RSC Chemical Biology

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: S. Takahashi, M. Hamada, H. Tateishi-Karimata and N. Sugimoto, *RSC Chem. Biol.*, 2025, DOI: 10.1039/D5CB00105F.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/rsc-chembio

View Article Online
DOI: 10.1039/D5CB00105F

Fitness landscapes and thermodynamic approaches to development of nucleic acids enzymes: From classical methods to AI integration

Shuntaro Takahashi[1,2*], Michiaki Hamada[3,4], Hisae Tateishi-Karimata[1,2], and Naoki Sugimoto[1*]

[1]FIBER (Frontier Institute for Biomolecular Engineering Research), Konan University, 7-1-20 Minatojima-Minamimachi, Chuo-ku, Kobe 650-0047, Japan
[2]FIRST (Graduate School of Frontiers of Innovative Research in Science and Technology), Konan University, 7-1-20 Minatojima-Minamimachi, Chuo-ku, Kobe 650-0047, Japan
[3]Department of Electrical Engineering and Bioscience, Graduate School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, 169-8555, Shinjuku-ku, Tokyo, Japan
[4]Cellular and Molecular Biotechnology Research Institute (CMB), National Institute of Advanced Industrial Science and Technology, 2-3-26 Aomi, Koto-ku, Tokyo 135-0064 Japan

Abstract

Nucleic acids (NA), namely DNA and RNA, dynamically fold and unfold to perform their functions in cells. Functional NAs include NA enzymes, such as ribozymes and DNAzymes. Their folding and target binding are governed by interactions between nucleobases, including base pairings, which follow thermodynamic principles. To elucidate biological mechanisms and enable diverse technical applications, it is essential to clarify the relationship between the primary sequence and the catalytic activity of NA enzymes. Unlike methods for predicting the stability of NA duplexes, which have been widely used for over half a century, predictive approaches for the catalytic activity of NA enzymes remain limited due to the low throughput of activity assays. However, recent advances in genome analysis and computational data science have significantly improved our understanding of the sequence-function relationship in NA enzymes. This article reviews the contributions of data-driven chemistry to understanding the reaction mechanisms of NA enzymes at the nucleotide level and predicting novel NA

View Article Online
DOI: 10.1039/D5CB00105F

enzymes with catalytic activity from sequence information. Furthermore, we discuss potential databases for predicting NA enzyme activity under various solution conditions and their integration with artificial intelligence for future applications.

1. Introduction

Program information is essential for designing reproducible systems and controlling their functions. In living cells, this information is held by nucleic acids (NA)—specifically, DNA and RNA—which are chemicals that store and transfer genetic information. Their nucleotide sequences are precisely replicated, transcribed, and translated to synthesise proteins that act as functional molecules. Genetic materials and their products are characterised by the fundamental rule that monomeric molecules, such as nucleotides and amino acids, are synthesised and function as one-dimensional macromolecules and units of information. From information decoding (via DNA transcription to RNA and RNA translation to protein) to functional expression (through protein folding), the transfer of genetic information relies on a series of molecular recognition events at the monomer level. The input and output of genetic information during transcription and translation follow the Watson–Crick base pairing rule, which unambiguously dictates how RNA and then protein are synthesised from a DNA sequence.[1] However, the functional expression of such information depends on the chemical properties of NAs as polymers. Since the formation and melting of secondary NA structures, including duplex association and dissociation, play key roles in regulating replication, transcription, and translation,[2-4] understanding their thermodynamic stability is crucial for programming and predicting their functions from sequences.

Proteins are the primary functional molecules in living organisms, but NAs themselves also exhibit functional activity. Recent advances, such as the Nobel Prize-winning methods for protein structure prediction, have made the design of functional proteins more feasible.[5] However, even with empirical knowledge of protein synthesis and folding, practical challenges persist in

View Article Online
DOI: 10.1039/D5CB00105F

deploying designed proteins for nanotechnology and medical applications.[6] On the other hand, NAs are easier to prepare chemically and biosynthetically than proteins, owing to their simple chemical composition and high water solubility. Therefore, research on their functional application in technology and medicine is highly active. Functional NAs were first discovered in self-splicing RNA for introns.[7, 8] Such catalytic RNA sequences called ribozymes are widespread across genomes and participate in biological processes such as tRNA processing,[9] rolling circle viral genome replication,[10] and peptide bond synthesis.[11] Various ribozyme families composed of different sequences and structures, such as hammerhead (HH),[12] hepatitis delta virus (HDV),[13] Varkud satellite (VS),[14] and hairpin,[15] have been isolated from cells and viruses. Moreover, bioinformatics approaches have contributed to the identification of other ribozyme families, including twister,[16] twister-sister,[17] hatchet,[17] and pistol[17]. HH, HDV, and twister family members have relatively small molecular sizes and high catalytic activities. Thus, these family members have frequently been used for *in vitro* and *in vivo* applications to generate RNAs with precise termini.[16, 18-20]

Each ribozyme has its own primary sequence, which folds into the secondary and tertiary structures that form its active site (Fig. 1a). Natural ribozymes mainly catalyse the cleavage of RNA strands (Fig. 1b). The catalytic activity of RNA led to the RNA world hypothesis,[21] which postulates that in prebiotic life, RNA not only replicates and transmits genetic information, but also catalyzes a number of metabolic reactions. *In vitro* selection techniques have generated ribozymes exhibiting various catalytic activities[22, 23] including phosphorylation, ligation, replication, aminoacylation, and Diels–Alder reactions.[24-27] Besides RNA, DNA has also been found to exhibit similar catalytic activity in the form of deoxyribozymes (DNAzymes), which act as catalysts for Pb^{2+}-dependent cleavage of RNA phophodiester bonds.[28] RNA-cleaving DNAzyme offers an attractive modality for targeting undruggable regions of the human genome,[29, 30] since the solid-phase chemical synthesis of DNA makes it a more inexpensibe and scalable material than RNA. Moreover, a their targeting sequences can be designed for any mRNA, enabling the identification of highly specific, low off-target candidates for safer nucleic acid therapeutics.[31, 32] NA enzymes combined with aptamers (having molecular recognition ability) and

View Article Online
DOI: 10.1039/D5CB00105F

complementary sequences (for DNA/RNA sequence recognition) have been actively investigated for applications such as biosensors and gene switches.[33-35] Thus, the development and use of NA enzymes, including ribozymes and DNAzymes, are of great interest in biotechnology.

NA enzymes exhibit enzymatic activity through the formation of secondary and tertiary structures from primary nucleotide sequences via intramolecular base pairing and other interactions. Therefore, understanding the correlation between the sequence and function of NA enzymes is of great importance from both fundamental and applied perspectives.[36] NAs, on the other hand, are subject to greater solution effects than proteins due to their nature as polyanions.[37] For example, the activity of NA enzymes is determined by their tertiary structure, accepting target molecules dominated by cations and interaction with cofactors such as divalent metal ions.[38] Notable targets of NA enzymes include DNA and RNA in cells, where the solution conditions are different from *in vitro* test tube conditions. The intracellular environment is both diverse and densely packed with biomolecules at concentrations ranging 50–400 g L^{-1}.[39] This molecular crowding profoundly affects nucelic acid conformation and stability.[40] As the effect of solution conditions on the structural stability and enzymatic activity of NA enzymes is highly complex, the systematic demonstration of sequences and environment-dependent reactions is required to determine which parts of the sequence are important for the function of an NA enzyme. Accumulation of enzymatic data creates a database of the properties of NA enzymes, which can provide information as parameters to predict how the sequences of these enzymes correlate with their catalytic activity. Therefore, progress in the analysis of large RNA and DNA datasets, together with computational approaches using machine learning and artificial intelligence (AI), is expected to advance the science and technology field considerably. In this review, we introduce such recent advances in the development of NA enzymes using data-driven analysis. Additionally, we analyse a novel dataset for determining the activities of these enzymes and discuss their utility for future predictions.

2. Development and improvement of NA enzymes using a large dataset

View Article Online
DOI: 10.1039/D5CB00105F

To determine the relationship between the sequence, reactant solution conditions, and activity of NA enzymes, a large dataset of enzyme activity derived from enzyme sequences is required. However, the analyses of enzymes with multiple sequences (mutations) usually rely on low-throughput experiments, including gel electrophoresis assays. To drastically expand the utility of NA enzymes, high-throughput assays must be developed to analyse their activity, sequence, and potential targets. Next-generation sequencing (NGS) is commonly used to analyse large numbers of sequences. NGS technology enables the simultaneous reading of millions of comparatively long NA sequences.[41, 42] NGS was originally developed for genome sequencing and has since been applied to the in vitro selection of NA aptamers and functional proteins that bind to or inhibit target proteins.[43, 44] NGS not only provides sequence data, but also the number of sequences from the read count, which enables the rapid analysis of sequence-function relationships in massive datasets.

The evolution process is propelled by mutations in genetic information. Consequently, the generated phenotype may fit more desired (or undesired) characteristics than the original. Thus, evolution toward the desired functions of NA enzymes relies on the “fitness” of the sequence and the efficiency of catalytic activity during each evolutional process.[45] NGS technology facilitates the visualization of these processes, aiding the understanding of how NA enzyme sequences achieve catalytic function from all possible sequences across the entire fitness landscape (Fig. 2).[46] The concept of a fitness landscape supports our understanding of evolutionary dynamics, describing a topography in which state variables such as genotype and phenotype are used as coordinates, and the height at each coordinate is the degree of adaptation.[47] When applied to NA enzymes, the fitness landscape contributes significantly to our understanding of the improvements in catalytic activity, as knowledge of this adaptive topography makes it possible to predict what evolutionary processes a sequence population will undergo.

The first study to create a fitness landscape for NA enzymes tested the class II ligase ribozyme.[48] The mutated ribozyme pool, containing $>10^{13}$ sequences via error-prone replication, was incubated with the immobilized substrate strand to select the active ribozymes. After single-round selection, NGS

View Article Online
DOI: 10.1039/D5CB00105F

data revealed the enrichment of the sequence containing a small Hamming distance, which is a measure of the difference between two strings of the same length compared to the original sequence (Fig. 3). Furthermore, the kinetically trapped incubation affected the enrichment of sequence reads of active ribozymes, which correlated with the observed catalytic rate constant (k_{obs}) obtained experimentally.[48] The large dataset of mutated ribozyme sequence and activity allowed the creation of a fitness landscape with >10^7 genotypes and phenotypes. It demonstrated the importance of sequences in the central bulge of the RNA and the distal end of paired region (helix) 3 (P3), along with other key residues characterised previously, in achieving maximal activity (Fig. 3).[49, 50] Thus, single or double mutations introduced through error-prone polymerase chain reaction or doped solid-phase synthesis enable the examination of fitness landscapes to identify the key sequence and structural determinants of NAzyme catalytic activity. This approach has been applied to various self-cleaving ribozymes owing to their small molecular sizes, which facilitate the efficient generation of mutants and direct assessment of mutational effects by reading enzyme activity.[51-55] For example, analysis of every single and double mutant of the *Osa-1-4* twister ribozyme from *Oryza sativa* (Fig. 4a) demonstrated its unexpected resilience to mutations, even with its compact and intricate structure. Notably, different structural components showed distinct levels of mutational sensitivity.[53] A recent comprehensive mutational study analysed five self-cleaving ribozymes, including CPEB3, HDV, hairpin, and hammerhead (Figs. 4b-e), in addition to twister ribozymes, providing strcutural information about the ribozymes, including their paired regions, unpaired loops, non-canonical structures, and tertiary structural contacts.[55] Additionally, NGS technology was used to study ribozyme evolution from random sequence and random structure space. In the case of Diels–Alderase ribozymes, increasing the selection pressure and analyzing the secondary structure through MFold prediction provided insights into how mutations can be rationally introduced to improve catalytic activity.[56] The NGS approach can be adopted not only for ribozymes but also for DNAzymes.[57] In a high-throughput kinetic analysis, 4,096 DNAzyme reactions were assayed simultaneously at multiple time points to determine the observed rate constants (k_{obs}) of 533 active mutants. These values were then

View Article Online
DOI: 10.1039/D5CB00105F

used to calculate activation energies (E_a), offering detailed insights into the mutational landscape of the DNAzymes. Deep sequencing enabled this quantitative view of the sequence–function relationship, which would not have been achievable with traditional assays.

The massive kinetic data on genotype and phenotype are also powerful tools for the analysis and development of NA enzymes triggered by ligand binding. One approach is to rationally develop an aptazyme, which is a combination of an aptamer and a self-cleaving ribozyme, to regulate translation by mRNA cleavage.[51, 58] In a previous study, all pairwise mutations in the *glmS* ribozyme triggered by glucosamine 6-phosphate (GlcN6P) were analysed using a custom-built fluorescent RNA array.[54] This array was derived using a combined approach involving ribozyme transcription on a sequencing tip and direct measurement of single-molecule fluorescence (detected using a total internal reflection fluorescence microscope). The advantage of this approach is its ability to monitor self-cleavage over short and long timescales, which enables the differentiation of both slow and fast self-cleaving variants. More recently, a kinetic sequencing (*k*-seq) technique was developed to perform a more accurate kinetic analysis of ribozymes using NGS.[59, 60] This technique provided the rate constants and maximum amplitude of the reaction without specialised instrumentation. The *k*-seq technique has been used to study the fitness landscape of self-aminoacylating and *glmS* ribozymes.[60, 61] By leveraging this approach, it is possible to systematically explore how biochemical factors, such as catalytic efficiency and the Michaelis constant, influence sequence conservation, even among partially active or inactive NA variants. As mentioned earlier, massive amounts of sequence data can be linked to each kinetic parameter through recent technological advances. The activity parameters describe the chemical mechanisms of the reactions, depending on the sequence and structure of the NA enzymes, with nucleotide-level resolution. Thus, if a sequence of NA enzymes responds well, comparing it (without additional experimentation) with the response of a mutant can provide insights into the mechanism of the NA enzyme and indicate the presence of novel structural rearrangements.

In contrast to the *in vitro* evolution of NA enzymes, naturally evolved NA enzymes, in which the full structural diversity is observed in many classes of

View Article Online
DOI: 10.1039/D5CB00105F

ribozymes found in nature, have also been targeted. Using massively parallel oligonucleotide synthesis, a diverse RNA pool was generated, enabling the direct functional testing of potential twister ribozyme sequences. This included over 1,600 previously reported putative twisters and approximately 1,000 new candidates derived from over a thousand different organisms.[62, 63] The Cleavage High-Throughput Assay, an NGS-based method for evaluating the activity of each potential sequence, revealed a broad structural tolerance to mutations.[64] These data about the relationships between the sequence diversity and activity of the twister ribozymes could advance the computational search of the active twister ribozymes, which identified the first intrinsically active twister ribozyme in mammals.[64] Although current studies are primarily focused on NA enzymes involved in cleavage, high throughput analysis has enabled the creation of large datasets for designing NA enzymes with other activities based on sequence information.

3. Prediction tool for NA enzyme sequence and function using a computational approach

One of the goals of studying the sequence-structure-function relationship of NA enzymes is to develop a prediction tool for determining the catalytic activity of NA enzymes from their sequence information. Classical bioinformatics approaches have developed ribozyme sequence finders from natural sequences. Conventional identification of ribozymes is perfomed using the Basic Local Alignment Search Tool, which detects sequence homology.[65] However, tertiary structures of the enzymes, which play a key role in their catalytic activity, must be considered for more accurate identification. Therefore, structure-based search programs are better suited for the identification of ribozymes. Computational tools such as RNAMotif,[66] Infernal,[67] and RNArobo[68] employ structural models of RNA secondary structures to systematically search for various ribozyme classes across sequence databases. These search programs have been successfully used to identify putative ribozymes in nature.[16, 69]

3.1 Prediction approach based on energetic calculation for secondary

View Article Online
DOI: 10.1039/D5CB00105F

structures of DNA and RNA

One of the major approaches to secondary (or even tertiary) structure prediction depends on the concept that a biomolecule folds into a certain structure with a minimum energy level based on its sequence information. Thus, the most stable structure can be inferred computationally by comparing the energy levels of all potential configurations. The most widely used prediction method based on the thermodynamic parameters of DNA and RNA stems (duplexes) derived from sequence information is the “nearest-neighbour (NN) model.” The NN model operates on the concept that the stability of a base pair is determined by its interactions with neighbouring base pairs. This model, developed by Tinoco Jr. et al., is extensively used to predict the thermostability of Watson–Crick duplexes, assuming that the duplexes exhibit two-state melting behavior.[70-72] Currently, the thermodynamic parameters for various duplexes—such as DNA/DNA, RNA/RNA, and RNA/DNA hybrids—have been established by Turner et al., along with contributions from other researchers, including our team.[73-75] In addition to the stems, the energetic penalties associated with mismatches, bulges, loops, dangling ends, and other motifs have been empirically determined.[76-95] With these expanded NN parameters, we can computationally identify the minimum energy structure from the possible conformations. MFold (UNAfold) remains the pioneering and most commonly used method for predicting DNA/ RNA secondary structures.[96-98] The algorithm evaluates all possible base pair combinations in the single-stranded sequence to determine the minimum free energy (MFE), ΔG_{min}, for the secondary structure. The sequence is represented as r_1, r_2, …, r_i, …, r_j, …, r_n, where r_i and r_j correspond to the ith and jth nucleotides ($1 \le i < j \le n$). The method then generates an "energy dot plot" where base pairs (r_i•r_j) within $\Delta\Delta G$ of ΔG_{min}, are plotted on a triangular grid (Fig. 5). The free energy increase, $\Delta\Delta G$, is determined by P/100 × $|\Delta G|$, where P is the "percent sub-optimality," which permits sub-optimal folding. The optimal folding with ΔG_{min} or sub-optimal folding with ΔG_{min} + $\Delta\Delta G$ is generated by selecting optimal or sub-optimal base pairs and computing their corresponding folding configurations.

Loops and bulges contribute destabilising or stabilising factors that influence overall structural stability. Accounting for these energy contributions

View Article Online
DOI: 10.1039/D5CB00105F

enhances the reliability of prediction systems.[88] Despite its widespread use, the average accuracy of secondary structure prediction is 73% (± 9%) for known canonical base pairs, indicating that the MFE approach alone has limited effectiveness in revealing true structures under different conditions. This can be attributed to the inability of the method to fully account for solution conditions, tertiary interactions, and protein binding effects. Therefore, the RNAstructure tool was developed for more reliable secondary structure predictions,[99, 100] integrating prediction constraints derived from experimental data—including Selective 2′ -Hydroxyl Acylation Analysed by Primer Extension (SHAPE), enzymatic cleavage, and chemical modification accessibility.[101]

Although the MFE approach is based on a simple concept, various structural combinations can be generated from a single DNA or RNA sequence, resulting in the formation of several structural motifs. Thus, predicting the secondary structure of relatively long chains is difficult. In particular, NA enzymes require the formation of a tertiary structure, which makes prediction complex. Machine learning and artificial intelligence (AI) tools have performed well in addressing these issues.[102] Besides MFE-based methods, non-MFE approaches relying on the centroid or maximum expected accuracy have also been developed.[103-105] With over 100,000 RNA sequences now available in databases,[106] the SPOT-RNA method, which is one of the representative predictions based on an advanced deep-learning technique, has been created to predict RNA secondary structures with exceptional accuracy.[107] E2Efold is an end-to-end deep learning model to directly predict the RNA base-pairing matrix for the RNA secondary structure prediction.[108] Ufold represents base matching on RNA sequences as "pseudo-image information" in a two-dimensional matrix.[109] However, the non-MFE approach often causes overfitting of machine learning by rich parameterisation.[110] These disadvantages have been minimised using a combination of MFE and non-MFE approaches to achieve greater prediction accuracy. LinearFold is based on a beam search teachnique to apply this algorithm to both machine-learned and Turner's thermodynamic models, resulting in fast and accurate prediction of the secondary structure from a long RNA strand.[111, 112] MXFold and MXFold2 generate folding scores, which are

View Article Online
DOI: 10.1039/D5CB00105F

calculated using a deep neural network incorporating Turner's NN free energy parameters.[113, 114] Therefore, these studies strongly suggest that incorporating thermodynamic information can enhance the robustness of deep learning-based RNA secondary structure predictions. The representative methods for prediction of RNA secondary structure are listed in Table 1.

3.2 Tools for the design of NA enzymes

The prediction of secondary structures provides valuable information for predicting ribozyme structure and function. Tools for the rational design of NA enzyme sequences have become a necessity, owing to the promising therapeutic applications of trans-acting ribozymes and DNAzymes for the cleavage of the target RNAs. Various computational tools for the design of NA enzymes have been developed up to now and surmarised in Table 2. For the catalytically active structure of NA enzyme and the substrate target, the energetic contribution of the formation of stems between NA enzyme and the substrate target is fundamental as well as the dissociation of each structured state (Fig. 6). Thus, the approach based on the MFE concept has been applied to identigy accessible target sites of NA enzyme using the tools of the secondary structure prediction.[115-119] For example, Sfold tool[120] was used to compute ΔG_{total} through consideration of the energy gain due to the complete intermolecular hybridization and the energy costs owing to structure alterations for both the target and the ribozyme calculated by $\Delta G_{total} = \Delta G_{hybrid} - \Delta G_{switch} - \Delta G_{disruption}$ (Fig. 6).[118] Another classical approach is called Aladdin,[119] which searches the optimized target sequences from the melting temperature of right and left arms of the HH ribozyme evaluated by MFold.[96] Although these tools are simple and useful, it does not contain off-target specificity analysis to find out the non-specific binding to the target region. RiboSoft is the first automated tool to design the HH ribozyme with consideration of the off-target effect, which can output some potential ribozyme sequences including the active one for silencing a disease-causing gene.[121] This tool is currently updated as RiboSoft 2.0 which can design different types of trans-acting conventional and allosteric ribozymes.[122]

In such applications, a trans-acting DNAzyme also fascinating owing to the higher chemical stability and easier chemical synthesis than RNA. DNAzyme

View Article Online
DOI: 10.1039/D5CB00105F

binds to the target RNA by forming stems as well as ribozyme, which are governed by thermodynamics, depending on the base components. DNAmoreDB and DNAzymeBuilder are pioneering web tools that can be used to design DNAzymes for any target sequence.[123, 124] DNAmoreDB is a comprehensive resource for DNAzymes that organises information such as sequences, selection conditions, catalysed reactions, kinetic parameters, substrates, cofactors, structural data (when available), and literature. The DNAzymeBuilder database includes the details of 44 RNA-cleaving and 93 DNA-cleaving DNAzymes, including those with RNA-like rA at the cleavage site, all of which function in a trans-cleavage manner and cleave intermolecular substrates. This internal database compiles extensive data on DNAzymes, including optimal reaction conditions, kinetic properties, types of catalysed reactions, sequence recognition, cleavage sites, and the necessary design elements to ensure optimal DNAzyme performance. Thus, predicted information on the target site, DNAzyme sequence, and catalytic activity can be obtained. To further enhance the prediction of DNAzyme activity, a machine learning approach was employed. This approach was used to identify DNAzymes capable of efficiently triaging thousands of potential molecules specific to a target RNA.[125] Based on logistic regression, the developpers trained the model on published and newly generated 10-23 DNAzyme activity data incorporating (1) the energetic parameters of the enzyme/target stems and (2) the DNAzyme secondary structure derived from NN parameters of RNA/DNA hybrids,[75] obtained using the secondary structure prediction tool.[101, 126] The analysis revealed that the binding free energy between the DNAzyme and its RNA target is the key factor influencing efficiency. However, other elements, such as the internal structure of the DNAzyme, also play a crucial role in determining its catalytic activity.[125] The machine learning approach is also trained the established database from DNAmoreDB, which is called SequenceCraft.[127] This approach was trained with the k_{obs} data from 178 RNA-cleaving DNAzymes together with varying experimental conditions, including cofactor type and concentration, pH, and temperature. In this platform, a dot-bracket notation of secondary structures calculated using MFold was used to generate a numerical vector, which ensures the good prediction accuracy of the k_{obs} values.

View Article Online
DOI: 10.1039/D5CB00105F

As shown earlier, machine learning and AI technologies have advanced and have been applied to the prediction of RNA secondary structures.[102] Moreover, creating a large dataset of NA enzymes can provide sufficient teaching data for AI to output accurate predictions (Fig. 2). One approach is to study the NGS data obtained from massive mutational analyses of NA enzymes to generate a fitness landscape. The fitness landscape can provide valuable information not only for elucidating the evolutionary process of ribozymes, but also for the rational design of novel enzymes. For instance, the AI technique involving NGS-based high-throughput data enabled the understanding of the F1*U ribozyme neutral network.[128] In this study, experimental evaluation of over 120,000 ribozyme sequences provided valuable empirical evidence that neutral networks can enhance the accessibility and predictability of the fitness landscape. In another study, the effects of higher-order mutations on the CPEB-3 ribozyme were also reported.[129]

Inverse folding has been studied for over a decade in the design of NA enzymes. RNAiFold was used as an example to design ten artificial cis-cleaving HH ribozymes by identifying RNA sequences whose MFE secondary structure corresponded to a user-defined target.[130] Each of the ribozymes demonstrated functionality in a cleavage assay. However, this method has some challenges in terms of accuracy and versatility because of the difficulty in predicting tertiary interactions of nucleotides. Therefore, advanced computational approaches are required. One of these approaches, a deep generative model, which has already been applied to protein design,[131] is an attractive pipeline for generating novel designs for NA enzymes. Recently, the world's first deep generative model for NA enzymes, RfamGen, was developed to support the design of artificial RNAs with desired functions and structures.[132] RfamGen combines a variational autoencoder, a method widely used in deep generative modelling, and a covariance model that can classify functional RNAs from information on RNA sequences and secondary structures. These features can be learned, and artificial sequences can be generated. Computer analysis and biochemical experiments confirmed that RfamGen could stably generate RNA sequences with a structure and function homologous to the learned RNA population. The performance of RfamGen was also attributed to the application of a covariance

View Article Online
DOI: 10.1039/D5CB00105F

model to a deep generative model. RfamGen was employed to generate 1,000 new sequences using the *glmS* ribozyme that cleaves its own RNA sequence by binding to small molecules. A comprehensive analysis of the generated RNA sequences was conducted on a large scale.[132] Interestingly, RfamGen showed a greater tendency to generate high-activity enzyme sequences than native sequences.[132] Therefore, recent advances in computational approaches using both machine learning and AI could demonstrate predictions for the development and generation of novel NA enzymes.

Recent AI-based predictions target not only RNA-RNA interaction but also RNA-protein interaction.[133] Moreover, a predictive tool for tertiary structures of NA enzymes has also been developed recently. Similar to how AlphaFold predicts protein structure from sequence information,[134] the machine learning and deep learning approach such as RhoFold+, RNA-Puzzles and trRosettaRNA can be used to develop a prediction tool for NA structures, including NA enzymes.[135-137] However, one issue is that the number of solved tertiary structures of NAs is small for machine learning and deep learning approaches, in contrast with the extensive datasets available for proteins. However, the simpler chemical characteristics of NAs compared to proteins may overcome this issue, allowing computational approaches to provide the necessary structural information.

4. New data base for prediction pipelines of NA zymes structure and activity

As mentioned earlier, the energetic calculation of the secondary structure is a fundamental process to predict the structure and activity of NA enzymes. Thermodynamically, six factors predominantly affect duplex stability (Fig. 7) [40] by governing base pair formation: (i) hydrogen bonds between base pairs—in Watson-Crick base pairs, the alignment of hydrogen bonding donors and acceptors contributes to their high selectivity and stability. (ii) Stacking interactions between adjacent base pairs—the aromatic rings of the bases, being electron-rich, engage in π stacking interactions with other aromatic rings. (iii) Conformational entropy—the transition from a single-stranded random coil structure to a helix results in an energetic penalty. (iv) Cation condensation—in addition to the structural factors, environmental factors should play a key role in determining the stability and structure of NAs. For instance, the influence of

View Article Online
DOI: 10.1039/D5CB00105F

cations on duplex formation is explained by the theory of counterion condensation. Polyanionic NA strands experience significant electrostatic repulsion, which makes their binding unfavourable. Cations neutralise the anionic charges, thereby promoting duplex formation. (v) Hydration—another important factor in duplex formation is the water molecules, which specifically hydrate the grooves and backbones of the structure. The extent of hydration depends on the structural conformation. (vi) Molecular crowding—crowders significantly affect the stability of NA structures. Moreover, the behaviour of NA stability and structure can be affected by the other physicochemical property of the solution such as the dieletric constant, viscosity, and so on. Because the conditions under which NA enzymes work differ between environments, such as test tubes and cells, these factors should be carefully considered.

To predict the NA enzyme activity in various conditions, the effect of the environment on the energetic contribution to the NA enzyme is also fundamental. As shown in Fig. 6, the targeting function via the duplex formation is driven by cation concentration, hydration and molecular crowding, whereas the catalysis function is affected by molecular crowding for the formation of the tertiary structure as well as the binding of the cofactor of the metal ion by the dielectric constant changes. In the classical secondary structure prediction of NAs based on the MFE approach, cation concentration corrections have been widely applied using improved NN parameters.[73, 138-140] For example, these corrections enabled the MFold database to predict structures at arbitrary NaCl concentrations.[96] However, current tools rely on Turner's NN parameters with such corrections. Therefore, the effects of hydration and crowding on NA stability have not yet been considered. Solutions under cellular conditions are densely packed with biomacromolecules such as proteins and NAs. Large numbers of small molecules, such as metabolites and metal ions, are also present. Hence, incorporating the new NN parameter datasets can expand the feasability of the MFE approach for predicting NA stability and structure, especially under cellular conditions. The total concentration of macromolecules has been estimated to reach 400 mg mL^{-1}, occupying approximately 40% of the intracellular space.[141] These *in vivo* crowded conditions are extremely different from the diluted conditions of standard *in vitro* systems. Therefore, it is important to understand the physicochemical

View Article Online
DOI: 10.1039/D5CB00105F

properties of NAs under molecular crowding. NA foldings and unfoldings occur in equilibrium, and are accompanied by structural changes and water interactions (Fig. 8a). The biophysical effects of molecular crowding are mostly based on the physicochemical properties of the crowders, which affect the volume and hydration effects of NA folding processes (Figs. 8b and 8c). The formation of a duplex makes the strand volume compact; a large cosolute tends to stabilise the duplex, whereas a small cosolute destroys it by effectively decreasing the water activity of the solution (as duplex formation accompanies hydration).[142] For NA enzyme reactions, the effect of crowders on the dielectric constant and viscosity plays an important role in the reaction kinetics.[143] Moreover, crowders can interact directly with NAs to stabilise the structure via CH-π interactions.[144] Therefore, the behaviour of NA structures is influenced by these biophysical factors under molecular crowding conditions.[145, 146]

To consider the effect of molecular crowding on dupex stability, the NN parameters for DNA duplexes (including self- and non-self- complementary strands) have been determined for buffers containing 100 mM NaCl with 40 wt% polyethylene glycol 200 (PEG200) (Table 3).[147] Compared to NN parameters under non-crowding condition, molecular crowding exhibited different effects on each NN parameter. Moreover, the relative destabilisation of NN with only GC pairs—d(CG/GC), d(GC/CG), and d(GG/CC)—was considerably larger than that of other NN pairs. This may be attributed to the low water activity caused by PEG200, as GC pairs require more water molecules for stabilisation compared to AT pairs.[148] The most remarkable difference was found among the initiation factors. The $\Delta H°$ and $\Delta S°$ corresponding to duplex initiation differed drastically under crowding conditions compared to that in the solution without cosolutes. This was because of the preferential hydration of terminal oligonucleotide pairs induced by the cosolute in the crowded environment.[147] Although this is the first report of NN parameters under crowding conditions, their application should not be limited to specific environments. For example, NN parameters under crowding have been only determined under 40% PEG condition with 100 mM NaCl for the DNA/DNA duplex[149] and under 20% PEG200 condition with 1 M NaCl for the RNA/RNA duplex.[150] Parameters for RNA/RNA duplexes in the Eco80 artificial cytoplasm, which contains 80% of *Escherichia coli* metabolites and biological

View Article Online
DOI: 10.1039/D5CB00105F

concentrations of metal ions, have also reported.[151] To generalise the available NN parameters for various cation and crowding conditions, each value of $\Delta G^\circ_{37,NN}$ for a duplex was considered as the sum of contributions from the bulk structure, cations, and crowders ($\Delta G^\circ_{37\ NN} = \Delta G^\circ_{37\ NN\ [bulk]} + \Delta G^\circ_{37\ NN\ [cation]} + \Delta G^\circ_{37\ NN\ [crowder]}$). The $\Delta G^\circ_{37\ NN\ [cation]}$ parameters at arbitrary concentrations of NaCl can be calculated from those measured at 1 M NaCl by applying the known dependence of [Na^+] for each NN base pair and regarding $\Delta G^\circ_{37\ NN\ [bulk]}$ as the value at 0 M [Na^+].[152] Figure 9 illustrates the scheme for obtaining NN parameters under the desired solution conditions. The $\Delta G^\circ_{37\ NN[crowder]}$ parameters can be determined from the linear function of changes in water activity Δa_w, as duplex (de)stabilisation ($\Delta\Delta G^\circ_{37}$) in the presence of crowders correlated linearly with changes in the excluded volume of cosolutes and water activity.[153] Based on this strategy, the improved NN parameters for any molecular environment can be obtained,[147] This approach of adjusting NN parameters according to cation and crowder conditions has been successfully applied to RNA/RNA and RNA/DNA duplexes.[154, 155] Table 3 shows the NN parameters of DNA/DNA, RNA/RNA, and RNA/DNA duplexes under 100 mM NaCl and 40 wt% PEG200 conditions. The parameters generally applied to different solutions for RNA/RNA duplexes are listed in Table 4. Thus, the latest NN parameters can be regarded as universal for predicting duplexes under arbitrary solution conditions. Examples of some predictions under different salt and crowding conditions are listed in Table 5.

Considering that the nucleolar environment is similar to that of PEG200,[156] the stability of the DNA duplex in Ddx4 liquid-liquid phase separation (LLPS)[157] was successfully predicted using the universal parameters derived from 50% PEG200 with 0.1 M NaCl conditions.[147] This approach clarifies that the nucleolar condition can be mimicked by the crowding conditions, allowing the investigation of various NA behaviours in the nucleolus using controlled solution conditions. These parameters could also accurately predict the stability of the RNA hairpin in the nucleus and cytosol,[154] and the efficiency of gene editing by CRISPR/Cas9.[155] Moreover, improvements in predicting the stability of GC- and AT-biased DNA duplexes enabled the prediction of the efficiency of G-quadruplex formations from GC-rich sequences, as well as the identification of the replication initiation region in genomic DNAs.[158] These approaches, which mimick local

View Article Online
DOI: 10.1039/D5CB00105F

cellular environments, can be a novel platform to assess the behaviour of NAs in such localised areas. For example, the mitochondrial environment in human cells induces G4 formation owing to the highly crowded conditions compared to the nucleus,[159] which can be regarded as a 60 wt% 1,3-propanediol (1,3 PDO) solution.[160] These findings indicate that the classical NN parameters obtained using 1 M NaCl solutions are not suitable for the prediction of either duplex stability or NA function based on duplex formations. Therefore, the newly obtained NN parameters can be key components for predicting and developing functional NA enzymes in specific environments, particularly under intracellular conditions.

In addition to the solution environment, compartmentalization also affects the solution conditions and NA behaviour. Since cellular environments are composed of lipid compartments, their effect on the structural stability of NAs and activity of NA enzymes should be critical from an evolutionary viewpoint, especially as a protocell model. The effects of the compartments on NA behaviour have been studied using reverse micelles and liposomes. Reverse micelles can create a nano-confinement space of variable size by changing the ratio of surfactants such as sodium bis(2-ethylhexyl)sulfosuccinate. These conditions in reverse micelles efficiently decrease DNA duplex stability.[161] Interestingly, non-duplex structures such as G-quadruplex and i-motif formations are promoted in the nano-confinement space by the reverse micelles.[162, 163] These findings indicate that solution compartments within the nanometre size range alter the solution environment, similar to molecular crowding, and cause the destabilisation of duplexes and stabilisation of non-duplexes. For NA enzyme activity, the excluded volume effect caused by compartmentalisation significantly promotes reaction activity. In one case, the reaction kinetics and conformational folding of hairpin ribozymes within a liposome were investigated.[164] The conditions inside the liposome (100 nm in diameter), prepared from a 1:1 mixture of oleic acid and 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine, enhanced both intermolecular and intramolecular RNA interactions. Simultaneously, it promoted the proper folding of tertiary structures, including the docked conformation of the active hairpin ribozyme and its characteristic triplex arrangement. Moreover, the misfolding rate of the active structure was reduced, contributing to the promotion

View Article Online
DOI: 10.1039/D5CB00105F

of ribozyme activity. A similar phenomenon was observed in the case of a self-aminoacylating ribozyme.[165]

Membrane-less compartments formed by LLPS also provide a confinement environment for NA enzymes. The Ddx4 LLPS decreases the DNA and RNA duplexes similar to PEG200-based *in vitro* crowding conditions.[157] Cationic polymers or peptides are used to form LLPS with HH, hairpin, R3C ligase, and 10–23 DNAzyme ribozymes, all of which are activated, unlike in bulk solutions.[166, 167] The unique LLPS conditions can simulate the unique ionic conditions required for ribozyme activity by concentrating ions such as Mg^{2+}s.[168] Although the physicochemical properties of LLPS are not clearly defined, ribozyme activation have been recreated in molecular crowding conditions.[169-171] Thus, future parameterization of solution properties mimicked by *in vitro* crowding conditions can provide a useful index to predict the stability and function of NA enzymes in the intracellular membrane or membrane-less organelles, including nucleus, mitochondria, nucleolus, and stress granules. For the application of NA enzymes in cells, their chemical modification is necessary to avoid degradation. As chemical modifications can affect both the stability and tertiary structure of NAs,[172, 173] a comprehensive analysis of the effects of chemical modifications on will also impact the development and functional prediction of NA enzymes.

Conclusion and perspective

In this article, we summarise the importance of thermodynamic parameters and future perspectives on NA enzyme prediction and development. The activity of an NA enzyme involves the folding of its structure, interaction with target NAs, and catalytic chemistry. A major milestone in NA enzyme technology is the development of tools for predicting enzymatic activity and structure from sequence information. In the case of proteins, an AI-based structural prediction program named AlphaFold has been developed and widely used, providing significant progress in the scientific and application fields. Recent advances in both AI-based and energy model-based approaches have extended beyond single RNA structure prediction to include interactions between RNA and other molecules, such as proteins and other RNAs. These interaction predictions are

View Article Online
DOI: 10.1039/D5CB00105F

particularly promising for the in vivo design of nucleic acid enzymes. Furthermore, the integration of AI with genomic functional data—such as sequence variants, regulatory elements, and epigenomic features—holds great potential to further enhance RNA design by enabling context-aware predictions tailored to cellular environments.

As computational structural predictions advance in the field of NAs, an accurate prediction model for NAs, similar to AlphaFold, may be established in the near future. However, two key issues, which the alphaFold algorithm does not account for, remain: (1) prediction of the *de novo* structure from primary sequence information and, more importantly, (2) effect of the molecular environment on structure formation. Since, NAs fold dynamically to form tertiary structures from single strands and are more sensitive to the environment than proteins, basic energetic information on how the environment affects their base pairing is required for accurately predicting the structure and activity of NA enzymes. While computational approaches, including AI and machine learning techniques, have progressed rapidly, the necessary experimentally obtained fundamental databases have not been adequately collected. However, as reviewed in this article, the key factors determining the folding and activity of NA enzymes at the sequence level have now been identified, based on the accumulation of chemical properties of NAs affecting their thermodynamics (Fig. 10). The elucidation of the underlying chemistry, driven by a large dataset collected under various critical situations, would be useful for designing AI and machine learning techniques to solve the structure and activity of NA enzymes within specific environments. Moreover, the *de novo* generation of NA enzymes that function actively in targeted environments could be realized without requiring a massive database of experimentally solved tertiary nucleic acid structures.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Conflicts of interest

View Article Online
DOI: 10.1039/D5CB00105F

The authors declare no competing interests.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS) (21K05283 and 24K01631), especially the Grant-in-Aid for Scientific Research (S) (22H04975), JSPS Core-to-Core Program (JPJSCCA20220005), Konan New Century Strategic Research Project, Asahi Glass Foundation, and Chubei Itoh Foundation.

References

1. F. H. Crick, *Symp. Soc. Exp. Biol.*, 1958, **12**, 8.
2. T. D. Yager and P. H. von Hippel, *Biochemistry*, 1991, **30**, 1097-1118.
3. J. A. Morin, F. J. Cao, J. M. Lázaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas and B. Ibarra, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 8115-8120.
4. S. E. Kolitz, J. E. Takacs and J. R. Lorsch, *RNA*, 2009, **15**, 138-152.
5. P. A. Vargas-Rosales and A. Caflisch, *RSC Med Chem*, 2025, DOI: 10.1039/d4md00869c.
6. R. Qing, S. Hao, E. Smorodina, D. Jin, A. Zalevsky and S. Zhang, *Chem. Rev.*, 2022, **122**, 14085-14179.
7. K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling and T. R. Cech, *Cell*, 1982, **31**, 147-157.
8. T. R. Cech, A. J. Zaug and P. J. Grabowski, *Cell*, 1981, **27**, 487-496.
9. C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace and S. Altman, *Cell*, 1983, **35**, 849-857.
10. E. L. Kristoffersen, M. Burman, A. Noy and P. Holliger, *eLife*, 2022, **11**, e75186.
11. P. Nissen, J. Hansen, N. Ban, P. B. Moore and T. A. Steitz, *Science*, 2000, **289**, 920-930.
12. A. C. Forster and R. H. Symons, *Cell*, 1987, **49**, 211-220.
13. M. Y. Kuo, L. Sharmeen, G. Dinter-Gottlieb and J. Taylor, *J. Virol.*, 1988, **62**, 4439-4444.

View Article Online
DOI: 10.1039/D5CB00105F

14. B. J. Saville and R. A. Collins, *Cell*, 1990, **61**, 685-696.
15. P. A. Feldstein, J. M. Buzayan and G. Bruening, *Gene*, 1989, **82**, 53-61.
16. A. Roth, Z. Weinberg, A. G. Chen, P. B. Kim, T. D. Ames and R. R. Breaker, *Nat Chem Biol*, 2014, **10**, 56-60.
17. Z. Weinberg, P. B. Kim, T. H. Chen, S. Li, K. A. Harris, C. E. Lünse and R. R. Breaker, *Nat Chem Biol*, 2015, **11**, 606-610.
18. A. R. Ferré-D'Amaré and J. A. Doudna, *Nucleic Acids Res.*, 1996, **24**, 977-978.
19. J. M. Avis, G. L. Conn and S. C. Walker, *Methods Mol. Biol.*, 2012, **941**, 83-98.
20. T. Zhang, Y. Gao, R. Wang and Y. Zhao, *Bio Protoc*, 2017, **7**, e2148.
21. W. Gilbert, *Nature*, 1986, **319**, 618-618.
22. C. Tuerk and L. Gold, *Science*, 1990, **249**, 505-510.
23. A. D. Ellington and J. W. Szostak, *Nature*, 1990, **346**, 818-822.
24. A. Jäschke, *Curr. Opin. Struct. Biol.*, 2001, **11**, 321-326.
25. T. J. Wilson and D. M. J. Lilley, *Wiley Interdiscip Rev RNA*, 2021, **12**, e1651.
26. M. J. Fedor and J. R. Williamson, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 399-412.
27. R. Micura and C. Höbartner, *Chemical Society Reviews*, 2020, **49**, 7331-7353.
28. R. R. Breaker and G. F. Joyce, *Chem. Biol.*, 1994, **1**, 223-229.
29. S. W. Santoro and G. F. Joyce, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 4262-4266.
30. S. W. Santoro and G. F. Joyce, *Biochemistry*, 1998, **37**, 13330-13342.
31. G. F. Joyce, *Methods Enzymol.*, 2001, **341**, 503-517.
32. A. A. Fokina, D. A. Stetsenko and J. C. François, *Expert Opin Biol Ther*, 2015, **15**, 689-711.
33. R. R. Breaker, *Curr. Opin. Biotechnol.*, 2002, **13**, 31-39.
34. S. K. Silverman, *Trends Biochem. Sci.*, 2016, **41**, 595-609.
35. A. Serganov and D. J. Patel, *Nat Rev Genet*, 2007, **8**, 776-790.
36. D. A. Lafontaine, D. G. Norman and D. M. Lilley, *EMBO J.*, 2001, **20**, 1415-1424.
37. D. M. Crothers, V. A. Bloomfield and I. Tinoco, *Nucleic acids: structures, properties, and functions*, University science books, 2000.
38. H. Saito and H. Suga, *Nucleic Acids Res.*, 2002, **30**, 5151-5159.
39. R. J. Ellis, *Curr. Opin. Struct. Biol.*, 2001, **11**, 114-119.
40. S. Takahashi and N. Sugimoto, *Chem. Soc. Rev.*, 2020, **49**, 8439-8468.
41. M. L. Metzker, *Nat. Rev. Genet.*, 2010, **11**, 31-46.

View Article Online
DOI: 10.1039/D5CB00105F

42. N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain and M. J. Pallen, *Nat. Biotechnol.*, 2012, **30**, 434-439.
43. M. Cho, Y. Xiao, J. Nie, R. Stewart, A. T. Csordas, S. S. Oh, J. A. Thomson and H. T. Soh, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 15373-15378.
44. M. A. Ditzler, M. J. Lange, D. Bose, C. A. Bottoms, K. F. Virkler, A. W. Sawyer, A. S. Whatley, W. Spollen, S. A. Givan and D. H. Burke, *Nucleic Acids Res.*, 2013, **41**, 1873-1884.
45. E. J. Hayden and A. Wagner, *Proc Biol Sci*, 2012, **279**, 3418-3425.
46. J. Maynard Smith, *Nature*, 1970, **225**, 563-564.
47. S. Kauffman and S. Levin, *J. Theor. Biol.*, 1987, **128**, 11-45.
48. J. N. Pitt and A. R. Ferré-D'Amaré, *Science*, 2010, **330**, 376-379.
49. E. H. Ekland, J. W. Szostak and D. P. Bartel, *Science*, 1995, **269**, 364-370.
50. J. N. Pitt and A. R. Ferré-D'Amaré, *J. Am. Chem. Soc.*, 2009, **131**, 3532-3540.
51. S. Kobori, Y. Nomura, A. Miu and Y. Yokobayashi, *Nucleic Acids Res.*, 2015, **43**, e85.
52. Y. Yokobayashi, *Acc. Chem. Res.*, 2020, **53**, 2903-2912.
53. S. Kobori and Y. Yokobayashi, *Angew. Chem. Int. Ed.*, 2016, **55**, 10354-10357.
54. J. O. L. Andreasson, A. Savinov, S. M. Block and W. J. Greenleaf, *Nat. Commun.*, 2020, **11**, 1663.
55. J. M. Roberts, J. D. Beck, T. B. Pollock, D. P. Bendixsen and E. J. Hayden, *eLife*, 2023, **12**, e80360.
56. S. Ameta, M. L. Winz, C. Previti and A. Jäschke, *Nucleic Acids Res.*, 2014, **42**, 1303-1310.
57. V. Dhamodharan, S. Kobori and Y. Yokobayashi, *ACS Chemical Biology*, 2017, **12**, 2940-2945.
58. Y. Nomura, H.-C. Chien and Y. Yokobayashi, *Chem. Commun.*, 2017, **53**, 12540-12543.
59. Y. Shen, A. Pressman, E. Janzen and I. A. Chen, *Nucleic Acids Res.*, 2021, **49**, e67.
60. A. D. Pressman, Z. Liu, E. Janzen, C. Blanco, U. F. Müller, G. F. Joyce, R. Pascal and I. A. Chen, *J. Am. Chem. Soc.*, 2019, **141**, 6213-6223.
61. L.-Eng D. Yu, Elise N. White and Sarah A. Woodson, *Nucleic Acids Res.*, 2024, **52**, 13340-13350.

View Article Online
DOI: 10.1039/D5CB00105F

62. M. A. Cleary, K. Kilian, Y. Wang, J. Bradshaw, G. Cavet, W. Ge, A. Kulkarni, P. J. Paddison, K. Chang, N. Sheth, E. Leproust, E. M. Coffey, J. Burchard, W. R. McCombie, P. Linsley and G. J. Hannon, *Nat. Methods*, 2004, **1**, 241-248.
63. S. Kosuri and G. M. Church, *Nat. Methods*, 2014, **11**, 499-507.
64. Lauren N. McKinley, McCauley O. Meyer, A. Sebastian, Benjamin K. Chang, Kyle J. Messina, I. Albert and Philip C. Bevilacqua, *Nucleic Acids Res.*, 2024, DOI: 10.1093/nar/gkae908.
65. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403-410.
66. T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case and R. Sampath, *Nucleic Acids Res.*, 2001, **29**, 4724-4735.
67. E. P. Nawrocki, D. L. Kolbe and S. R. Eddy, *Bioinformatics*, 2009, **25**, 1335-1337.
68. L. Rampášek, R. M. Jimenez, A. Lupták, T. Vinař and B. Brejová, *BMC Bioinformatics*, 2016, **17**, 216.
69. B. D. Lee, U. Neri, S. Roux, Y. I. Wolf, A. P. Camargo, M. Krupovic, P. Simmonds, N. Kyrpides, U. Gophna, V. V. Dolja and E. V. Koonin, *Cell*, 2023, **186**, 646-661.e644.
70. I. Tinoco, Jr., O. C. Uhlenbeck and M. D. Levine, *Nature*, 1971, **230**, 362-367.
71. P. N. Borer, B. Dengler, I. Tinoco, Jr. and O. C. Uhlenbeck, *J. Mol. Biol.*, 1974, **86**, 843-853.
72. M. Andronescu, A. Condon, D. H. Turner and D. H. Mathews, *Methods Mol. Biol.*, 2014, **1097**, 45-70.
73. J. SantaLucia, Jr., *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 1460-1465.
74. T. Xia, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner, *Biochemistry*, 1998, **37**, 14719-14735.
75. N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama and M. Sasaki, *Biochemistry*, 1995, **34**, 11211-11216.
76. S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson and D. H. Turner, *Proc. Natl. Acad. Sci. U. S. A.*, 1986, **83**, 9373-9377.
77. C. E. Longfellow, R. Kierzek and D. H. Turner, *Biochemistry*, 1990, **29**, 278-285.
78. D. R. Groebe and O. C. Uhlenbeck, *Biochemistry*, 1989, **28**, 742-747.
79. J. M. Blose, M. L. Manni, K. A. Klapec, Y. Stranger-Jones, A. C. Zyra, V. Sim, C. A. Griffith, J. D. Long and M. J. Serra, *Biochemistry*, 2007, **46**, 15123-15135.

View Article Online
DOI: 10.1039/D5CB00105F

80. H. T. Allawi and J. SantaLucia, Jr., *Biochemistry*, 1997, **36**, 10581-10594.
81. H. T. Allawi and J. SantaLucia, Jr., *Biochemistry*, 1998, **37**, 2170-2179.
82. H. T. Allawi and J. SantaLucia, Jr., *Nucleic Acids Res.*, 1998, **26**, 2694-2701.
83. H. T. Allawi and J. SantaLucia, Jr., *Biochemistry*, 1998, **37**, 9435-9444.
84. N. Peyret, P. A. Seneviratne, H. T. Allawi and J. SantaLucia, Jr., *Biochemistry*, 1999, **38**, 3468-3477.
85. L. He, R. Kierzek, J. SantaLucia, Jr., A. E. Walter and D. H. Turner, *Biochemistry*, 1991, **30**, 11124-11132.
86. J. SantaLucia, Jr., R. Kierzek and D. H. Turner, *Biochemistry*, 1991, **30**, 8242-8251.
87. M. Wu, J. A. McDowell and D. H. Turner, *Biochemistry*, 1995, **34**, 3204-3211.
88. D. H. Mathews, J. Sabina, M. Zuker and D. H. Turner, *J. Mol. Biol.*, 1999, **288**, 911-940.
89. S. M. Freier, R. Kierzek, M. H. Caruthers, T. Neilson and D. H. Turner, *Biochemistry*, 1986, **25**, 3209-3213.
90. D. R. Hickey and D. H. Turner, *Biochemistry*, 1985, **24**, 3987-3991.
91. N. Sugimoto, R. Kierzek and D. H. Turner, *Biochemistry*, 1987, **26**, 4559-4562.
92. M. J. Serra, T. J. Axenson and D. H. Turner, *Biochemistry*, 1994, **33**, 14289-14296.
93. S. M. Freier, D. Alkema, A. Sinclair, T. Neilson and D. H. Turner, *Biochemistry*, 1985, **24**, 4533-4539.
94. S. Bommarito, N. Peyret and J. SantaLucia, Jr., *Nucleic Acids Res.*, 2000, **28**, 1929-1934.
95. N. Sugimoto, R. Kierzek and D. H. Turner, *Biochemistry*, 1987, **26**, 4554-4558.
96. M. Zuker, *Nucleic Acids Res.*, 2003, **31**, 3406-3415.
97. M. Zuker and P. Stiegler, *Nucleic Acids Res.*, 1981, **9**, 133-148.
98. M. Zuker, *Science*, 1989, **244**, 48-52.
99. S. Bellaousov, J. S. Reuter, M. G. Seetin and D. H. Mathews, *Nucleic Acids Res.*, 2013, **41**, W471-474.
100. S. E. Ali, A. Mittal and D. H. Mathews, *Curr Protoc*, 2023, **3**, e846.
101. D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker and D. H. Turner, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 7287-7292.
102. K. Sato and M. Hamada, *Briefings in Bioinformatics*, 2023, **24**.

View Article Online
DOI: 10.1039/D5CB00105F

103. L. E. Carvalho and C. E. Lawrence, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 3209-3214.
104. C. B. Do, D. A. Woods and S. Batzoglou, *Bioinformatics*, 2006, **22**, e90-e98.
105. M. Hamada, H. Kiryu, K. Sato, T. Mituyama and K. Asai, *Bioinformatics*, 2008, **25**, 465-473.
106. P. Danaee, M. Rouches, M. Wiley, D. Deng, L. Huang and D. Hendrix, *Nucleic Acids Res.*, 2018, **46**, 5381-5394.
107. J. Singh, J. Hanson, K. Paliwal and Y. Zhou, *Nat. Commun.*, 2019, **10**, 5407.
108. X. Chen, Y. Li, R. Umarov, X. Gao and L. Song, *arXiv preprint arXiv:2002.05810*, 2020.
109. L. Fu, Y. Cao, J. Wu, Q. Peng, Q. Nie and X. Xie, *Nucleic Acids Res.*, 2022, **50**, e14-e14.
110. E. Rivas, R. Lang and S. R. Eddy, *RNA*, 2012, **18**, 193-212.
111. L. Huang, H. Zhang, D. Deng, K. Zhao, K. Liu, D. A. Hendrix and D. H. Mathews, *Bioinformatics*, 2019, **35**, i295-i304.
112. H. Zhang, L. Zhang, D. H. Mathews and L. Huang, *Bioinformatics*, 2020, **36**, i258-i267.
113. M. Akiyama, K. Sato and Y. Sakakibara, *Journal of Bioinformatics and Computational Biology*, 2018, **16**, 1840025.
114. K. Sato, M. Akiyama and Y. Sakakibara, *Nat. Commun.*, 2021, **12**, 941.
115. R. B. Denman, *Biotechniques*, 1993, **15**, 1090-1095.
116. G. Sczakiel and M. Tabler, *Methods Mol. Biol.*, 1997, **74**, 11-15.
117. W. James and E. Cowe, *Methods Mol. Biol.*, 1997, **74**, 17-26.
118. Y. Shao, S. Wu, C. Y. Chan, J. R. Klapper, E. Schneider and Y. Ding, *BMC Bioinformatics*, 2007, **8**, 469.
119. A. Mercatanti, C. Lande and L. Citti, *Methods Mol. Biol.*, 2012, **848**, 337-356.
120. Y. Ding and C. E. Lawrence, *Nucleic Acids Res.*, 2003, **31**, 7280-7301.
121. N. Kharma, L. Varin, A. Abu-Baker, J. Ouellet, S. Najeh, M. R. Ehdaeivand, G. Belmonte, A. Ambri, G. Rouleau and J. Perreault, *Nucleic Acids Res.*, 2016, **44**, e39.
122. S. Najeh, N. Kharma, T. Vaudry-Read, A. Haurie, C. Paslawski, D. Adams, S. Ferreira and J. Perreault, *bioRxiv*, 2023, DOI: 10.1101/2023.09.30.560155, 2023.2009.2030.560155.

View Article Online
DOI: 10.1039/D5CB00105F

123. R. Mohammadi-Arani, F. Javadi-Zarnaghi, P. Boccaletto, J. M. Bujnicki and A. Ponce-Salvatierra, *Nucleic Acids Res.*, 2022, **50**, W261-w265.

124. A. Ponce-Salvatierra, P. Boccaletto and J. M. Bujnicki, *Nucleic Acids Res.*, 2020, **49**, D76-D81.

125. A. C. Pine, G. N. Brooke and A. Marco, *NAR Genom Bioinform*, 2023, **5**, lqac098.

126. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler and I. L. Hofacker, *Algorithms Mol Biol*, 2011, **6**, 26.

127. M. Eremeyeva, Y. Din, N. Shirokii and N. Serov, *BMC Bioinformatics*, 2025, **26**, 2.

128. R. Rotrattanadumrong and Y. Yokobayashi, *Nat. Commun.*, 2022, **13**, 4847.

129. J. D. Beck, J. M. Roberts, J. M. Kitzhaber, A. Trapp, E. Serra, F. Spezzano and E. J. Hayden, *Front Mol Biosci*, 2022, **9**, 893864.

130. I. Dotu, J. A. Garcia-Martin, B. L. Slinger, V. Mechery, M. M. Meyer and P. Clote, *Nucleic Acids Res.*, 2014, **42**, 11752-11762.

131. A. Strokach and P. M. Kim, *Curr. Opin. Struct. Biol.*, 2022, **72**, 226-236.

132. S. Sumi, M. Hamada and H. Saito, *Nature Methods*, 2024, **21**, 435-443.

133. G. Pepe, R. Appierdo, C. Carrino, F. Ballesio, M. Helmer-Citterich and P. F. Gherardini, *Front Mol Biosci*, 2022, **9**, 1000205.

134. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583-589.

135. T. Shen, Z. Hu, S. Sun, D. Liu, F. Wong, J. Wang, J. Chen, Y. Wang, L. Hong, J. Xiao, L. Zheng, T. Krishnamoorthi, I. King, S. Wang, P. Yin, J. J. Collins and Y. Li, *Nat. Methods*, 2024, **21**, 2287-2298.

136. F. Bu, Y. Adam, R. W. Adamiak, M. Antczak, B. R. H. de Aquino, N. G. Badepally, R. T. Batey, E. F. Baulin, P. Boinski, M. J. Boniecki, J. M. Bujnicki, K. A. Carpenter, J. Chacon, S.-J. Chen, W. Chiu, P. Cordero, N. K. Das, R. Das, W. K. Dawson, F. DiMaio, F. Ding, A.-C. Dock-Bregeon, N. V. Dokholyan, R. O. Dror, S. Dunin-Horkawicz, S. Eismann, E. Ennifar, R. Esmaeeli, M. A. Farsani, A. R. Ferré-D'Amaré, C. Geniesse, G. E. Ghanim, H. V. Guzman, I. V. Hood, L. Huang, D.

View Article Online
DOI: 10.1039/D5CB00105F

S. Jain, F. Jaryani, L. Jin, A. Joshi, M. Karelina, J. S. Kieft, W. Kladwang, S. Kmiecik, D. Koirala, M. Kollmann, R. C. Kretsch, M. Kurciński, J. Li, S. Li, M. Magnus, B. Masquida, S. N. Moafinejad, A. Mondal, S. Mukherjee, T. H. D. Nguyen, G. Nikolaev, C. Nithin, G. Nye, I. P. N. Pandaranadar Jeyeram, A. Perez, P. Pham, J. A. Piccirilli, S. P. Pilla, R. Pluta, S. Poblete, A. Ponce-Salvatierra, M. Popenda, L. Popenda, F. Pucci, R. Rangan, A. Ray, A. Ren, J. Sarzynska, C. M. Sha, F. Stefaniak, Z. Su, K. C. Suddala, M. Szachniuk, R. Townshend, R. J. Trachman, J. Wang, W. Wang, A. Watkins, T. K. Wirecki, Y. Xiao, P. Xiong, Y. Xiong, J. Yang, J. D. Yesselman, J. Zhang, Y. Zhang, Z. Zhang, Y. Zhou, T. Zok, D. Zhang, S. Zhang, A. Żyła, E. Westhof and Z. Miao, *Nat. Methods*, 2024, DOI: 10.1038/s41592-024-02543-9.

137. W. Wang, C. Feng, R. Han, Z. Wang, L. Ye, Z. Du, H. Wei, F. Zhang, Z. Peng and J. Yang, *Nat. Commun.*, 2023, **14**, 7266.
138. R. Owczarzy, Y. You, B. G. Moreira, J. A. Manthey, L. Huang, M. A. Behlke and J. A. Walder, *Biochemistry*, 2004, **43**, 3537-3554.
139. V. Basilio Barbosa, E. de Oliveira Martins and G. Weber, *Biophys. Chem.*, 2019, **251**, 106189.
140. G. Weber, *Bioinformatics*, 2015, **31**, 871-877.
141. S. B. Zimmerman and S. O. Trach, *J. Mol. Biol.*, 1991, **222**, 599-620.
142. S.-i. Nakano, H. Karimata, T. Ohmichi, J. Kawakami and N. Sugimoto, *J. Am. Chem. Soc.*, 2004, **126**, 14330-14331.
143. S.-i. Nakano, M. Horita, M. Kobayashi and N. Sugimoto, *Catalysts*, 2017, **7**, 355.
144. H. Tateishi-Karimata, T. Ohyama, T. Muraoka, P. Podbevsek, A. M. Wawro, S. Tanaka, S. I. Nakano, K. Kinbara, J. Plavec and N. Sugimoto, *Nucleic Acids Res.*, 2017, **45**, 7021-7030.
145. S. Nakano, D. Miyoshi and N. Sugimoto, *Chem. Rev.*, 2014, **114**, 2733-2758.
146. A. P. Minton, *J. Biol. Chem.*, 2001, **276**, 10577-10580.
147. S. Ghosh, S. Takahashi, T. Ohyama, T. Endoh, H. Tateishi-Karimata and N. Sugimoto, *Proc Natl Acad Sci U S A*, 2020, **117**, 14194-14201.
148. E. Rozners and J. Moulder, *Nucleic Acids Res.*, 2004, **32**, 248-254.
149. S. Ghosh, S. Takahashi, T. Endoh, H. Tateishi-Karimata, S. Hazra and N. Sugimoto, *Nucleic Acids Res.*, 2019, **47**, 3284-3294.
150. M. S. Adams and B. M. Znosko, *Nucleic Acids Res.*, 2019, **47**, 3658-3666.

View Article Online
DOI: 10.1039/D5CB00105F

151. J. P. Sieg, E. A. Jolley, M. J. Huot, P. Babitzke and P. C. Bevilacqua, *Nucleic Acids Res.*, 2023, **51**, 11298-11317.
152. J. M. Huguet, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante and F. Ritort, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 15431-15436.
153. S. Nakano, H. Karimata, T. Ohmichi, J. Kawakami and N. Sugimoto, *J. Am. Chem. Soc.*, 2004, **126**, 14330-14331.
154. S. Ghosh, S. Takahashi, D. Banerjee, T. Ohyama, T. Endoh, H. Tateishi-Karimata and N. Sugimoto, *Nucleic Acids Res.*, 2023, **51**, 4101-4111.
155. D. Banerjee, H. Tateishi-Karimata, M. Toplishek, T. Ohyama, S. Ghosh, S. Takahashi, M. Trajkovski, J. Plavec and N. Sugimoto, *J. Am. Chem. Soc.*, 2023, **145**, 23503-23518.
156. S. Takahashi, J. Yamamoto, A. Kitamura, M. Kinjo and N. Sugimoto, *Anal. Chem.*, 2019, **91**, 2586-2590.
157. T. J. Nott, T. D. Craggs and A. J. Baldwin, *Nat. Chem.*, 2016, **8**, 569-575.
158. S. Ghosh, S. Takahashi, T. Ohyama, L. Liu and N. Sugimoto, *J. Am. Chem. Soc.*, 2024, **146**, 32479-32497.
159. E. P. Bulthuis, C. E. J. Dieteren, J. Bergmans, J. Berkhout, J. A. Wagenaars, E. M. A. van de Westerlo, E. Podhumljak, M. A. Hink, L. F. B. Hesp, H. S. Rosa, A. N. Malik, M. K. Lindert, P. Willems, H. Gardeniers, W. K. den Otter, M. J. W. Adjobo-Hermans and W. J. H. Koopman, *EMBO J.*, 2023, **42**, e108533.
160. L. Liu, S. Takahashi, S. Ghosh, T. Endoh, N. Yoshinaga, K. Numata and N. Sugimoto, *Communications Chemistry*, 2025, In press.
161. L.-C. Park, T. Maruyama and M. Goto, *Analyst*, 2003, **128**, 161-165.
162. S. Pramanik, S. Nagatoishi and N. Sugimoto, *Chem. Commun.*, 2012, **48**, 4815-4817.
163. L. Khamari and S. Mukherjee, *The Journal of Physical Chemistry Letters*, 2022, **13**, 8169-8176.
164. H. Peng, A. Lelievre, K. Landenfeld, S. Müller and I. A. Chen, *Curr. Biol.*, 2022, **32**, 86-96.e86.
165. Y. C. Lai, Z. Liu and I. A. Chen, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**.
166. R. R. Poudyal, C. D. Keating and P. C. Bevilacqua, *ACS Chemical Biology*, 2019, **14**, 1243-1248.
167. R. R. Poudyal, R. M. Guth-Metzler, A. J. Veenis, E. A. Frankel, C. D. Keating and

View Article Online
DOI: 10.1039/D5CB00105F

P. C. Bevilacqua, *Nat. Commun.*, 2019, **10**, 490.

168. J. M. Iglesias-Artola, B. Drobot, M. Kar, A. W. Fritsch, H. Mutschler, T. Y. Dora Tang and M. Kreysing, *Nature Chemistry*, 2022, **14**, 407-416.
169. S. DasGupta, S. Zhang and J. W. Szostak, *ACS central science*, 2023, **9**, 1670-1678.
170. B. P. Paudel and D. Rueda, *J. Am. Chem. Soc.*, 2014, **136**, 16700-16703.
171. S. Nakano, H. T. Karimata, Y. Kitagawa and N. Sugimoto, *J. Am. Chem. Soc.*, 2009, **131**, 16881-16888.
172. D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker and D. H. Turner, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 7287-7292.
173. E. Kierzek, X. Zhang, R. M. Watson, S. D. Kennedy, M. Szabat, R. Kierzek and D. H. Mathews, *Nat. Commun.*, 2022, **13**, 1271.
174. R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler and I. L. Hofacker, *Algorithms Mol. Biol.*, 2011, **6**, 26.
175. I. L. Hofacker, *Nucleic Acids Res.*, 2003, **31**, 3429-3431.
176. J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks and N. A. Pierce, *J. Comput. Chem.*, 2011, **32**, 170-173.
177. E. Rivas and S. R. Eddy, *J. Mol. Biol.*, 1999, **285**, 2053-2068.
178. S. Zakov, Y. Goldberg, M. Elhadad and M. Ziv-Ukelson, *J. Comput. Biol.*, 2011, **18**, 1525-1542.
179. M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews and K. P. Murphy, *Bioinformatics*, 2007, **23**, i19-i28.
180. J. Ren, B. Rastegari, A. Condon and H. H. Hoos, *RNA*, 2005, **11**, 1494-1504.
181. H. K. Wayment-Steele, W. Kladwang, A. I. Strom, J. Lee, A. Treuille, A. Becka and R. Das, *Nat. Methods*, 2022, **19**, 1234-1242.
182. B. Knudsen and J. Hein, *Nucleic Acids Res.*, 2003, **31**, 3423-3428.
183. R. D. Dowell and S. R. Eddy, *BMC Bioinformatics*, 2004, **5**, 71.
184. T. Gong, F. Ju and D. Bu, *Communications Biology*, 2024, **7**, 297.
185. R. J. Penić, T. Vlašić, R. G. Huber, Y. Wan and M. Šikić, *arXiv preprint arXiv:2403.00043*, 2024.
186. J. K. Franke, F. Runge, R. Köksal, R. Backofen and F. Hutter, *bioRxiv*, 2024, 2024.2002. 2012.579881.
187. S. Pramanik, S. Nagatoishi, S. Saxena, J. Bhattacharyya and N. Sugimoto, *J.*

View Article Online
DOI: 10.1039/D5CB00105F

Phys. Chem. B, 2011, **115**, 13862-13872.

188. I. Ferreira, E. A. Jolley, B. M. Znosko and G. Weber, *Chem Phys*, 2019, **521**, 69-76.

Table 1. Prediction of RNA and DNA secondary structure based on nearest-neighbour (NN) and non-NN models

Method	Concept	Feature	Ref
NN model based prediction			
MFold/UNAFold	MFE-based thermodynamics	The most commonly used prediction tool and has now been replaced by UNAFold.	96-98
RNAfold	MFE-based thermodynamics	Availability to compute the equilibrium partition functions and base-pairing probabilities.	174, 175
Sfold	MFE-based thermodynamics	Sampling all possible structures in the Boltzmann ensemble of secondary structures.	120
RNAstructure	MFE-based thermodynamics	Database using alternative set of thermodynamic parameters compared to MFold and SHAPE data.	99, 100
LinearX tool	MFE-based thermodynamics	Fast prediction of the secondary structure from a long RNA strand using the beam search technique.	111, 112
NUPACK	MFE-based thermodynamics	Applicable to the prediction of pseudoknot structure.	176
PKNOTS	MFE-based thermodynamics	Applicable to the prediction of pseudoknot structure.	177
CONTRAfold	Machine learning with conditional log-linear models	Trained by the nearest neighbor models (without NN parameters) for solved RNA secondary structures with parameters corresponding to free energy.	104
ContextFold	Machine learning with max-margin framework	Trained by the nearest neighbor models (without NN parameters) using fine-grained RNA structures.	178
CentroidFold	Non MFE-based thermodynamics	The maximum expected accuracy approach	105
SimFold	MFE-based thermodynamics & machine learning with constraint generation and Boltzmann likelihood	Trained by large sets of structural as well as NN parameters for predicting the secondary structure of RNA with thermodynamic parameters.	179
MXfold	MFE-based thermodynamics & machine learning with max-margin framework	Combination of NN parameters with the structural data of RNAs trained by a method called structured support vector machine for precise prediction of substrutures.	113
MXfold2	MFE-based thermodynamics & machine learning with max-margin framework and deep learning	Application of deep learning to learn RNA folding (helix stacking, helix opening, helix closing, and unpaired region) scores based on NN parameters	114
HotKnots	MFE-based thermodynamics & machine learning with constraint generation	Applicable to the prediction of pseudoknot structure.	180
EternaFold	Multi task machine learning methods	Trained by NN model based on different data types, SHAPE structures, and riboswitch-ligand binding affinity data for the accurate prediction of RNA structures.	181
Non-NN based model			

View Article Online
DOI: 10.1039/D5CB00105F

Pfold	Probabilistic generative model using stochastic context-free grammars	Utilizing simple context-free grammars.	182
CONUS	Probabilistic generative model using stochastic context-free grammars	Comparing nine lightweight grammars for RNA secondary structure prediction.	183
TORNADO	Probabilistic generative model using stochastic context-free grammars	Describing various RNA grammars including NN models	110
SPOT-RNA	Deep learning based model	Using an ensemble of ultra-deep hybrid networks and pre-trained with a large set of non-redundant RNAs.	107
E2Efold	Deep learning based model	Predicting the probability of each nucleotide match by machine learning without any NN parameters	108
Ufold	Deep learning based model	Representing base matching on RNA sequences as "pseudo-image information" in a two-dimensional matrix.	109
Knotfold	Deep learning based model	Applicable to the prediction of pseudoknot structure.	184
RiNALMo	Deep learning based model	Utilizing the 650 million parameters RNA language model	185
RNAformer	Deep learning based model	Facilitating the application of axial attention like AlphaFold protein prediction.	186

Table 2. Representative tools for the design of NA enzymes

Method	Design	Feature	Ref
Sfold	RNA-cleaving ribozymes	Scoring of the complex formation of ribozyme and target RNA based on ΔG values obtained from the MFE-based secondary structure prediction.	118
Aladdin	HH ribozyme	Optimization of the stem stability of the ribozyme-target complex obtained from the MFE-based secondary structure prediction.	119
RiboSoft	RNA-cleaving ribozymes	Output of potential sequences with automatically minimizing the off-target effect.	121, 122
DNAzymeBuilder	RNA-cleaving DNAzyme	The sequence design of DNAzymes based on NN parameters and	123
NAR Genom Bioinform 2023	10–23 DNAzyme	The sequence design of DNAzymes using NN parameters with machine learning based on the stem stability and the internal structure of the DNAzyme.	125
SequenceCraft	RNA-cleaving DNAzyme	Machine learning algorithms capable of predicting DNAzyme sequence and the potential rate constants based on various sequence-, cofactor-, and buffer-related factors.	127
Nat. Commun. 2024	Ligase ribozyme	Providing the fitness landscape to rationally design the novel ribozymes.	128
RNAiFold	Various types of NA enzymes (HH ribozyme was demonstrated.)	Generation of ribozyme sequence under the concept to of inverse folding based on the MFE-based secondary structure prediction.	130
RfamGen	Various types of NA enzymes (*glmS* ribozyme was demonstrated.)	Generation of novel ribozyme sequences by deep learning of characteristics of a group of RNAs with specific functional and structural features.	132

View Article Online
DOI: 10.1039/D5CB00105F

View Article Online
DOI: 10.1039/D5CB00105F

Table 3. NN parameters for DNA duplexes in 40 wt% PEG200 and 100 mM NaCl at 37ºC[a]

Sequence	ΔH°_{NN} (kcal mol^{-1})	ΔS°_{NN} (cal mol^{-1} K^{-1})	$\Delta G^\circ_{37\ NN}$ (kcal mol^{-1})
DNA/DNA			
d(AA/TT)	–6.5 ± 0.3	–19.2 ± 0.8	–0.55 ± 0.07
d(AT/TA)	–9.4 ± 0.3	–29.4 ± 0.8	–0.28 ± 0.05
d(TA/AT)	–4.3 ± 0.5	–13.3 ± 1.3	–0.16 ± 0.14
d(CA/GT)	–13.1 ± 0.1	–38.8 ± 0.1	–1.00 ± 0.05
d(GT/CA)	–9.2 ± 0.1	–26.8 ± 0.1	–0.89 ± 0.01
d(CT/GA)	–3.4 ± 0.6	–7.9 ± 1.6	–0.91 ± 0.11
d(GA/CT)	–4.9 ± 0.7	–13.0 ± 2.1	–0.87 ± 0.06
d(CG/GC)	–6.4 ± 0.7	–16.1 ± 2.0	–1.38 ± 0.12
d(GC/CG)	–4.2 ± 0.7	–9.3 ± 2.0	–1.31 ± 0.06
d(GG/CC)	–4.0 ± 0.6	–8.9 ± 2.0	–1.25 ± 0.03
Initiation per GC	–10.1 ± 0.2	–35.1 ± 0.5	0.76 ± 0.06
Initiation per AT	–2.9 ± 0.3	–12.7 ± 0.9	1.00 ± 0.07
Self-complementary	0	–1.4	0.40
Non-self-complementary	0	0	0
RNA/RNA			
r(AA/UU)	-10.0 ± 0.1	-30.4 ± 0.2	-0.57 ± 0.05
r(AU/UA)	-10.1 ± 0.1	-30.8 ± 0.1	-0.55 ± 0.03
r(UA/AU)	-11.1 ± 0.4	-31.5 ± 0.8	-1.33 ± 0.09
r(CA/GU)	-12.1 ± 0.3	-32.1 ± 0.6	-2.14 ± 0.08
r(GU/CA)	-10.7 ± 0.2	-28.7 ± 0.1	-1.80 ± 0.16
r(CU/GA)	-11.2 ± 0.3	-30.4 ± 0.3	-1.77 ± 0.17
r(GA/CU)	-11.7 ± 0.2	-30.5 ± 0.3	-2.24 ± 0.07
r(CG/GC)	-11.1 ± 0.5	-28.8 ± 1.1	-2.16 ± 0.16
r(GC/CG)	-13.8 ± 0.1	-34.6 ± 0.1	-3.07 ± 0.01
r(GG/CC)	-14.8 ± 0.1	-38.4 ± 0.1	-2.89 ± 0.06
initiation	4.6 ± 2.0	-2.9 ± 6.1	5.50 ± 0.11
per terminal AU	6.5 ± 0.1	18.2 ± 0.1	0.85 ± 0.92
self-complementary	0	–1.4	0.43
non-self-complementary	0	0	0
RNA/DNA			
rAA/dTT	–6.6 ± 0.4	–19.8 ± 0.2	–0.46 ± 0.05
rAC/dGT	–8.3 ± 0.3	–23.0 ± 0.2	–1.17 ± 0.06
rAG/dCT	–7.7 ± 0.0	–20.7 ± 0.1	–1.28 ± 0.01
rAU/dAT	–7.6 ± 0.5	–24.0 ± 0.1	–0.16 ± 0.03
rCA/dTG	–9.0 ± 0.3	–27.2 ± 0.3	–0.56 ± 0.09
rCC/dGG	–8.1 ± 0.1	–20.5 ± 0.0	–1.74 ± 0.01
rCG/dCG	–7.8 ± 0.2	–21.5 ± 0.1	–1.13 ± 0.02
rCU/dAG	–5.3 ± 0.1	–16.5 ± 0.1	–0.18 ± 0.02
rGA/dTC	–6.8 ± 0.2	–19.2 ± 0.1	–0.85 ± 0.02

View Article Online
DOI: 10.1039/D5CB00105F

rGC/dGC	–8.6 ± 0.0	–21.7 ± 0.1	–1.87 ± 0.03
rGG/dCC	–11.5 ± 0.3	–30.9 ± 1.4	–1.92 ± 0.00
rGU/dAC	–7.1 ± 0.4	–20.2 ± 0.6	–0.83 ± 0.03
rUA/dTA	–7.3 ± 0.5	–22.9 ± 0.1	–0.20 ± 0.03
rUC/dGA	–5.7 ± 0.6	–13.9 ± 0.2	–1.39 ± 0.05
rUG/dCA	–8.0 ± 0.5	–21.5 ± 0.1	–1.33 ± 0.04
rUU/dAA	–7.2 ± 0.1	–22.7 ± 0.1	–0.16 ± 0.04
init. per rG-dC or rC-dG	–5.0 ± 0.3	–19.7 ± 0.3	1.11 ± 0.16
init. per rA-dT or rU-dA	–3.0 ± 0.1	–13.9 ± 0.4	1.31 ± 0.12

[a]Experiments were conducted in 10 mM Na_2HPO_4, 1 mM Na_2EDTA, 100 mM NaCl, and 40 wt% PEG200 at pH 7.0.

Table 4. DNA/DNA and RNA/RNA NN parameters for $\Delta G^\circ_{37\ NN\ [cation]}$ and $\Delta G^\circ_{37\ NN\ [crowder]}$ in 100 mM NaCl, with prefactors (m_{cs}) for different cosolutes[a]

Sequence	$\Delta G^\circ_{37\ NN\ [cation]}$[b] (kcal mol^{-1})	$\Delta G^\circ_{37\ NN\ [crowder]}$[c] (kcal mol^{-1})	m_{cs}[d] (kcal mol^{-1})		
DNA/DNA			PEG/1,2 DME	EG/GLY	1,3PDO/ 2-ME
d(AA/TT)	-0.65	0.10	2.0	0.7	1.3
d(AT/TA)	-0.60	0.32	6.4	2.2	4.2
d(TA/AT)	-0.36	0.20	4.0	1.4	2.6
d(CA/GT)	-1.23	0.23	4.6	1.6	3.0
d(GT/CA)	-1.20	0.31	6.2	2.2	4.1
d(CT/GA)	-1.11	0.20	4.0	1.4	2.6
d(GA/CT)	-0.93	0.06	1.2	0.4	0.8
d(CG/GC)	-1.85	0.47	9.4	3.3	6.2
d(GC/CG)	-2.05	0.72	14.4	5.0	9.5
d(GG/CC)	-1.69	0.44	8.8	3.0	5.8
Initiation per GC	0.98	-0.22	-4.4	-1.5	-2.9
Initiation per AT	1.03	-0.03	-0.6	-0.2	-0.4
RNA/RNA		$\Delta G^\circ_{37\ NN\ ev\ [crowder]}$[e] (kcal mol^{-1})	$\Delta G^\circ_{37\ NN\ wa\ [crowder]}$[e] (kcal mol^{-1})	PEG/2-ME/1,2 DME	EG/GLY/1,3 PDO
r(AA/UU)	-0.77	-0.22	0.35	7.1	2.9
r(AU/UA)	-0.52	-0.22	0.19	3.9	1.6
r(UA/AU)	-1.25	-0.22	0.19	3.9	1.6
r(CA/GU)	-1.77	-0.22	-0.14	-2.9	-1.2
r(GU/CA)	-2.08	-0.22	0.56	11.4	4.7
r(CU/GA)	-1.76	-0.22	0.25	5.1	2.1
r(GA/CU)	-2.20	-0.22	0.20	4.1	1.7
r(CG/GC)	-2.16	-0.22	0.24	4.9	2.0
r(GC/CG)	-3.24	-0.22	0.45	9.2	3.8

View Article Online
DOI: 10.1039/D5CB00105F

r(GG/CC)	-3.08	-0.22	0.43	8.8	3.6
initiation	-0.77	-0.22	1.63	33.3	13.7
per terminal AU	-0.52	NA[f]	0.40	8.2	3.4

[a]Correction factor for self-complementary sequences is 0.4 kcal mol^{-1} for all cosolutes, as it is independent of crowding environments.

[b]Cation concentration is 100 mM Na^+.

[c]Crowder condition is 40 wt% PEG200.

[d]Different cofactors were used for each crowder: polyethylene glycol (PEG), 2-methoxy ethanol (2-ME), 1,2-dimethoxyethane (1,2 DME), ethylene glycol (EG), glycerol (GLY), and 1,3-propanediol (1,3 PDO)

[e]In the case of RNA/RNA, the excluded volume effect and water activity contribution should be considered separately for accurate prediction.

[f]Excluded volume effect for terminal AU pairs was not considered to avoid overestimation, as it had already been considered for initiation.

Table 5. Experimental and predicted thermodynamic parameters for DNA and RNA sequences under different cosolute and salt conditions

Sequence	Solution[a]	Measured ΔG°_{37} (kcal mol^{-1})	Ref	Predicted ΔG°_{37} (kcal mol^{-1})
d(GAGGTCGT)	10 wt% PEG200 at 1 M NaCl	–8.4 ± 0.1	[153]	–8.3
	20 wt% PEG200 at 1 M NaCl	–7.7 ± 0.1		–7.9
	30 wt% PEG200 at 1 M NaCl	–7.0 ± 0.3		–7.0
d(ATGCGCAT)	20 wt% PEG1000 at 1 M NaCl	–8.2 ± 0.3	[187]	–8.1
	20 wt% PEG6000 at 1 M NaCl	–8.6 ± 0.6		–8.4
d(CCGTACGG)	20 wt% EG at 100 mM NaCl	–7.2 ± 0.8	[147]	–6.7
	20 wt% 1,3 PDO at 100 mM NaCl	–6.6 ± 0.3		–6.2
d(CCGTAACGTTGG)	20 wt% EG at 100 mM NaCl	–10.9 ± 0.8	[147]	–10.5
	20 wt% 1,3 PDO at 100 mM NaCl	–10.8 ± 0.9		–9.9
r(GGCUCAAUUGAC)	10 wt% PEG200 at 100 mM NaCl	-15.1 ± 0.8	[154]	-15.4
	20 wt% PEG200 at 100 mM NaCl	-14.8 ± 0.6		-14.7
	30 wt% PEG200 at 100 mM NaCl	-14.0 ± 0.7		-14.0
	40 wt% PEG200 at 100 mM NaCl	-13.8 ± 0.6		-13.7
	20 wt% EG at 100 mM NaCl	-14.7 ± 0.6		-14.7
	20 wt% PEG2000 at 100 mM NaCl	-16.9 ± 0.8		-16.3
	20 wt% PEG8000 at 100 mM NaCl	-16.7 ± 0.4		-16.5
	10 wt% PEG200 at 1 M NaCl	-18.9 ± 0.8		-17.9
	20 wt% PEG200 at 1 M NaCl	-18.3 ± 0.6		-17.2
r(GGAUCGAUCC)	20 wt% EG at 100 mM NaCl	-12.7 ± 0.7	[154]	-13.1

View Article Online
DOI: 10.1039/D5CB00105F

r(AUCAGCUGAU)	20 wt% EG at 100 mM NaCl	-9.9 ± 0.6	154	-9.9
r(GGCUCAAUUGAC)	20 wt% EG at 100 mM NaCl	-14.4 ± 0.6	154	-15.0
r(GAUCCGGAUC)	20 wt% 1,3 PDO at 100 mM NaCl	-14.5 ± 0.7	154	-12.9
r(GGCUCAAUUGAC)	20 wt% 1,3 PDO at 100 mM NaCl	-14.7 ± 0.6	154	-14.7
r(GCUAUG)	20 vol% PEG200 at 1 M NaCl	-5.2 ± 0.2	150	-5.2
r(AGAUAUCU)	20 vol% PEG200 at 1 M NaCl	-5.7 ± 0.1	150	-5.6
r(UUAUCGAUAA)	20 vol% PEG200 at 1 M NaCl	-6.9 ± 0.0	150	-6.8

[a]Experiments were performed in a buffer containing 10 mM Na_2HPO_4 (pH 7.0), 1 mM Na_2EDTA, and NaCl.

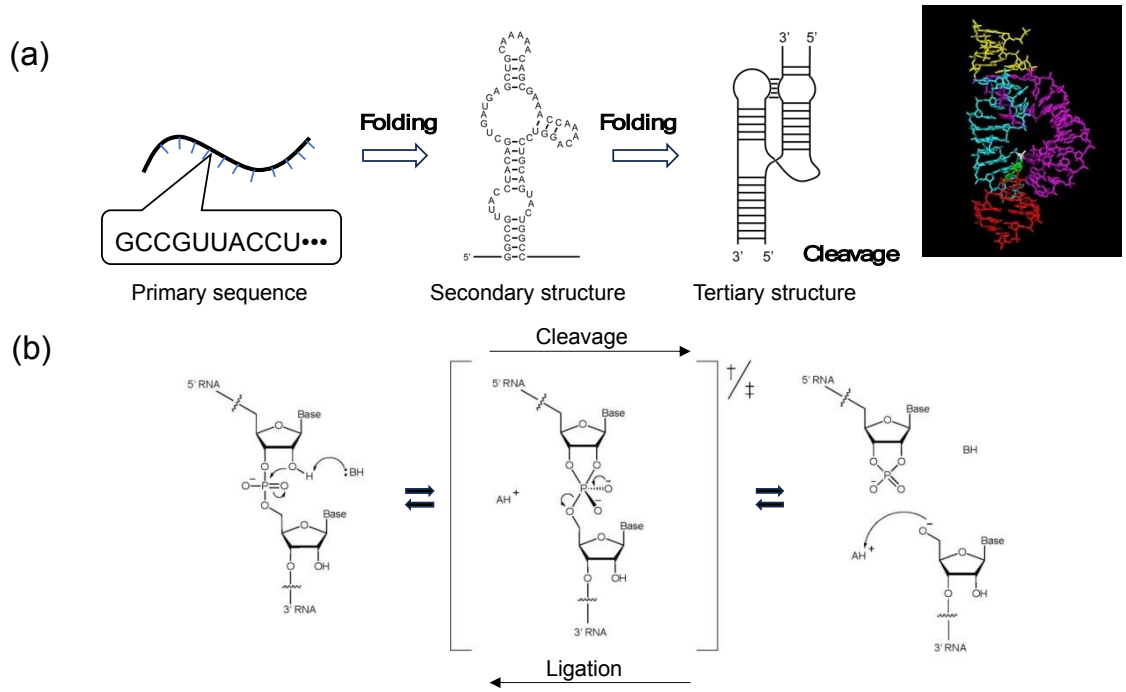


Fig. 1 (a) Folding process from the primary sequence to the secondary and tertiary structures to form an active ribozyme. The structures depict a HH ribozyme. (b) Chemical mechanism of cleavage and ligation by nucelic acid (NA) enzymes, showing catalysis by a general acid (AH^+) and base (:B).

View Article Online
DOI: 10.1039/D5CB00105F

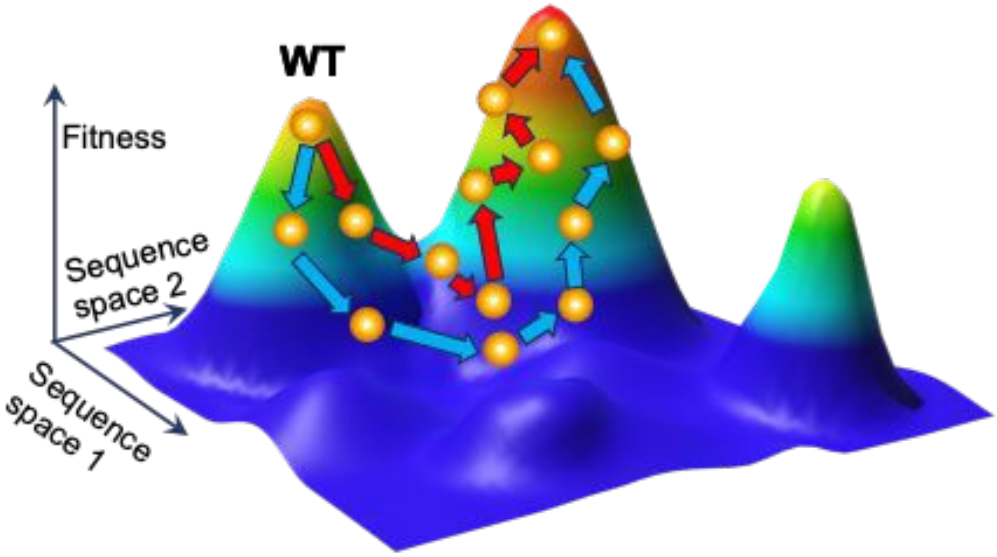


Fig. 2 Conceptual illustration of the fitness landscape of a NA enzyme. Most wild-types (WT) are located on or near the top of isolated fitness peaks, where only a few mutational steps lead to a significant reduction in fitness. Red arrows indicate the evolution process driven by genetic mutations. Blue arrows indicate a different evolutional process discovered by AI technique, which predicts the structure and function of NA enzyme efficiently. The sequence space of NAs is represented by a dimension with four possible nucleobases at each position along the NA chain. To visualize these vast sequence spaces, the sequence information is usually compressed into two dimensions using principal component analysis.

View Article Online
DOI: 10.1039/D5CB00105F

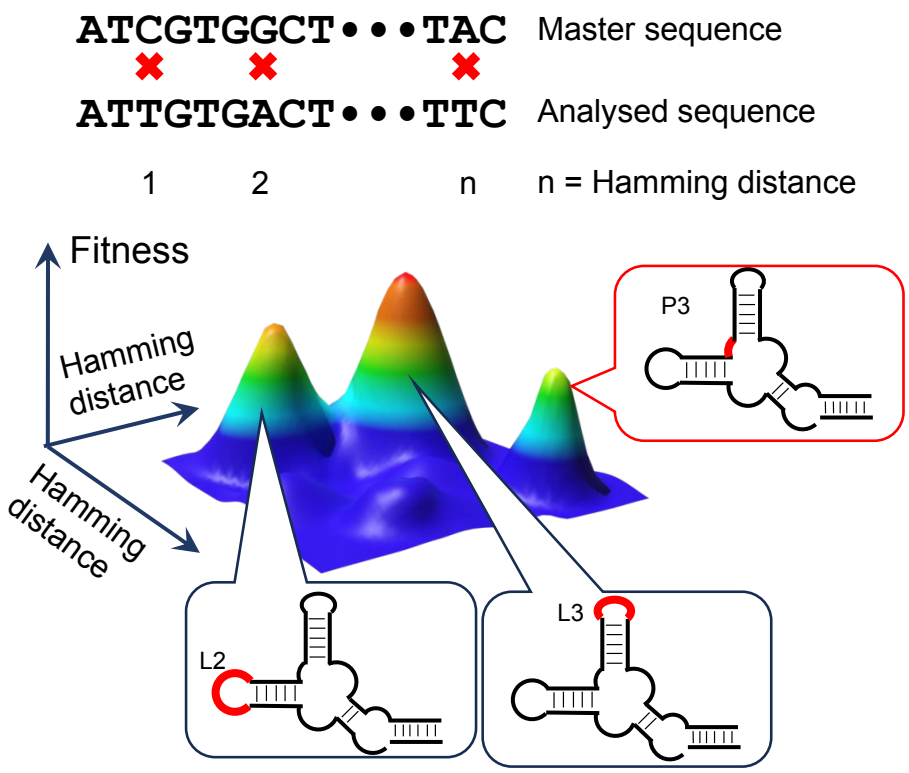
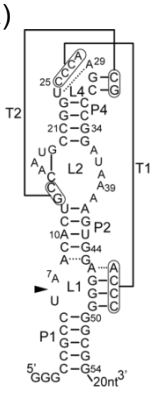


Fig. 3 Schematic illustration of population structure before and after one round of *in vitro* selection of Ref 48. The experimentally constructed fitness landscape clarified that the distal end of paired region (helix) 3 (P3) is a key residue for the class II ligase ribozyme, which had not been identified.

View Article Online
DOI: 10.1039/D5CB00105F

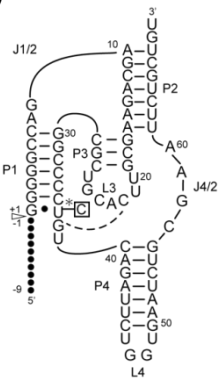
(a)



The active site centered in a double-pseudoknot with two two long-range tertiary interactions

Found in nature with 2,700 examples observed across bacteria, fungi, plants, and animals.

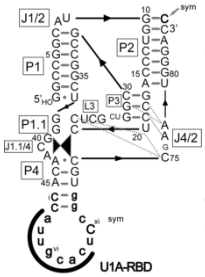
(b)



Located in the second intron of the CPEB3 gene which belongs to a family of genes regulating mRNA polyadenylation.

Highly conserved and found only in mammals.

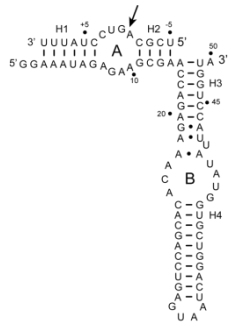
(c)



Five helical segments connected by a double pseudoknot.

Necessary for viral replication of Hepatitis delta virus in human cells.

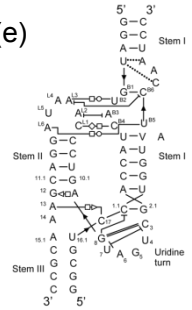
(d)



Two domains, each consisting of two short base paired helices separated by an internal loop.

No requirement of metal ions for the catalysis.

(e)



Composed of three base paired helices, separated by short linkers of conserved sequence

No requirement of metal ions for the catalysis.

Fig. 4 Secondary structures of representative ribozymes. (a) Twister (Osa-1-4), (b) CPEB3, (c) HDV, (d) hairpin, and (e) HH ribozymes. The typical physicochemical and biological characteristics of each ribozyme are shown in the figure.

View Article Online
DOI: 10.1039/D5CB00105F

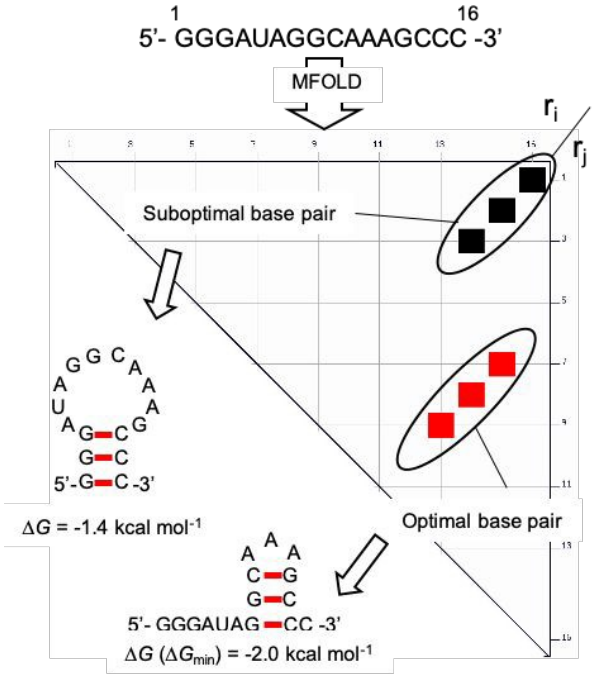


Fig. 5 Schematic illustration of MFold prediction.[40] The energy dot plot displays the predicted optimal structure and one sub-optimal structure for the example sequence shown in the figure.

View Article Online
DOI: 10.1039/D5CB00105F

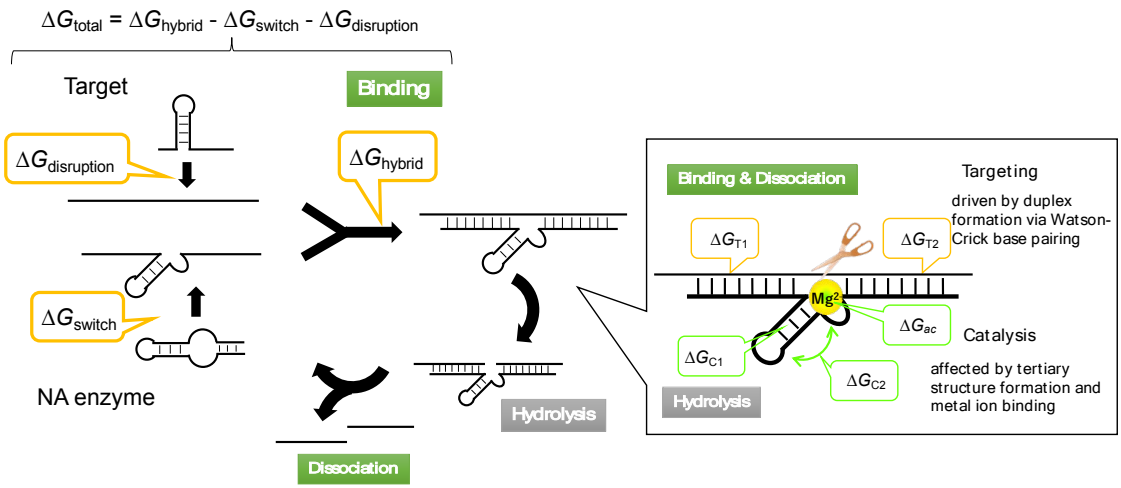


Fig. 6 Schematic illustration of energetic contributions of NA enzyme reaction. The energetic contribution of NA structure formation affects the target disruption ($\Delta G_{disruption}$), the conformational switch of NA enzyme strand (ΔG_{switch}), the binding to (dissociation from) a target (ΔG_{hybrid} related to ΔG_{T1} and ΔG_{T2}) and folding of the core of NA enzyme (ΔG_{C1} and ΔG_{C2}). The cofactor binding of Mg^{2+} is essential for the tertiary structure of the active center (ΔG_{ac}) for the catalysis.

View Article Online
DOI: 10.1039/D5CB00105F

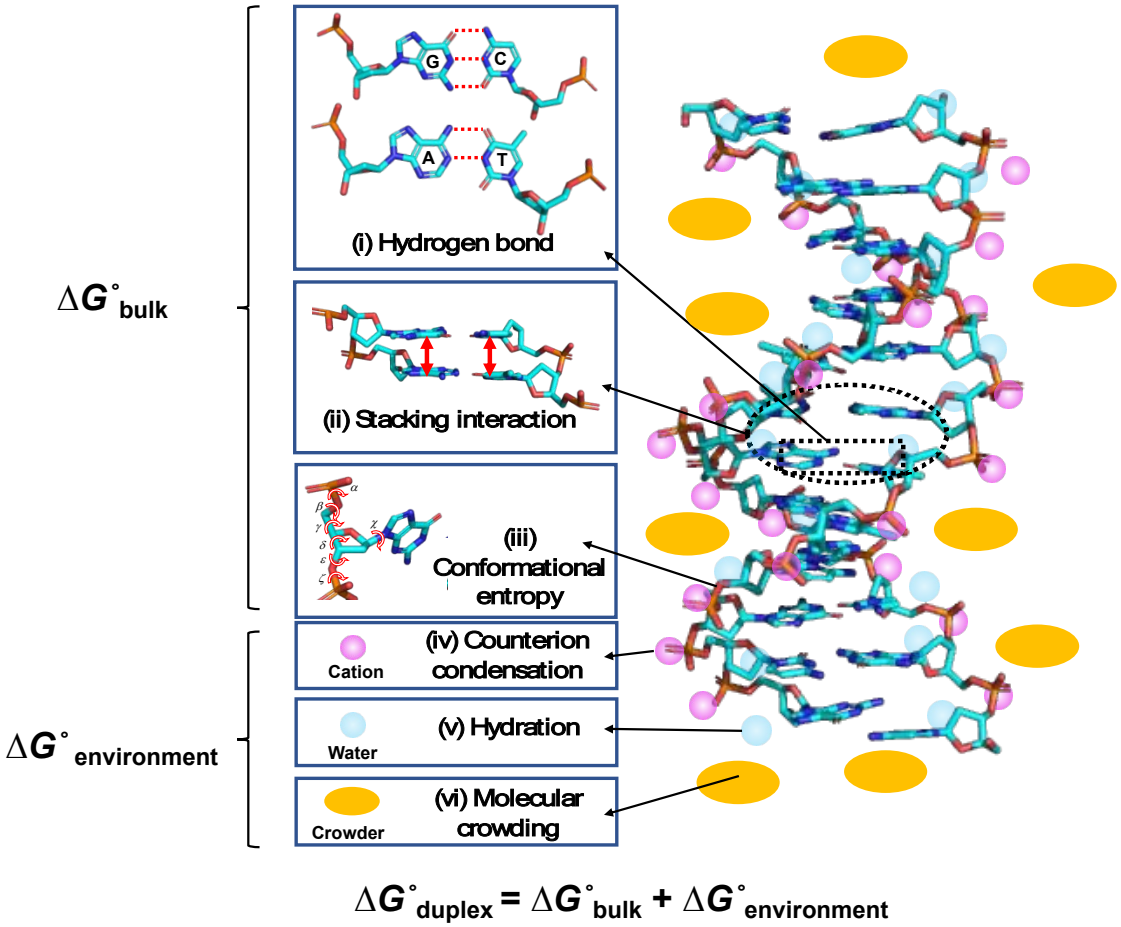


Fig. 7 Factors determining the thermodynamic stability of a canonical NA duplex. The duplex stability is determined by the bulk factors (hydrogen bondings, stacking interactions, and conformational entropy) and environmental factors (cation, water, and crowder interactions).

View Article Online
DOI: 10.1039/D5CB00105F

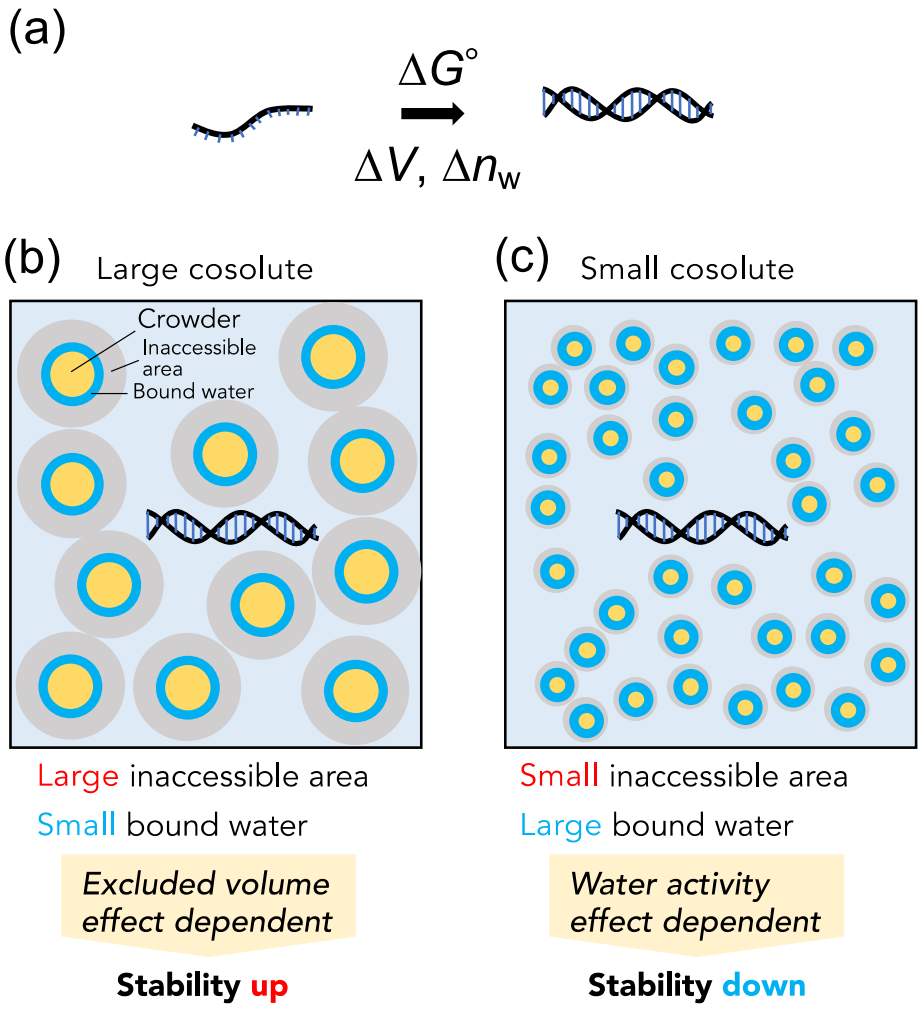


Fig. 8 Schematic illustrations of the effect of crowders on the stability of NA structures. (a) Formation of NA structure, which occurs with volumetric (ΔV) and hydration (Δn_w) changes. Physical contributions of (b) large and (c) small cosolutes in an aqueous solution. Additonal factors such as changes in dielectric constant, viscosity, and direct interactions with crowders influence NA stability.

View Article Online
DOI: 10.1039/D5CB00105F

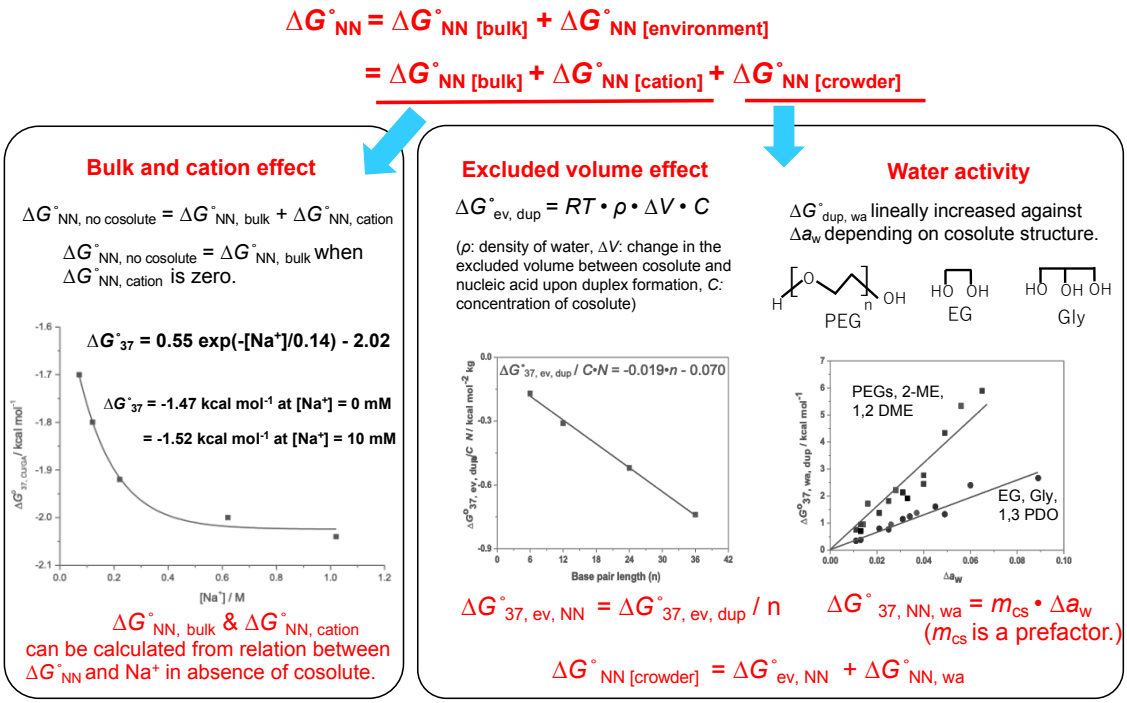


Fig. 9 Schematic representation of the contributions of bulk structure, cations, and crowders to NN parameters. (Left) Variation of $\Delta G^\circ_{37,\ NN}$ against the Na^+ concentration, based on data from Weber's report[188]. (Mid) Plot showing the excluded volume effect for RNA-PEG interactions against base pair length. (Right) Plot of the contribution of water activity on the stability of r(GAUUACGCCUG) against Δa_w.

View Article Online
DOI: 10.1039/D5CB00105F

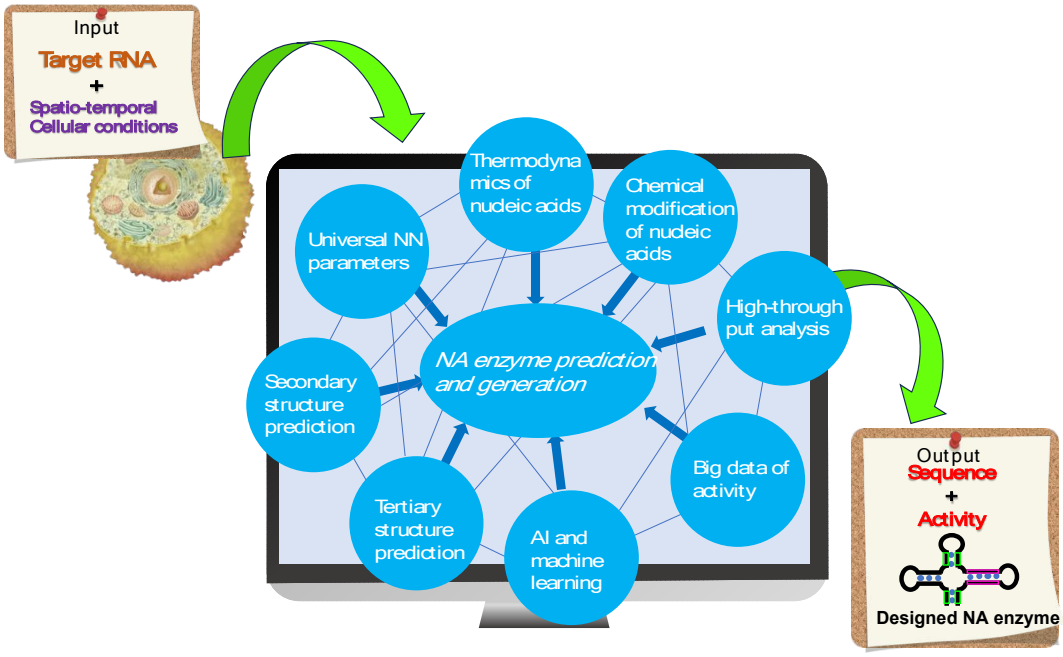


Fig. 10 Required elements for the rational design of active NA enzymes in cells from NA sequence and environment information.

View Article Online
DOI: 10.1039/D5CB00105F

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.