA construction 3.7-fold faster than non-iterative Riboseek and over 100-fold faster than Infernal and RNAcmap. The lightweight variant RNAflash-L achieves sub-second average search times per query while producing higher-quality MSAs than nhmmer and non-iterative Riboseek. The results show that RNAflash makes sensitive RNA homology search and MSA construction more scalable, while RNAflash-L provides an efficient source of MSAs for downstream RNA structure prediction.

Keywords: RNA homology search, remote homology detection, RNA language models, multiple sequence alignment, contrastive learning

1 Introduction

Identifying homologous RNA sequences is a fundamental task in RNA bioinformatics. The task is a prerequisite for constructing multiple sequence alignments (MSAs), which are widely used in downstream *in silico* analysis, such as structure prediction [1, 2, 3], representation learning [3] and database construction [4]. However, homology search methods generally face a trade-off between sensitivity and computational efficiency. The most computationally efficient algorithms are sequence-based, relying on dynamic programming and scoring schemes to assess the sequence similarity. A representative example is BLASTN [5], which applies k-mer indexing and hash-based search to achieve high speed but can miss homologs with low sequence identity to the query. Nhmmer [6] improves the sensitivity with a hidden Markov model (HMM) constructed from query sequences or query MSAs, but the inference time significantly increases. Riboseek [7] is a recently proposed method designed to balance the sensitivity and speed, but it is still limited by sequence information.

In contrast, structure-aware approaches improve the sensitivity by incorporating RNA structure information, which is generally more conserved than primary sequences. Infernal [8] , a widely used tool for RNA homology search, employs covariance models to capture both sequence and structural conservation, enabling the detection of remote homologs in sequence databases without structural annotations. There are some recently proposed methods such as SHAPEwarp [9] and bpRNA-CosMoS [10], yet their performance has not yet been extensively validated. LocARNA [11], US-align [12] and RTM-align [13] can directly measure the similarity at the secondary or tertiary structure level, but the time complexity and the lack of scalable RNA structure databases constrain their practical use. Another class of methods focuses on combined search strategies, including RNAcmap [14] and rMSA [15], which use the aforementioned tools to progressively search and construct MSAs. While these approaches can achieve high sensitivity, they are not ideal for large-scale search due to computational costs.

The trade-off between sensitivity and speed in current methods motivates the need for alternatives that achieve both goals. Embedding-based approaches, driven by advances in deep learning and large language models, have recently emerged as a promising solution. For protein homology search, methods such as DHR [16] and TM-Vec[17] have demonstrated that embedding-based retrieval can achieve both high sensitivity and efficiency in homolog detection. Despite such successful implementation in protein homology search, embedding-based algorithms remain relatively underexplored in the RNA field. Existing work such as ERAST [18] adapts embedding-based retrieval to nucleotide sequences, but focuses primarily on species taxonomy and is not intended for MSA construction. In addition, converting retrieved candidates into an MSA requires filtering false positives and aligning homologs that may share little sequence identity. This motivates a pairwise model that learns to assess the homology and align nucleotides jointly, using predicted secondary structure to guide the alignment.

To address these gaps, we present RNAflash, a fast structure-aware RNA homology search framework consisting of two components: a retrieval bi-encoder (RNAflash-R) and a cross-encoder for reranking and alignment (RNAflash-A). Specifically, RNAflash-R uses contrastively trained RNA embeddings to retrieve candidates from a pre-indexed database. RNAflash-A then reranks these candidates and uses predicted secondary structure to guide their alignment with the query, assembling the retained alignments into an MSA. We show that RNAflash achieves higher sensitivity than sequence-based methods and substantially improves RNA language models' ability to capture homology relationships in the learned embedding space. It can retrieve structurally conserved sequences with low sequence identity to the query and align them in a structure-aware manner, producing MSAs with stronger evolutionary signals for secondary structure analysis. Compared with conventional algorithms, RNAflash is more scalable, and we additionally provide a lightweight pipeline for fast RNA MSA construction, which was used to support RNA 3D structure

prediction with promising performance.

2 Results

2.1 Overview of RNAflash for fast and sensitive RNA homology search and alignment

RNAflash addresses the speed-sensitivity trade-off by splitting homology search into two stages: fast embedding-based retrieval with RNAflash-R, followed by reranking and alignment with RNAflash-A (Fig. 1a). The first stage is performed by RNAflash-R, a bi-encoder that embeds each sequence independently (Fig. 1b). The encoded query is then matched against a pre-encoded target database (using the same bi-encoder) via cosine similarity. To speed up the RNA homology retrieval, the database embeddings are organized into an inverted-file (IVF) index using Faiss [19], enabling parallel search in sublinear time. For each query, the top k sequences are passed to the next stage.

The second stage is performed by RNAflash-A, a cross-encoder that jointly processes a query–candidate pair to produce a reranking score and an alignment matrix, from which a pairwise alignment is obtained by dynamic programming traceback (Fig. 1c). Candidates are ranked by their predicted homology scores, and those below the threshold are discarded. The retained pairwise alignments are then merged using the query as a common reference to construct an MSA (Fig. 1a). This two-stage design combines efficient embedding-based retrieval with structure-informed alignments, enabling searches against millions of sequences, while restricting the more expensive pairwise computation to the top-ranked candidates.

The two modules are trained separately. We choose Rfam [4] as the training data, because its family assignments provide a credible and biologically meaningful definition of RNA homology. Moreover, Rfam families are accompanied by curated MSAs, making them well suited for evaluating both homology search and alignment. RNAflash-R builds on ERNIE-RNA, a structure-aware language model [20], and is trained on randomly extended sequences with a momentum contrast objective (MoCo) [21], which pulls the representations of homologous sequences together while pushing those of non-homologous sequences apart (Fig. 1b). RNAflash-A uses a secondary-structure-prediction variant of the same backbone (Fig. 1c) and is trained jointly for reranking and alignment. It computes a cross feature from the token-level embeddings of both sequences, and produces the reranking score and the alignment matrix using two independent heads. The predicted secondary structure is injected into the alignment head as a lightweight augmentation so that structurally conserved regions can be aligned even at low sequence identity.

2.2 RNAflash-R outperforms sequence-based approaches in retrieval sensitivity

We first evaluated the retrieval stage in isolation, asking whether the embedding-based retrieval (RNAflash-R) could outperform conventional non-iterative sequence-based methods in homolog detection sensitivity. We compared RNAflash-R with four methods: MMseqs2 [22], BLASTN [5], nhmmer [6] and Riboseek [7], which collectively span sequence-only search. The dataset was a test set of 155 families held out from the training set, each containing at least one experimentally determined tertiary structure. To ensure data quality, we restricted each family to its seed-alignment sequences, which were manually curated by the Rfam team. One sequence per family was randomly chosen as the query, yielding 155 queries, each searched against the remaining sequences in the test set.

The sensitivity of the methods was measured by recall, with true positives defined by retrieved sequences belonging to the same family as the query. Because the methods return scores on different, non-comparable scales, we compared them by ranking rather than at any shared threshold. Each query's database sequences

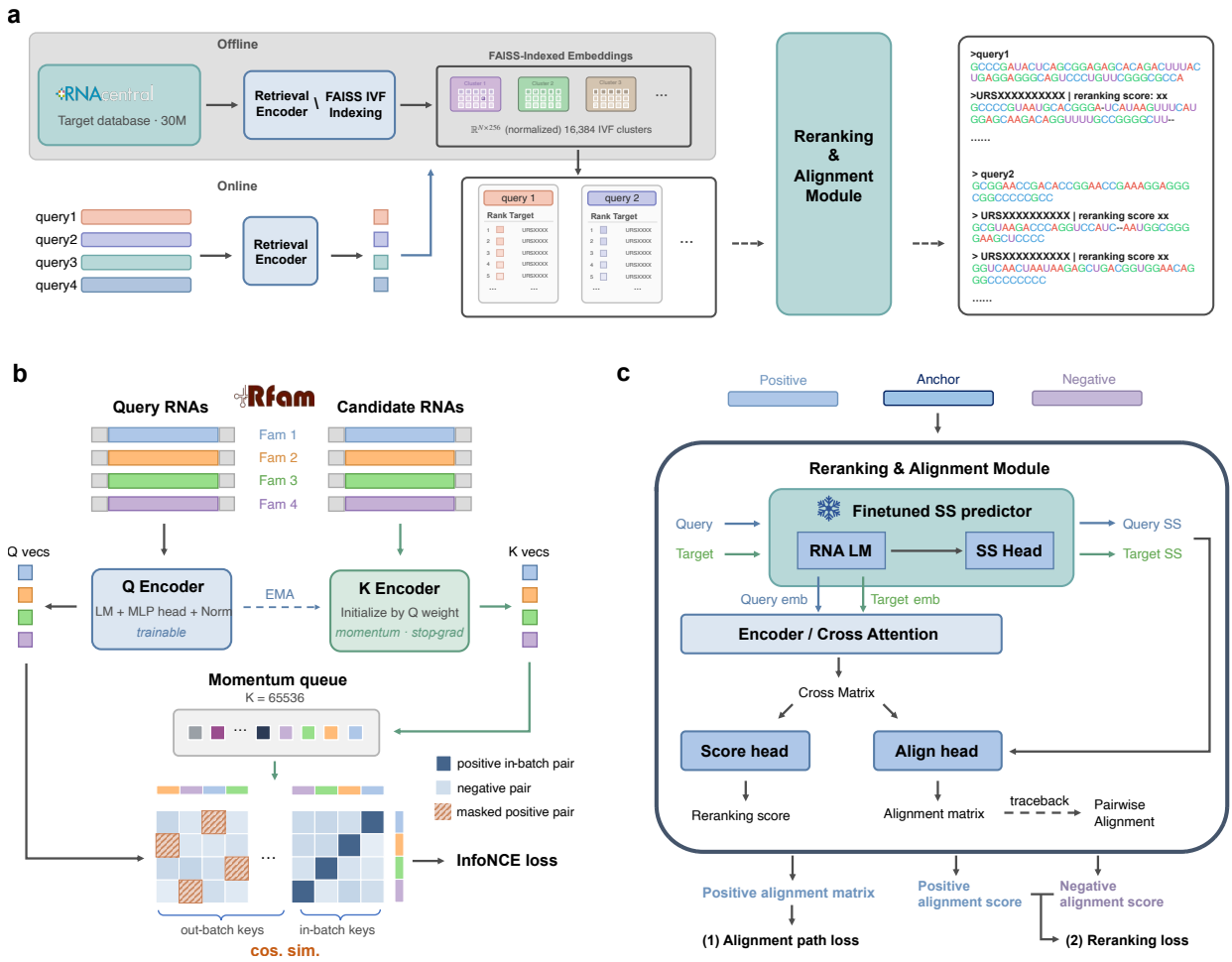


Figure 1: The overall framework of RNAflash. (a) Search pipeline. The query is encoded once and matched against the pre-encoded, IVF-indexed database. Then the top-ranked candidates are reranked by RNAflash-A and their alignments are assembled into an MSA. Further details are provided in the Methods section. (b) RNAflash-R, the retrieval encoder (bi-encoder), for RNA homology search. Two sequences are sampled from each family, extended with random nucleotide flanks at both ends, and encoded separately by the query (Q) and key (K) encoders into dense vectors. The cosine similarity between a pair of vectors is used in the MoCo objective to pull homologous pairs together and push non-homologous pairs apart. (c) RNAflash-A, the cross-encoder, for RNA alignment. An anchor is randomly sampled from a family as the query, and several in-family and out-of-family sequences are chosen as targets. The module learns to generate a reranking score and an alignment matrix from cross features. The alignment matrix is augmented by predicted secondary structure (SS), and alignment is obtained by dynamic programming traceback from the matrix.

were ranked by each method's native metric (E-value for the four benchmark methods, and cosine similarity for RNAflash-R), and we measured recall within the top-k retrieved sequences (recall@k, Fig. 2a). RNAflash-R attained the highest recall@k at most cutoffs, ranking second only to Riboseek at $k = 5, 10, 20$.

Because a fixed cutoff treats families of different sizes unevenly, we additionally evaluated each query at a cutoff matched to its own family size (Recall@FS, Fig. 2b). RNAflash-R again achieved the highest recall@FS, showing that its advantage persisted when retrieval depth was adjusted to family size.

For evaluation of RNAflash-R's ability to balance sensitivity and precision, we also measured recall up to the first false positive (Recall@1FP, Fig. 2c). Each ranking was truncated at its first non-homolog, and recall was computed over this error-free prefix. RNAflash-R outperformed all baselines, recovering on average 75.5% of a query's homologs before the first unrelated sequence appeared. Notably, Riboseek was also trained on Rfam, and its training data included part of our test families, which RNAflash-R was never exposed to during training. Nevertheless, RNAflash-R surpassed Riboseek on recall@FS and recall@1FP. This suggests that RNAflash-R's advantage reflects generalization to unseen families rather than familiarity with the test families. Moreover, the high sensitivity of RNAflash-R depends less on which sequence is taken as the query. To validate this, we repeated the evaluation using every sequence in the test set as a query and computed recall@1FP for each (Fig. 2d). RNAflash-R ranked first under both a macro average over families and a micro average over sequences, showing that its performance is robust to the choice of queries.

To examine retrieval performance within families of low sequence identity, we selected RF00032 (Histone 3' UTR stem-loop) as a representative case study (Fig. 2e). The family exhibits high sequence diversity, with 49 sequences in its seed alignment, and 80% of sequences retained after clustering at 70% sequence identity using MMseqs2, indicating limited redundancy among family members. RNAflash-R outperformed all competing methods across recall levels, achieving the highest area under the precision–recall curve. While sequence-based approaches like nhmmer and Riboseek effectively retrieved homologs in relatively conserved families, their performance degraded in this highly diverse family, where sequence identity provides limited discrimination. In contrast, RNAflash-R preserved its retrieval capability, demonstrating its potential for identifying homologs with reduced sequence identity.

Furthermore, consistent with its retrieval performance, RNAflash-R grouped homologous sequences more closely in the embedding space while those belonging to different families were farther apart (Fig. 2f). Notably, this sensitivity is not inherited from the backbone alone. We used six different RNA language models as the backbone and trained them with the same MoCo objective and training data (Fig. 2g, Supplementary Figure 1). MoCo substantially improved retrieval sensitivity for every backbone, whereas the choice of backbone still mattered. ERNIE-RNA, which incorporates a structural prior [20], gave the highest sensitivity after MoCo training.

2.3 RNAflash-A achieves efficient reranking and structure-aware alignment

Whether RNAflash-A supports MSA construction from embedding-retrieved sequences depends on two capabilities: reranking the retrieved candidates to remove false positives, and aligning the retained candidates. We evaluated these in turn on the Rfam test set, where curated seed alignments and homolog relationships provide a reference, and then examined MSA construction itself at scale, where curated reference alignments were unavailable.

Running on the same Rfam test set as the previous section, RNAflash-A achieved a higher recall@1FP than RNAflash-R and other structure-based algorithms, with 49.7% of families reaching a perfect recall@1FP, showing that it places more of a query's homologs above the first false positive (Fig. 3a). We also measured the precision of different reranking methods under varying depth (precision@k, Fig. 3b). We ranked all

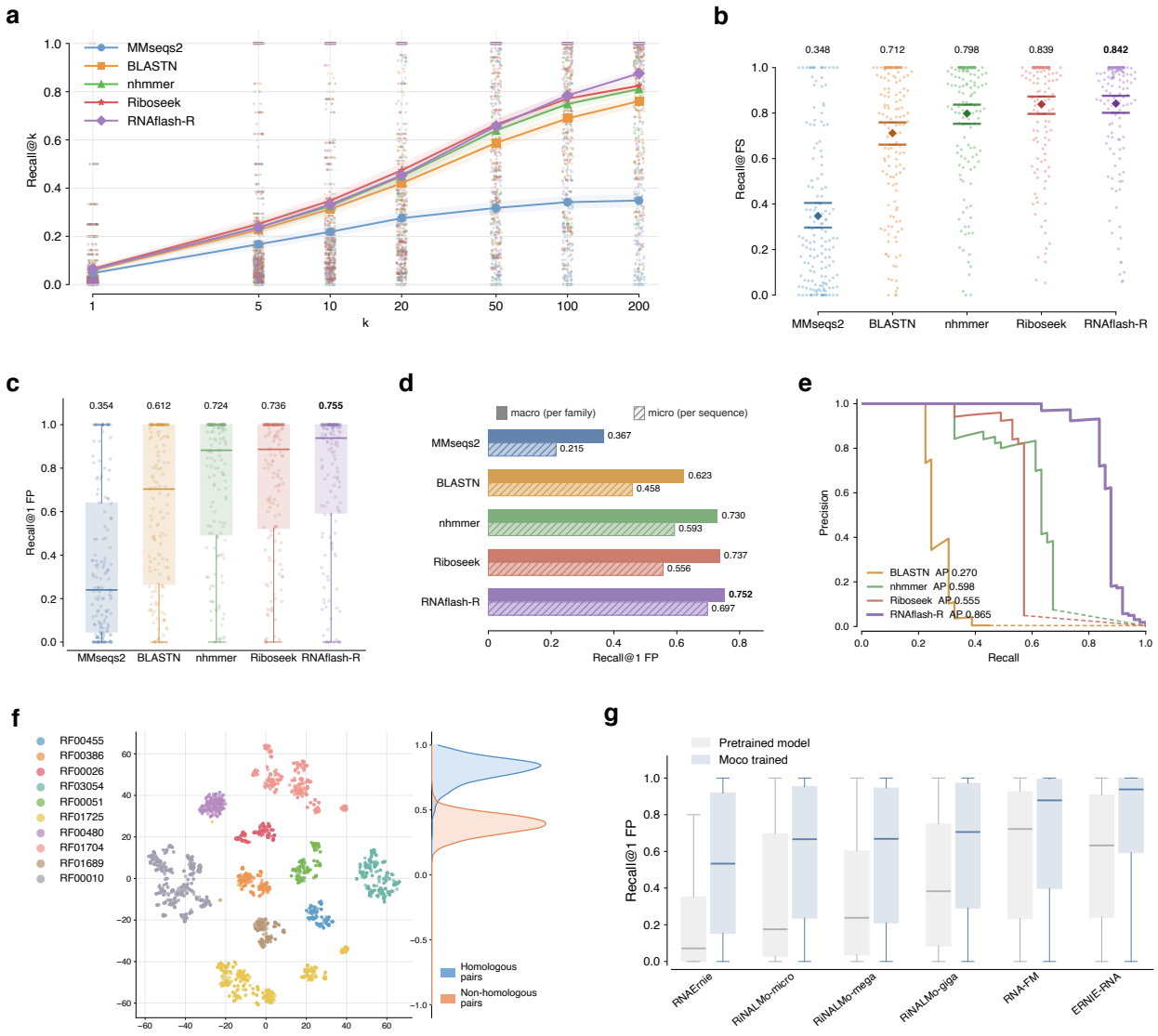


Figure 2: RNAflash-R (retrieval encoder) outperforms sequence-based methods. (a) Average recall at a fixed cutoff k, where $k \in \{1, 5, 10, 20, 50, 100, 200\}$. (b) Recall of 155 queries with a cutoff of family size (recall@FS). Each semi-transparent dot represents a query, and horizontal lines indicate the 95% BCa bootstrap confidence interval for the mean (10,000 resamples). (c) Box plot of recall@1FP for the 155 queries. (d) Macro average and micro average of recall@1FP for all sequences in the test set. (e) A case study on RF00032. For methods failing to retrieve all homologous sequences, all unretrieved sequences (including non-homologous ones) were assigned the lowest possible score and treated as bottom-ranked. (f) Visualization of learned RNA sequence embeddings. Left, t-SNE projection of sequence embeddings from the test set. Right, kernel density estimates (KDEs) of pairwise cosine similarity for homologous pairs and non-homologous pairs. (g) Ablation study on different RNA language models' sensitivity before and after MoCo training, measured by recall@1FP.

target sequences by each method's E-value (nhmmer, Infernal) or score (RNAflash-A). RNAflash-A had precision comparable to Infernal at $k = 1$ and the highest precision at $k = 5$–200. Its advantage was small at shallow ranks and became more pronounced as k increased, indicating fewer false positives among the top-ranked candidates at greater depths. We further evaluated the alignment performance using reference seed alignments from the Rfam test set (Fig. 3c). RNAflash-A achieved an average F1 of 0.807, compared with 0.740 for nhmmer and 0.793 for Infernal.

To examine the MSA construction performance of RNAflash-A following RNAflash-R retrieval, we built MSAs from candidates produced by the retrieval stage rather than from curated alignments. We used 174 single-chain RNAs with experimentally determined structures (resolution ≤ 4.0 Å, length > 20 nt) from RNAsolo [23] as queries and retrieved the top-1000 targets for each from a 30-million-sequence RNAcentral [24] database with RNAflash-R. The same candidate list was passed to nhmmer, Infernal and RNAflash-A to test the MSA construction ability.

We evaluated the MSA quality by number of effective sequences (Neff) and MSA-based secondary structure prediction. The number of effective sequences was calculated with NEFFy [25], which measures the number of non-redundant sequences in an MSA. The MSA-based structure prediction was performed by plmc [26] and R-scape [27], which assign scores to pairs of alignment columns by coupling analysis and covariation analysis on an MSA respectively. RNAflash-A required an average of 9.67 seconds per query, whereas Infernal required $26.7\times$ as much time (Fig. 3d). RNAflash-A produced MSAs with higher Neff than nhmmer and Infernal (Fig. 3e) and higher AUPRCs in both coupling and covariation analyses (Fig. 3f, g). Nhmmer was faster than RNAflash-A, but its MSAs had lower effective depth and weaker evolutionary signals. Its speed nevertheless makes it an attractive option when throughput is a priority, motivating the lightweight pipeline evaluated below.

We also studied whether the lightweight structure augmentation is necessary in MSA construction by retraining RNAflash-A with an identical backbone, architecture and training data, but with the secondary-structure augmentation removed (Fig. 3h). Removing structure augmentation substantially reduced AUPRCs in both coupling and covariation analyses, highlighting the importance of augmentation in achieving the alignment quality reported above.

2.4 RNAflash enables fast sensitive homology search for RNA MSA construction

We next evaluated RNAflash as a complete pipeline for RNA homology search and MSA construction on RNAcentral database. We compared RNAflash with sequence-based methods (nhmmer and Riboseek), the CM-based method Infernal, and RNAcmap, which combines sequence- and structure-based searches. Each method performed its own search and alignment steps, and by default, RNAflash-R retrieved the top 1,000 candidates for each query, which were then filtered and aligned by RNAflash-A.

On the full query set, RNAflash produced MSAs with higher Neff than the sequence-based methods (Fig. 4a). The RNAflash-derived MSAs also showed stronger evolutionary coupling and covariation signals, with higher AUPRCs than those generated by sequence-based methods and Infernal. RNAcmap achieved the highest AUPRCs in both analyses, followed by RNAflash (Fig. 4b).

We further examined performance on 22 RNAsolo queries without detectable matches to Rfam covariance models in our cmscan search, referred to here as Rfam-unassigned queries. RNAflash maintained similar Neff values across the two query groups, while those of competing methods dropped substantially on the Rfam-unassigned subset (Fig. 4c). RNAflash also achieved the highest coupling and covariation AUPRCs among the methods tested (Fig. 4d). Infernal showed the largest performance decline on this subset, with performance approaching that of Riboseek. This pattern may partly reflect the CM-based criterion used to select these queries: the absence of detectable Rfam CM matches may identify queries that are also

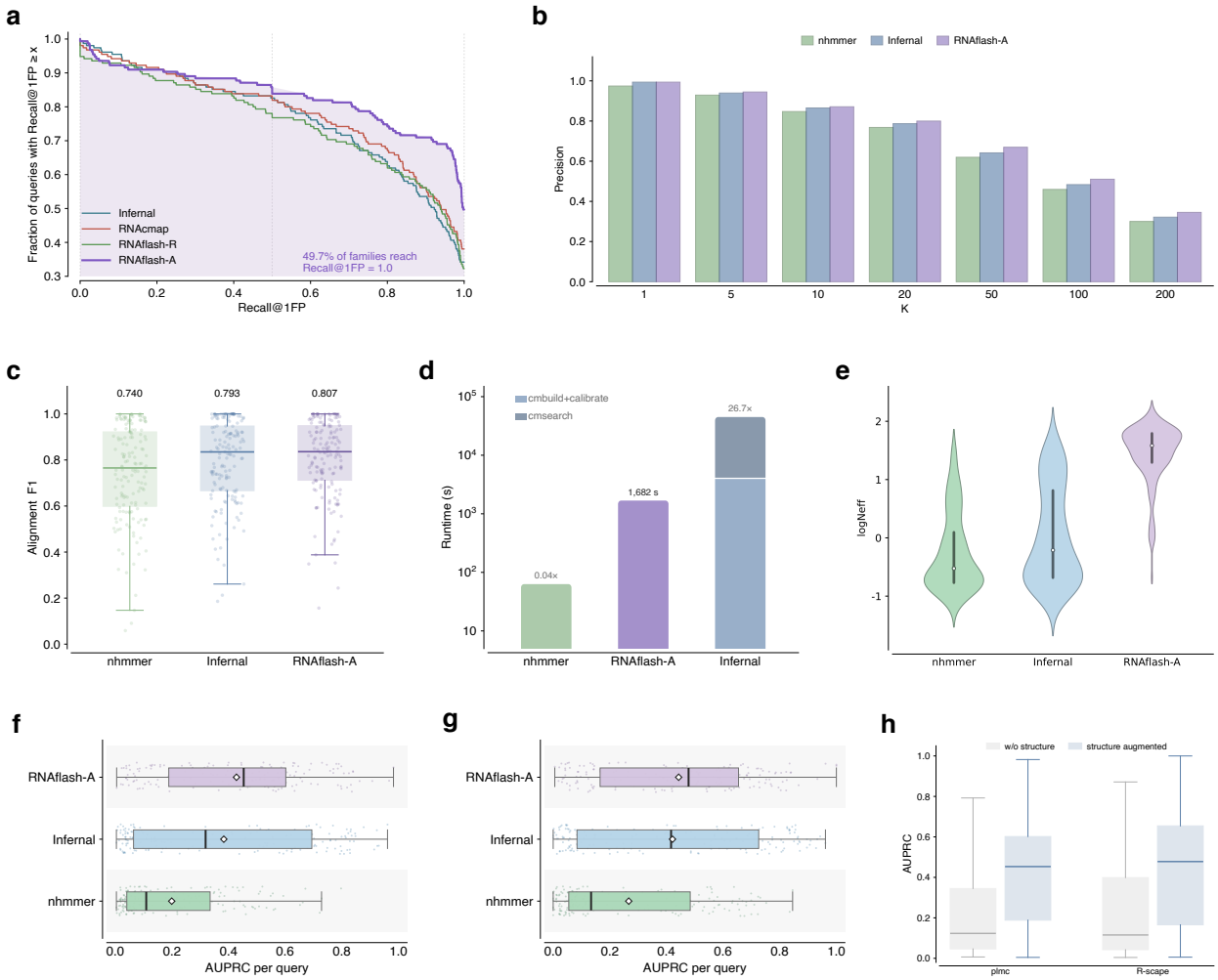


Figure 3: Performance of RNAflash-A in reranking and alignment. (a) Cumulative distribution of recall@1FP over the 155 Rfam test families. (b) Precision within the top-k ranked sequences (precision@k) for nhmmer, Infernal and RNAflash-A. (c) F1 of the predicted pairwise alignments against the Rfam seed alignments for the same three methods. The mean F1 of each method is above the corresponding box. (d) Runtime of reranking-alignment algorithms on 174 queries' top-1000 candidates retrieved by RNAflash-R. Fold changes are relative to RNAflash-A. (e) Log-Number of effective sequences in MSAs (logNeff). Neff was normalized by the square root of the query length ($\sqrt{L}$), and then $\log 10$-transformed. (f, g) Coupling analysis and covariation analysis of the generated MSAs. For each MSA, plmc and R-scape generated scores for pairs of alignment columns. Nucleotide pairs with a sequence distance of 4 nt or fewer were excluded, and AUPRC was calculated against the reference secondary structure to assess MSA quality. The diamond denotes mean, and the black line denotes median. (h) Ablation study on the structure augmentation (identical backbone, architecture and training data, augmentation removed), evaluated by the AUPRC of plmc and R-scape.

challenging for CM-based homology search. RNAflash's performance on this subset suggests that its utility extends beyond queries readily assigned to known Rfam families.

Moreover, RNAflash completed search and MSA construction for all 174 queries in 1,682 seconds, the shortest end-to-end runtime among the methods compared (Fig. 4e). It was 3.7-fold faster than non-iterative Riboseek and approximately 10-fold faster than iterative Riboseek, even though Riboseek was GPU-accelerated on the same device in this benchmark. Infernal and RNAcmap required over $100\times$ and $142\times$ the runtime of RNAflash, respectively, highlighting the computational advantage of RNAflash's two-stage pipeline. Notably, when used for retrieval alone instead of MSA construction, RNAflash completed all 174 queries in 12 seconds (0.07 seconds per query on average, with the Faiss index resident in cache). Also, the retrieval time did not change significantly with the increase of retrieval depth (Supplementary Figure 2). This low retrieval cost also supports using RNAflash-R independently as an ultra-fast search tool for large databases when ranked homolog candidates are sufficient and an MSA is not required.

The retrieval depth k determines how many candidates are passed to RNAflash-A and therefore provides a way to adjust the balance between MSA quality and computational cost. We varied k from 500 to 10,000 to examine this trade-off (Fig. 4f). Increasing k improved coupling and covariation AUPRCs and increased Neff, although MSA construction took longer as more candidates required pairwise processing. At the default setting, RNAflash was fast enough to make homology search scalable. When k increased to 5000, it achieved MSA quality comparable to RNAcmap while remaining faster than iterative Riboseek.

2.5 RNAflash provides a lightweight route for large-scale RNA homology search

Nhmmer is a widely used tool for homolog detection and MSA construction, but its runtime grows with the size of the target database, because its filter still visits every sequence before the HMM alignment stages. As discussed in Section 2.3, nhmmer is fast on a small candidate set retrieved by RNAflash-R, although the resulting MSAs had lower Neff and weaker evolutionary signals than those constructed by RNAflash-A. Therefore, we tested whether the retrieval stage of RNAflash could accelerate nhmmer by narrowing its search to the retrieved candidates, and whether this lightweight configuration, which we call RNAflash-L, improves nhmmer's output. The lower alignment cost allowed us to retrieve more candidates, and we selected $k = 5{,}000$, beyond which gains in coupling and covariation AUPRCs plateaued (Supplementary Figure 3).

To assess whether RNAflash retrieval can recover sequences identified by nhmmer search, we selected the queries in the RNAsolo dataset for which nhmmer returned hits when searching the full RNAcentral database. For the 141 queries, we performed RNAflash-R retrieval against the same database and measured the fraction of nhmmer hits recovered by RNAflash-R (Fig. 5a). In 63 queries, RNAflash-R recovered all the nhmmer-retrieved sequences, and in half of the queries, it recovered more than 98% of nhmmer hits. Restricting the search to a small high-confidence set also improves sensitivity through E-values (Supplementary Figure 4), which scale with the size of the search space. For the other 33 queries for which nhmmer returned no hits from RNAcentral, our lightweight pipeline returned at least one homolog in 21 queries (Fig. 5b), including more than 100 hits for two queries.

Beyond improving hit recovery, RNAflash-L completed search and MSA construction in less than one second per query on average (Fig. 5c). We further assessed whether the improved hit recovery translated into better MSA quality (Fig. 5d). For most queries in the RNAsolo dataset, MSAs generated by RNAflash-L were at least as deep as those generated by standalone nhmmer, with higher effective sequence counts and stronger coupling and covariation signals. Interestingly, the improvements were not confined to queries with deeper MSAs. For some queries, RNAflash-L retained fewer sequences yet achieved higher Neff and AUPRCs, suggesting that its retained sequences provided more non-redundant evolutionary informa-

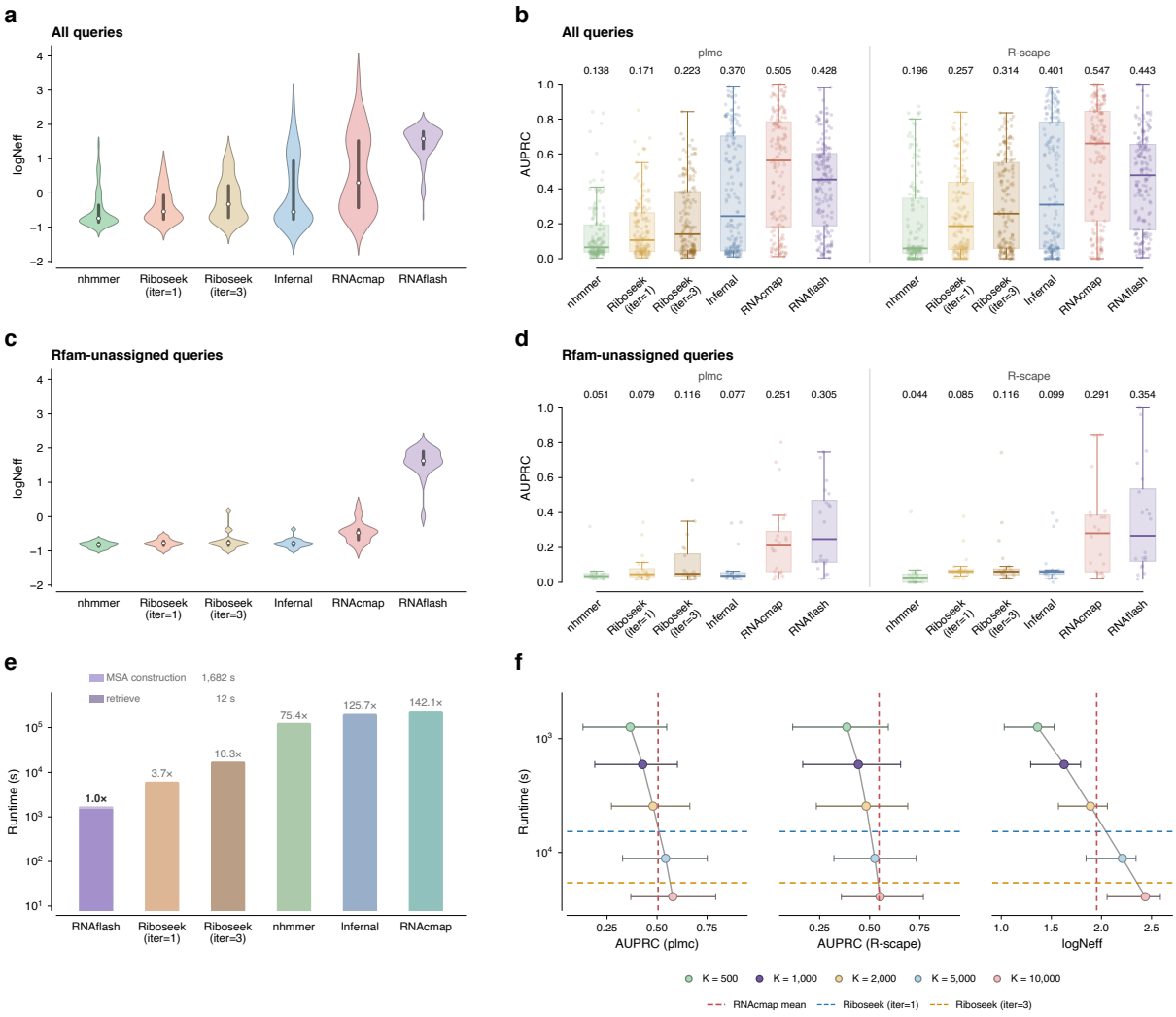


Figure 4: Performance of RNAflash full pipeline on RNAcentral database. (a) Number of effective sequences (Neff) of the MSAs for the 174 RNAsolo queries, on a log-10 scale. (b) AUPRC of coupling and covariation analysis on the same MSAs (queries without annotated base pairs were excluded). Average AUPRC is presented above each box. (c, d) Neff and AUPRC restricted to queries without detectable matches to Rfam covariance models. (e) End-to-end runtime for the 174 queries. Retrieval and MSA construction time are denoted by light and dark segments. (f) MSA quality against runtime with the increase of retrieval depth k, represented by interquartile range and mean value. Dashed horizontal lines denote the runtimes of the two Riboseek configurations, and the dashed vertical line marks the mean AUPRC of RNAcmap.

tion. RNAflash-L also outperformed non-iterative Riboseek in Neff and coupling and covariation AUPRCs (Fig. 5e). Among the three pipelines, RNAflash-L also had the shortest runtime (Fig. 5c). These findings illustrate how learned retrieval can extend the capabilities of current homology search tools, making informative RNA MSAs more accessible for large-scale downstream analyses.

Furthermore, given the use of nhmmer-derived MSAs in RNA structure prediction pipelines, RNAflash-L could serve as an alternative MSA generation workflow compatible with Protenix-v2 [28] and AlphaFold3 [29]. On the RNAsolo benchmark, RNAflash-L gave Protenix-v2 the highest TM-score and lowest RMSD among the MSA sources tested. This combination also outperformed all tested AlphaFold3 configurations on both metrics (Supplementary Table 1). On the Foldbench dataset [30], RNAflash-L also gave Protenix-v2 the highest TM-score among the MSA sources tested (Supplementary Table 2).

3 Discussion

To address the trade-off between speed and sensitivity in conventional algorithms, we introduced RNAflash, a two-stage pipeline for RNA homology search and MSA construction. By combining embedding-based retrieval with structure-informed reranking and alignment, it made sensitive RNA homology search more scalable.

On held-out Rfam families, RNAflash-R improved homolog retrieval over sequence-based methods, while the contrastive training framework improved all tested RNA language-model backbones. RNAflash-A produced accurate alignments and MSAs with stronger evolutionary signals than nhmmer and Infernal on the same retrieved candidates. The complete RNAflash pipeline demonstrated higher sensitivity and MSA quality than sequence-based methods and Infernal, with a considerably shorter execution time. The complementary RNAflash-L configuration improved nhmmer-based searches and benefited Protenix-v2 structure prediction. The latter may reflect compatibility with the nhmmer-derived MSAs used to train Protenix-v2. Together, these results show how learned representations and pairwise structural information can support both large-scale retrieval and downstream MSA construction.

Several considerations define the scope of these results. The benchmarks covered 155 structurally characterized Rfam families and 174 RNAsolo queries, and the ERNIE-RNA backbone restricts the current implementation to sequences of up to 1,022 nt. Because reference alignments were unavailable for the large-database searches, MSA quality was assessed using effective sequence counts and coupling and covariation signals. RNAflash also trades retrieval depth for substantially lower runtime, and the retrieval depth should be adjusted to balance MSA quality and computational cost, with RNAflash-L providing a lighter alternative.

Overall, these results position RNAflash as a solution for scalable retrieval and alignment of remote RNA homologs. It makes evolutionary information more accessible for RNAs whose homologs may be difficult to detect by sequence identity alone, and provides a promising option for MSA-based analyses.

4 Methods

4.1 Benchmark datasets

Rfam training set. We use Rfam to train our models. The Rfam database is a collection of non-coding RNA families, and sequences within a family are treated as homologous. For RNAflash-R, all sequences in a family are used, and we remove duplicate sequences within each family and exclude families containing

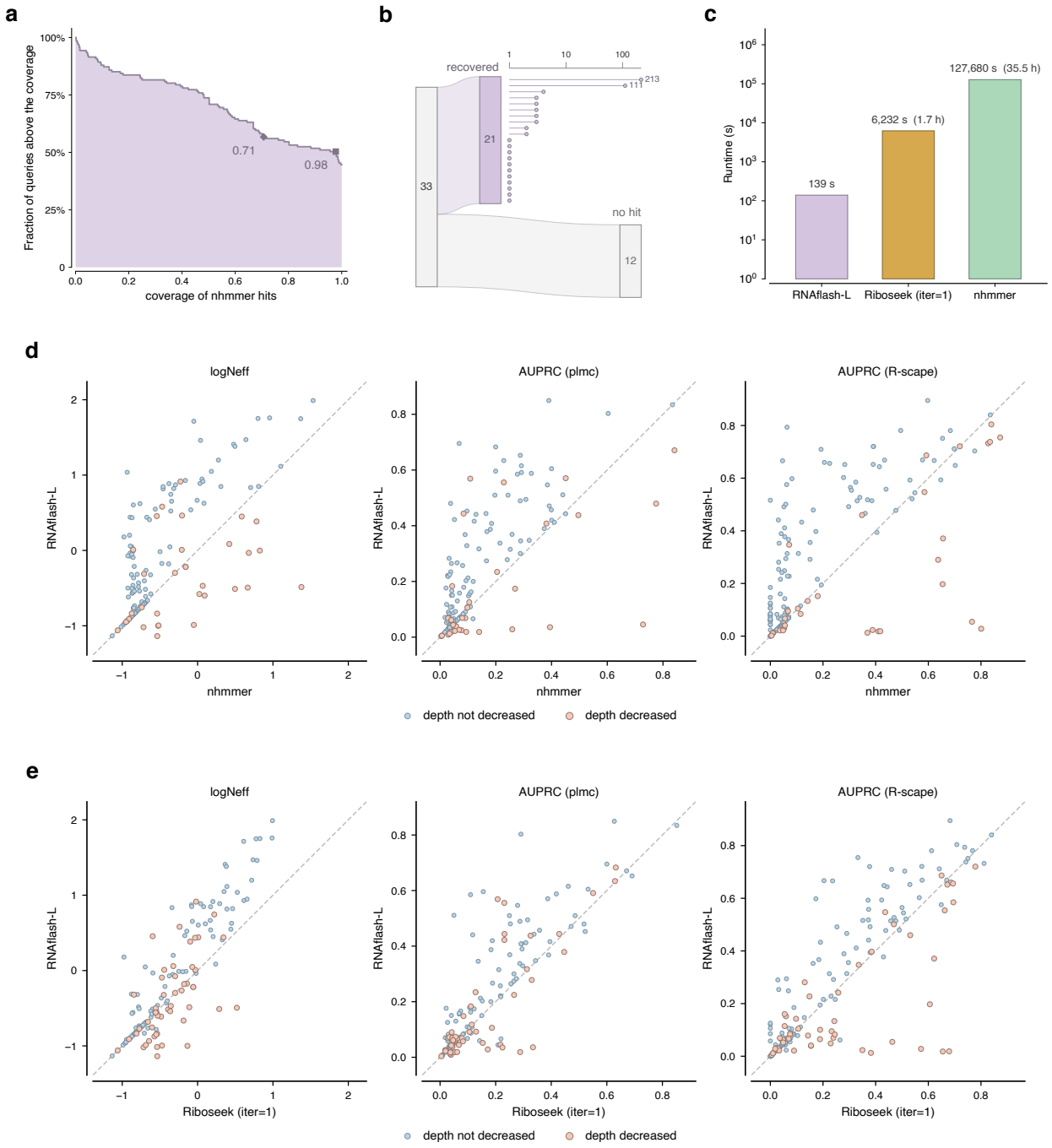


Figure 5: Lightweight pipeline of RNAflash. (a) Fraction of nhmmer's hits recovered by RNAflash-R. The diamond and square denote mean and median, respectively. (b) Number of RNAflash-L's hits in queries for which nhmmer returned no hit from the full database. (c) Runtime of search pipelines against the full database. (d) Improvements in Neff and coupling and covariation AUPRCs relative to standalone nhmmer. Points are coloured by whether the number of sequences retained by RNAflash-L was not lower (blue) or lower (orange) than for default nhmmer. (e) Comparison of Neff, coupling analysis and covariation analysis AUPRC between RNAflash-L and non-iterative Riboseek.

only one unique sequence. RNAflash-A only used seed alignments which are manually curated to ensure data quality. To separate training and test families, the Rfam training set includes only families without experimentally determined 3D structures and contains 4,031 families.

Rfam validation set. 10% of RNA families (402 families) in Rfam training set are randomly held out for model validation, ensuring that the training and validation sets contain disjoint families. The other 3,629 families are used for training.

Rfam test set. Families with at least one experimentally determined 3D structure are held out as the test set (155 families). This family-level split ensures that training and test sets contain no overlapping families, allowing us to evaluate whether RNAflash generalizes to homologous relationships within families unseen during training. Restricting the test set to structurally characterized families also matches our later evaluation, where all query RNAs have known tertiary structures.

RNA 2D structure prediction benchmark. For evaluation of MSA quality, we use a benchmark set collected from RNAsolo, which is a curated database of RNA three-dimensional structures that integrates experimentally determined RNA-containing structures from the Protein Data Bank (PDB) and provides standardized sequence, 3D structure and 3D-structure-derived secondary structure information for RNA molecules. We choose the representative RNA set provided by RNAsolo with a resolution of 4.0 Å or better. Structures containing multiple RNA chains or RNA assemblies composed of multiple interacting subunits are excluded, and redundant sequences are removed using MMseqs2 with a sequence identity threshold of 80%. The filtering steps yield a final set of 174 RNA queries. Eight queries have no annotated base pairs and are excluded from coupling and covariation analyses.

RNA 3D structure prediction benchmark. The 2D structure dataset is also used as a tertiary structure prediction benchmark, but only sequences released after 2022 are selected to prevent data leakage, as AlphaFold3 and Protenix-v2 are both trained with structures released before 2022. The dataset consists of 51 structures. We also use Foldbench, a dataset designed to evaluate the accuracy and generalizability of structure prediction models, and select the 12 single-chain RNAs in the dataset for evaluation.

4.2 RNAflash-R for fast RNA search

RNAflash-R, the retrieval encoder in RNAflash, uses pretrained ERNIE-RNA[20] as its backbone. ERNIE-RNA is a BERT-based RNA language model pre-trained on non-coding RNA sequences from RNAcentral[24]. It incorporates a base-pairing-informed, all-against-all pairwise attention bias into the self-attention mechanism, thereby learning structure-enhanced representations without requiring explicit secondary-structure annotations and providing both token embeddings and attention maps.

For an RNA sequence q of length L, the ERNIE-RNA backbone first extracts nucleotide embeddings $\mathbf{H} \in \mathbb{R}^{L\times d}$. The embeddings are projected to 256 dimensions through an MLP, averaged over sequence length and L2-normalized to obtain a dense sequence embedding $\mathbf{q}$. Queries longer than 1022 nt are truncated to 1022 to fit the maximum input length of ERNIE-RNA. The sequence similarity is computed as the dot product

$$s(q, d_i) = \mathbf{q}^T \mathbf{d}_i, \tag{1}$$

where $\mathbf{d}_i$ is the sequence embedding of the i-th sequence in the database. Before homolog retrieval, the database is encoded using the same pipeline, with sequences longer than 1022 discarded. Also, a FAISS inverted-file (IVF) index is constructed over the database vectors, yielding 16,384 learned clusters. At retrieval time, the search proceeds at the cluster level. The database index first locates the nearest clusters and then retrieves the top k closest vectors within these clusters. The retrieved vectors are finally mapped back to the original sequences, which are written to an output file.

To enable homology retrieval, we fine-tuned the entire model under a MoCo-style contrastive framework. A FIFO queue $\{(\mathbf{k}_n, \text{fam}(n))\}_{n=1}^{K}$ stores K sequence embeddings with their family labels. Queued embeddings from the query's family are excluded from the negative set. The query encoder, including its backbone and projection head, is updated by backpropagation, while the key encoder is initialized from the query encoder and updated only through the exponential moving average (EMA) as

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q,\ m = 0.999, \tag{2}$$

where θ_k and θ_q are the parameters of the query encoder and key encoder, and m is the update momentum. Each training batch contains B sequence pairs $\{(q_i, k_i)\}_{i=1}^{B}$ from different homologous families, where q_i and k_i are sequences randomly chosen from the same family, used as the query and key sequences, respectively. At each training step, the query encoder and key encoder map query sequences and key sequences to sequence embeddings, $\{\mathbf{q}_i\}_{i=1}^{B}$ and $\{\mathbf{k}_i\}_{i=1}^{B}$. For query q_i, the positive is its in-pair key k_i, and the negatives are the other $B-1$ keys in the batch, together with negative keys from a FIFO queue. The InfoNCE loss for pair i is

$$\mathcal{L}_i = -\log \frac{\exp(\mathbf{q}_i^T \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^T \mathbf{k}_i/\tau) + \sum_{j\neq i}^{B} \exp(\mathbf{q}_i^T \mathbf{k}_j/\tau) + \sum_{\text{fam}(n)\neq\text{fam}(i)}^{K} \exp(\mathbf{q}_i^T \mathbf{k}_n/\tau)}, \tag{3}$$

where τ is a learnable temperature parameter, and the total loss is averaged over the B pairs.

Unlike nhmmer and other algorithms, RNAflash-R compares averaged embeddings of entire sequences without explicitly performing local alignment. Therefore, we additionally introduce random nucleotide flanks during training to encourage retrieval of sequences containing homologous regions despite unrelated flanking subsequences. Random nucleotide flanks are added to both sequences in a sampled pair with probability 0.5. For each sequence, the total flank length is sampled uniformly from zero to 20% of the original sequence length and capped to keep the augmented sequence within 1,022 nt. The flanks are randomly distributed between the 5' and 3' ends, with each added nucleotide sampled uniformly from A, U, C and G. This augmentation improves the recovery of nhmmer hits across retrieval depths (Supplementary Figure 5).

4.3 RNAflash-A for fast RNA alignment

RNAflash-A is the reranking and alignment module. It uses the secondary-structure-prediction version of ERNIE-RNA as the backbone. The backbone is frozen, and we train a downstream module for reranking and alignment. The module consists of a transformer encoder, a cross-attention module, and two MLP heads for reranking and alignment.

At inference time, given a query sequence q and target sequences $\{t_i\}_{i=1}^{N}$, RNAflash-A first extracts nucleotide embeddings $H_q \in \mathbb{R}^{L_q \times d}$, $H_{t_i} \in \mathbb{R}^{L_{t_i} \times d}$ and predicts secondary structure using the backbone. The nucleotide embeddings are projected by the transformer encoder, and then merged into a cross feature $C_i \in \mathbb{R}^{L_q \times L_{t_i} \times h}$ using multi-head cross-attention scoring. The residue-pair alignment score matrix S_i and reranking logit r_i are computed from C_i by alignment head and reranking head. S_i is then augmented by secondary structure in a lightweight approach:

$$\tilde{S}_i = S_i + P_q S_i P_{t_i}, \tag{4}$$

where P_q and P_{t_i} denote the predicted secondary structure pairing matrices of the query and target sequence. The reranking logit is converted into a reranking score $p_i = \sigma(r_i)$ and targets with $p_i <$ threshold are filtered out. For the remaining targets, the optimal residue correspondence is obtained by dynamic programming on $\tilde{S}_i$, followed by alignment construction. The retained targets are finally ranked according to p_i.

Algorithm 1 Inference of RNAflash-A

Require: Query sequence q, target sequences $\mathcal{T} = \{t_i\}_{i=1}^{N}$, model f, threshold τ
Ensure: Ranked target sequences with alignments

$(H_q, A_q) \leftarrow \mathrm{Backbone}(q)$
$P_q \leftarrow \mathrm{SSHead}(A_q)$
$E_q \leftarrow \mathrm{Encoder}(H_q)$
$(H_t, A_t) \leftarrow \mathrm{Backbone}(\mathcal{T})$
$E_t \leftarrow \mathrm{Encoder}(H_t)$
$C \leftarrow \mathrm{CrossAttention}(E_q, E_t)$
$(S, r) \leftarrow f_{\mathrm{head}}(C)$
$P_t \leftarrow \mathrm{SSHead}(A_t)$
$\tilde{S} \leftarrow S + P_q S P_t$
$p \leftarrow \sigma(r)$
$\mathcal{R} \leftarrow \varnothing$
for each target t_i with $p_i \geq \tau$ **do**
  $M_i \leftarrow \mathrm{DP} - \mathrm{Decode}(\tilde{S}_i)$ {Algorithm 2}
  $(q_i, t_i') \leftarrow \mathrm{BuildAlignment}(q, t_i, M_i)$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(q_i, t_i', p_i)\}$
end for
$\mathcal{R} \leftarrow \mathrm{SortDescending}(\mathcal{R}, p)$
return $\mathcal{R}$

During training, each sample contains a query RNA sequence q, a set of homologous targets $\{t_i^+\}_{i=1}^{B}$, an equal-sized set of non-homologous targets $\{t_i^-\}_{i=1}^{B}$ and reference alignment matrices for homologous pairs. RNAflash-A predicts a reranking score for each query-target pair, and a residue-pair alignment score matrix for homologous pairs. Let $A_i \in \{0,1\}^{L_q \times L_{t_i^+}}$ be the reference alignment matrix, $\hat{A}_i = \sigma(\tilde{S}_i^+)$ the predicted alignment probabilities, and $p_i^{\pm} = \sigma(r_i^{\pm})$ the homology probabilities. We define weighted BCE as

$$\ell_w(y, p) = -wy \log p - (1 - y) \log(1 - p), \tag{5}$$

and the three losses are defined as

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{L_q L_{t_i^+}} \sum_{u=1}^{L_q} \sum_{v=1}^{L_{t_i^+}} \ell_{100}(A_{iuv}, \hat{A}_{iuv}), \tag{6}$$

$$\mathcal{L}_{\mathrm{hom}} = \frac{1}{2B} \sum_{i=1}^{B} \left[\ell_1(1, p_i^+) + \ell_1(0, p_i^-) \right], \tag{7}$$

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{B} \sum_{i=1}^{B} \max\left(0, \mu - r_i^+ + r_i^-\right), \tag{8}$$

where μ is the ranking margin, and padding and special tokens are excluded from alignment loss. The total objective is $\mathcal{L} = \lambda_a \mathcal{L}_{\mathrm{align}} + \lambda_h \mathcal{L}_{\mathrm{hom}} + \lambda_r \mathcal{L}_{\mathrm{rank}}$.

4.4 Benchmark baseline methods

To evaluate the sensitivity of RNAflash in homologous RNA retrieval and the quality of the resulting MSAs, we use six established methods for comparison:

Algorithm 2 Dynamic programming decoder for residue correspondences

Require: Alignment score matrix $S \in \mathbb{R}^{m \times n}$ for the m query and n target nucleotides
Ensure: Ordered set of matched nucleotide positions M
  Initialize score matrix $D \in \mathbb{R}^{(m+1)\times(n+1)}$ and traceback matrix T to zero
  for $i = 1$ to m **do**
    for $j = 1$ to n **do**
      $u \leftarrow D_{i-1,j}$; $l \leftarrow D_{i,j-1}$; $d \leftarrow D_{i-1,j-1} + S_{ij}$
      if $u \geq l$ and $u \geq d$ **then**
        $D_{ij} \leftarrow u$; $T_{ij} \leftarrow \text{up}$
      else if $l \geq u$ and $l \geq d$ **then**
        $D_{ij} \leftarrow l$; $T_{ij} \leftarrow \text{left}$
      else
        $D_{ij} \leftarrow d$; $T_{ij} \leftarrow \text{diag}$
      end if
    end for
  end for
  $M \leftarrow [\,]$; $(i, j) \leftarrow (m, n)$
  while $i > 0$ and $j > 0$ **do**
    if $T_{ij} = \text{diag}$ **then**
      Append (i, j) to M; $(i, j) \leftarrow (i - 1, j - 1)$
    else if $T_{ij} = \text{left}$ **then**
      $j \leftarrow j - 1$
    else
      $i \leftarrow i - 1$
    end if
  end while
  return $\text{Reverse}(M)$

- MMseqs2. MMseqs2 is a fast sequence search tool for efficient large-scale sequence comparison. It combines k-mer-based prefiltering and ungapped alignment with local sequence alignment to identify homologous sequences in large databases.
- BLASTN. BLASTN is a widely used nucleotide sequence search algorithm that identifies local sequence similarities through a heuristic seed-and-extend strategy. It detects homologous regions based on sequence information without explicitly considering RNA secondary structure.
- nhmmer. nhmmer is a profile hidden Markov model (HMM)-based nucleotide homology search tool implemented in the HMMER package. It uses probabilistic sequence profiles to identify homologous regions, providing improved sensitivity for detecting divergent nucleotide sequences.
- Riboseek. Riboseek is a nucleotide homology search and alignment method that represents sequences as overlapping di-mers. It uses a pipeline similar to that of MMseqs2 and supports iterative profile refinement to improve the detection of remote homologs.
- Infernal. Infernal is an RNA homology search and alignment tool based on covariance models (CMs), which jointly capture sequence conservation and RNA secondary structure. By modeling conserved base pairs, it enables structure-aware detection and alignment of homologous RNAs.

- RNAcmap. RNAcmap is an automated pipeline for RNA homolog search and MSA generation, originally developed for evolutionary coupling-based contact prediction. It combines an initial BLASTN search with secondary structure prediction and subsequent covariance model-based searching using Infernal to identify additional RNA homologs.

To ensure a fair comparison, thresholds of all algorithms were set above their defaults in the retrieval test. The E-value threshold is set to 10 for MMseqs2 (its default threshold is 1e-3), and 10000 for other algorithms. For database search, the parameters of each algorithm are set as default. The structure used in Infernal is predicted by RNAfold, which is also the source of structure information in RNAcmap. All experiments were conducted on a server equipped with two Intel Xeon Silver 4216 processors running at 2.10 GHz (32 physical cores and 64 threads in total), approximately 188 GiB of system memory, and 1 NVIDIA GeForce RTX 3090 GPU, running Ubuntu 16.04.6 LTS.

4.5 Performance metrics

To evaluate the quality of the generated MSAs, we use NEFFy, plmc and R-scape. NEFFy is a tool for computing the effective sequence count in an MSA. We normalize this count by the square root of the query length:

$$\text{Neff} = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} I[S_{m,n} \geq \theta]}, \tag{9}$$

where L is the length of the query sequence, N is the depth of the MSA, $S_{m,n}$ is the similarity between sequence m and sequence n in the MSA, θ is the similarity threshold. When $S_{m,n} \geq \theta$, $I[S_{m,n} \geq \theta] = 1$. Otherwise, $I[S_{m,n} \geq \theta] = 0$.

Plmc is used to analyze evolutionary couplings in an MSA. It fits a global maximum-entropy model to the input MSA by pseudolikelihood maximization with L2 regularization, yielding single-site fields and pairwise coupling parameters. R-scape is used to assess whether the MSA provides statistically significant evidence for an evolutionarily conserved RNA secondary structure. It computes pairwise covariation statistics for alignment columns and evaluates their significance against simulated null alignments. Both methods generate scores for pairs of alignment columns, and these scores are used to calculate an AUPRC against true secondary structure.

Code Availability

Source Code of RNAflash is available on GitHub (https://github.com/AcetoneChen/RNAflash).

Data Availability

Datasets for the evaluation of RNAflash are available at https://zenodo.org/records/23599961.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62473257, 62573293), and the Science and Technology Commission of Shanghai Municipality (24ZR1435300, 24510714300).

Author Contributions

X.P. conceived and supervised the project. B.C. collected and preprocessed the data and developed and trained the model under the methodological guidance of X.P. B.C., S.W., W.Z., Y.X., and X.P. drafted the manuscript, with contributions from H.B.S. All authors reviewed, revised, and approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.

References

[1] Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng, and Jianyi Yang. trrosettarna: automated prediction of rna 3d structure with transformer network. *Nature communications*, 14(1):7266, 2023.

[2] Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024.

[3] Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic acids research*, 52(1):e3–e3, 2024.

[4] Nancy Ontiveros-Palacios, Emma Cooke, Eric P Nawrocki, Sandra Triebel, Manja Marz, Elena Rivas, Sam Griffiths-Jones, Anton I Petrov, Alex Bateman, and Blake Sweeney. Rfam 15: Rna families database in 2025. *Nucleic acids research*, 53(D1):D258–D267, 2025.

[5] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1): 421, 2009.

[6] Travis J Wheeler and Sean R Eddy. nhmmer: Dna homology search with profile hmms. *Bioinformatics*, 29(19):2487–2489, 2013.

[7] Sukhwan Park, Kieran Didi, Andrew Favor, Anton Bushuiev, Soohyun Kim, Milot Mirdita, and Martin Steinegger. Fast remote nucleotide sequence alignment with riboseek. *bioRxiv*, pages 2026–07, 2026.

[8] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[9] Niek R Scholten, Dennis Haandrikman, Joshua O Tolhuis, Edoardo Morandi, and Danny Incarnato. Shapewarp-web: sequence-agnostic search for structurally homologous rna regions across databases of chemical probing data. *Nucleic Acids Research*, 52(W1):W362–W367, 2024.

[10] Brittany Lasher and David A Hendrix. bprna-cosmos: a robust and efficient rna structural comparison method using k-mer based cosine similarity. *Bioinformatics*, 41(4):btaf108, 2025.

[11] Sebastian Will, Tejal Joshi, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Locarna-p: accurate boundary prediction and improved detection of structural rnas. *Rna*, 18(5):900–914, 2012.

[12] Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods*, 19(9):1109–1115, 2022.

[13] Zijie Qiu, Sheng Xu, Junkang Wei, Tao Shen, and Siqi Sun. Rtm-align: an improved rna alignment tool with enhanced short sequence performance via post-standardization and fragment alignment. *bioRxiv*, pages 2024–05, 2024.

[14] Ke Chen, Thomas Litfin, Jaswinder Singh, Jian Zhan, and Yaoqi Zhou. Mars and rnacmap3: the master database of all possible rna sequences integrated with rnacmap for rna homology search. *Genomics, proteomics & bioinformatics*, 22(1):qzae018, 2024.

[15] Chengxin Zhang, Yang Zhang, and Anna Marie Pyle. rmsa: a sequence search and alignment algorithm to improve rna structure modeling. *Journal of Molecular Biology*, 435(14):167904, 2023.

[16] Liang Hong, Zhihang Hu, Siqi Sun, Xiangru Tang, Jiuming Wang, Qingxiong Tan, Liangzhen Zheng, Sheng Wang, Sheng Xu, Irwin King, et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nature Biotechnology*, 43(6):983–995, 2025.

[17] Tymor Hamamsy, James T Morton, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Robert Blackwell, Charlie EM Strauss, Julia Koehler Leman, Kyunghyun Cho, and Richard Bonneau. Tm-vec: template modeling vectors for fast homology detection and alignment. *bioRxiv*, pages 2022–07, 2022.

[18] Yinuo Jiang, Bing He, Zihan Wu, Fang Wang, Tianxu Lv, Yuran Jia, Yu Zhao, Chenchen Qin, Huajun Chen, Qiang Zhang, et al. Scalable homology detection with erast. *Nature Biotechnology*, pages 1–9, 2026.

[19] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025.

[20] Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, et al. Ernie-rna: an rna language model with structure-enhanced representations. *Nature Communications*, 16(1):10076, 2025.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[22] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[23] Bartosz Adamczyk, Pawel Boinski, Marta Szachniuk, and Maciej Antczak. Rnasolo 2.0: multimodal database to study rnas, their structural families and intermolecular interfaces. *Journal of Molecular Biology*, page 169570, 2025.

[24] The RNAcentral Consortium. Rnacentral in 2026: genes and literature integration. *Nucleic Acids Research*, 54(D1):D303–D313, 2026.

[25] Maryam Haghani, Debswapna Bhattacharya, and TM Murali. Neffy: a versatile tool for computing the number of effective sequences. *Bioinformatics*, 42(6):btaf222, 2026.

[26] Caleb Weinreb, Adam J Riesselman, John B Ingraham, Torsten Gross, Chris Sander, and Debora S Marks. 3d rna and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975, 2016.

[27] Elena Rivas, Jody Clements, and Sean R Eddy. A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. *Nature methods*, 14(1):45–48, 2017.

[28] Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, et al. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, pages 2026–04, 2026.

[29] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[30] Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 17(1):442, 2025.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf