nature reviews genetics

https://doi.org/10.1038/s41576-022-00567-5

Review article

Check for updates

High-throughput biochemistry in RNA sequence space: predicting structure and function

Emil Marklund [1], Yuxi Ke [2] & William J. Greenleaf [1]

Abstract

RNAs are central to fundamental biological processes in all known organisms. The set of possible intramolecular interactions of RNA nucleotides defines the range of alternative structural conformations of a specific RNA that can coexist, and these structures enable functional catalytic properties of RNAs and/or their productive intermolecular interactions with other RNAs or proteins. However, the immense combinatorial space of potential RNA sequences has precluded predictive mapping between RNA sequence and molecular structure and function. Recent advances in high-throughput approaches in vitro have enabled quantitative thermodynamic and kinetic measurements of RNA–RNA and RNA–protein interactions, across hundreds of thousands of sequence variations. In this Review, we explore these techniques, how they can be used to understand RNA function and how they might form the foundations of an accurate model to predict the structure and function of an RNA directly from its nucleotide sequence. The experimental techniques and modelling frameworks discussed here are also highly relevant for the sampling of sequence–structure–function space of DNAs and proteins.

Sections

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Bioengineering, Stanford University, Stanford, CA, USA. e-mail: wjg@stanford.edu

Introduction

Sequence-specific RNA–RNA and RNA–protein interactions are essential for cellular regulation and function. Although RNAs have diverse biological functions, they are all composed of the same four nucleotides, and thus the regulatory roles of RNAs must be determined by the combinatorial code built by these nucleotide building blocks. The RNA sequence determines the structural conformations that are available to the molecule and its interactions with RNA-binding proteins. RNA interacts with itself through a diverse set of intramolecular interactions, including both canonical Watson–Crick base pairing and non-canonical base pairing, to form helices, loops and bulges that define the overall three-dimensional structure of the molecule[1] (Fig. 1). This structural diversity, which can often change dynamically in response to the binding of cellular signals such as ligands or proteins, in turn defines the diverse set of functions that a specific RNA might have[2–7]. Furthermore, RNA does not often exist in a single conformation in physiological conditions, but rather as a range of alternative conformational states (the ensemble) that coexist for systems at equilibrium or as a constantly changing ensemble of conformations for systems out of equilibrium, in which some conformational states are more or less likely to be observed[2,8,9]. Therefore, the linkage of RNA sequence to structure is complex and does not involve a single solution but rather the definition of one or more ensembles of conformational states, in which each state is assigned a relative weighting. To derive these conformational ensembles from RNA sequence is thus challenging and is fundamentally a combinatorial problem, such that we might anticipate that a large diversity of RNA sequence space must be sampled and characterized before appropriate predictive models can be built.

Traditional methods in structural biology and biochemistry — whereby RNA structure has been studied by X-ray crystallography and NMR spectroscopy[10–13], and RNA–RNA interactions have been studied using different biochemical assays[14–17] — often have relatively low throughput, which makes it challenging to use these methods to sample the vast combinatorial RNA sequence space with any degree of completeness. More recently, various sequencing-based methods have been developed to probe RNA structure[18–21] and RNA–protein binding[22–24] in vivo with high throughput. These in vivo methods are not discussed in detail here but have been covered thoroughly in a recent review[25]. In short, the in vivo methods reveal the structural and binding properties of RNAs at one time point, while averaging over some or all of the RNA structures in the ensemble of conformations that are present. Therefore, in their current implementation, these assays do not quantify the occupancies and transition rates of all naturally occurring RNA structural states and they cannot easily be interpreted to determine the thermodynamic and kinetic parameters that regulate the structure and function of RNAs in the cell. New high-throughput, quantitative biochemical methods are now beginning to be used to address these limitations.

Here, we discuss how next-generation sequencing technologies enable quantitative measurements in vitro of the structural features and molecular functions of millions of biomolecular RNA sequence mutants at the same time. We start by explaining the methodology, highlighting the differences from more traditional biochemical assays for measuring biomolecular conformation, affinities and kinetics. Next, we review how these methods have been used so far to study RNA–RNA interactions, RNA–protein interactions, RNA-guided protein interactions, RNA–small molecule interactions and catalytic RNAs. We provide an outlook on how the large amounts of quantitative binding data that are generated might be used to enable quantitative predictions of RNA conformational states, stability and regulatory functions directly from the nucleotide sequence. Finally, we discuss the future extension of these methods to enable the investigation of different biochemical characteristics, to measure single molecules, to further scale up in terms of throughput and to improve availability.

The procedure of studying binding to DNA, RNA or protein with high throughput on next-generation sequencing chips has been implemented by different groups with some variations in the details of experimental protocols. The protocol for studying binding to DNA has been referred to as HiTS-FLIP (high-throughput sequencing–fluorescent ligand interaction profiling)[26]. For RNA, two highly related methods are HiTS-RAP (high-throughput sequencing and RNA affinity profiling)[27] and RNA-MaP (quantitative analysis of RNA on a massively parallel array)[28]. For studying protein mutants, the method has been referred to as Prot-MaP (protein display on a massively parallel array)[29]. In this Review, we refer to these classes of techniques as DNA, RNA and protein array technologies, respectively. We focus our discussion on RNA-related applications, but our conclusions about the utility of and next steps for these applications are also highly relevant for functional sampling of DNA sequence space and protein sequence space.

Methodologies

Although next-generation sequencing has markedly expanded the known universe of RNA sequences that exist in cells, it has not provided insight into the structure and function of these molecules. To attempt to address this limitation, RNA array technologies use the same high-throughput methods that are used to identify these sequences to assess their structure or function.

Biochemistry on sequencing chips

The RNA sequences to be studied by RNA array are first designed as a DNA library, later to be transcribed to RNA directly on the sequencing flow cell. The number of molecular variants that can be screened simultaneously — in other words, the size of the DNA library — has ranged from 10^3 to 10^7 sequence variants[30,31]. The DNA library is then amplified and sequenced using Illumina 'sequencing by synthesis' technology. First, individual DNA molecules are immobilized on a flow cell surface and each individual DNA is amplified to generate a cluster of ~1,000 identical DNA molecules. Next, the DNA sequence of each cluster is determined using sequencing by synthesis, by flowing in fluorescent nucleotides that are incorporated into the growing complementary DNA strand such that each cycle of sequencing determines the identity of one nucleotide in each DNA cluster. Sequencing is carried out for 30–600 cycles on the Illumina sequencer to determine the identity of each DNA cluster on the array (Fig. 2a). For studies of a biomolecule of interest that binds DNA, binding to the DNA clusters can be studied directly. Indeed, this DNA array approach was the first implementation of a high-throughput biophysical measurement on a sequencing chip, involving binding of the yeast transcription factor GCn4 to a library of DNA sites[27]. High-throughput measurements of DNA binding are not limited to Illumina sequencing technology and have also been made on the DNA nanoballs that are used in the BGISEQ-500 sequencing platform[32].

To assay RNA, the DNA array on a sequenced flow cell is converted to RNA by in situ transcription (Fig. 2b). First, before transcription can take place, the single-stranded DNA (ssDNA) on the array surface is made double stranded by annealing a primer and extending the complementary strand with a DNA polymerase. After transcription of the nascent RNA by RNA polymerase, a 'roadblock' is introduced to

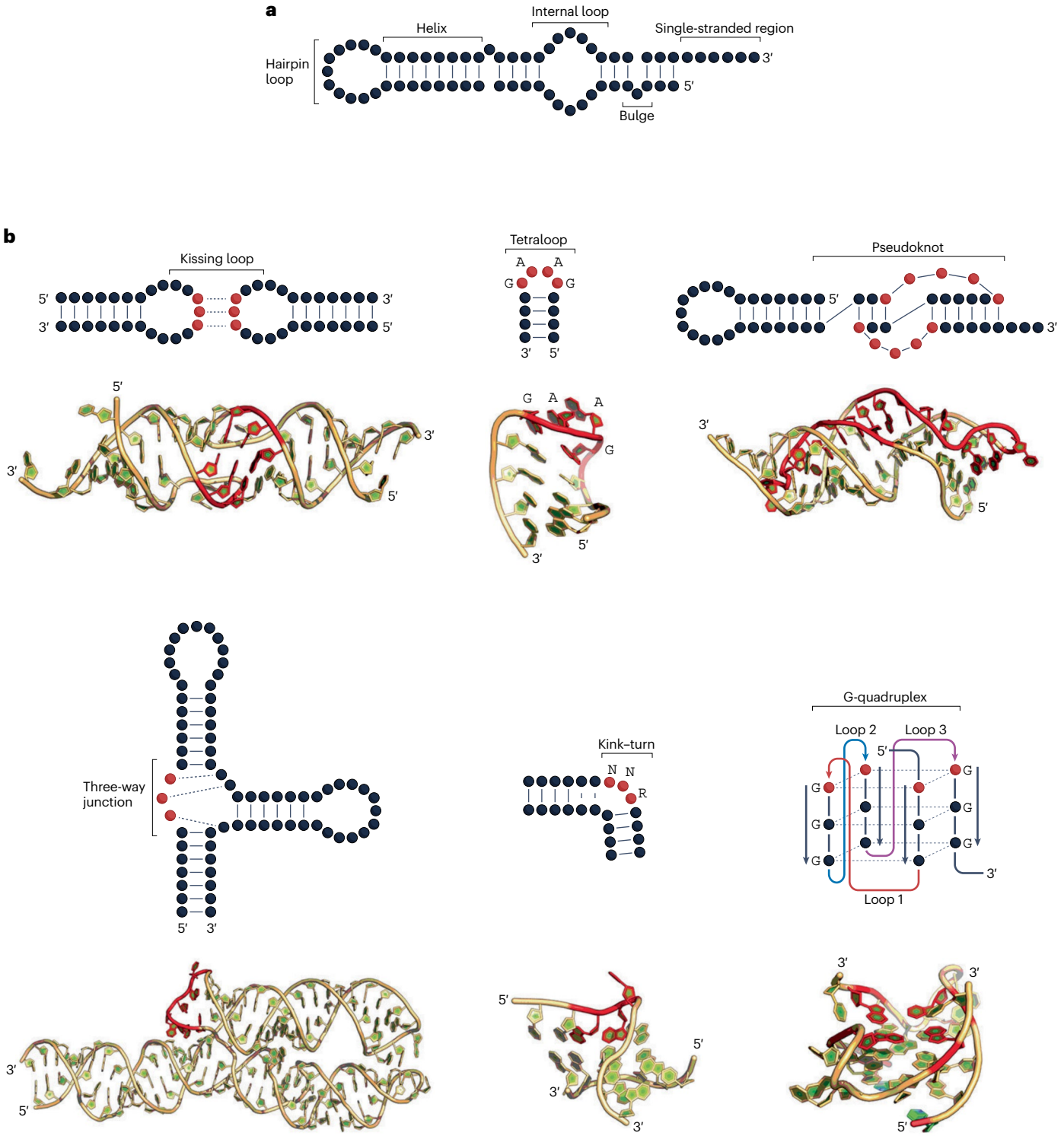
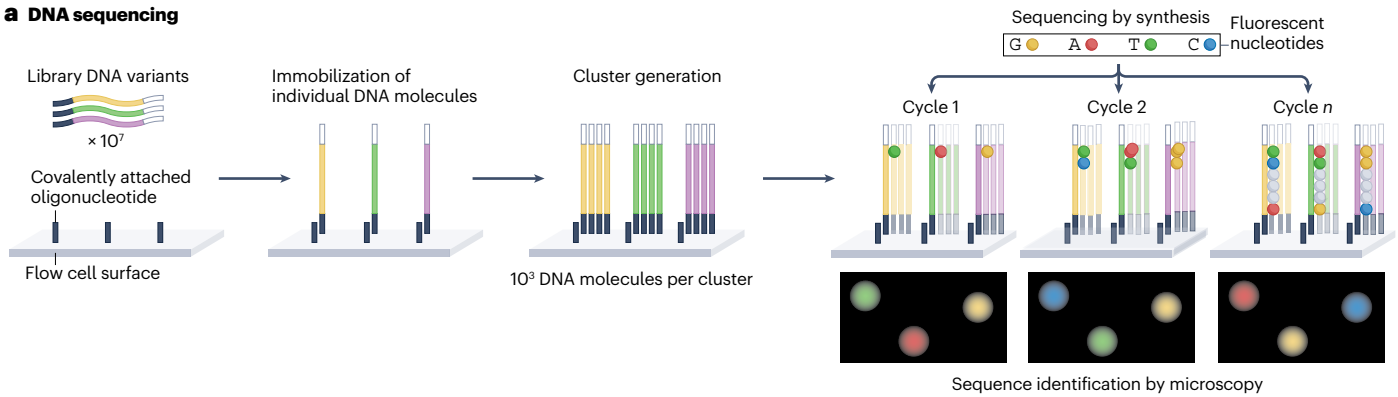
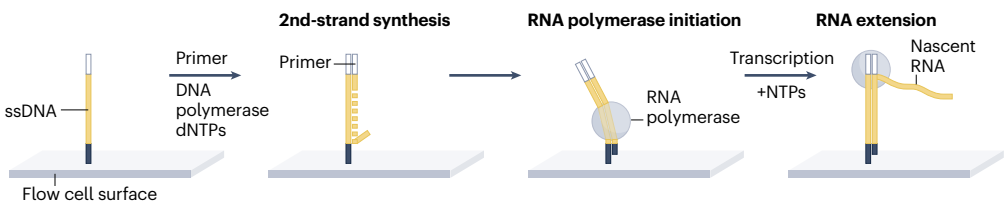
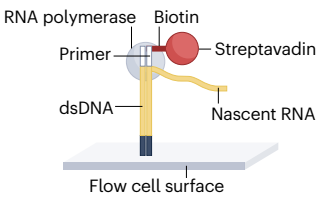
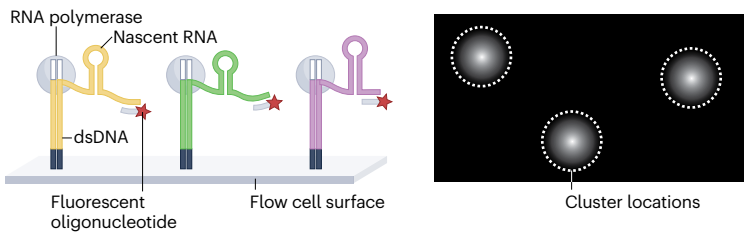


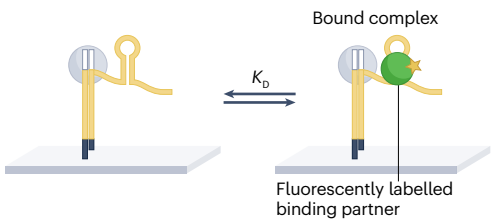
Fig. 1 | Diversity of RNA secondary and tertiary structures. a, Schematic of common RNA secondary structure motifs, comprising hairpin loops, helices, internal loops, bulges and single-stranded regions. **b**, Examples of RNA secondary structure motifs and the corresponding tertiary structure motifs from crystal structures: 'kissing loop' structure of dimeric HIV-1 RNAs (Protein Data Bank (PDB) ID: 1K9W)[103], tetraloop structure of signal recognition particle (SRP) RNA from *Pyrococcus furiosus* (PDB ID: 2F87)[104], pseudoknot structure of human telomerase RNA (PDB ID: 2K96)[105], three-way junction formed by human 7SL RNA (PDB ID: 1MFQ)[106], kink–turn structure of SAM-I riboswitch RNA from *Thermoanaerobacter tengcongensis* (PDB ID: 3IQN)[107] and G-quadruplex structure of human telomeric RNA (PDB ID: 3IBK)[108]. Red nucleotides highlight the distinct structural motifs. Reprinted from ref. [7], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

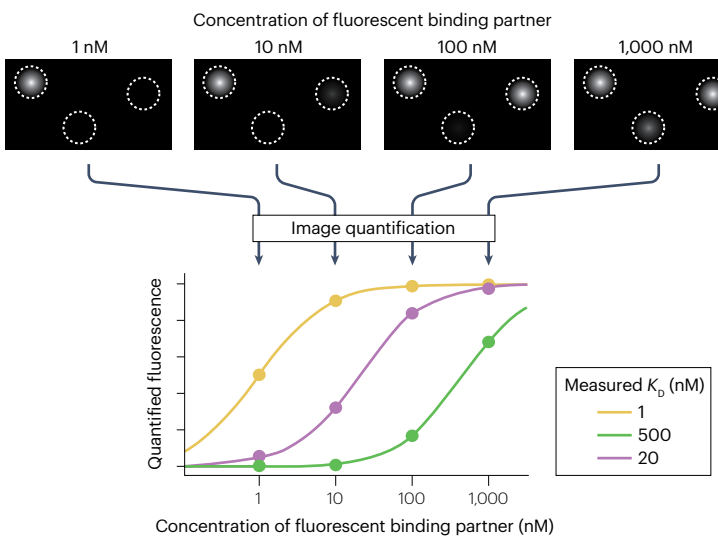

a DNA sequencing
Library DNA variants
× 10^7
Covalently attached oligonucleotide
Flow cell surface
Immobilization of individual DNA molecules
Cluster generation
10^3 DNA molecules per cluster
Sequencing by synthesis
G A T C
Fluorescent nucleotides
Cycle 1
Cycle 2
Cycle n
Sequence identification by microscopy


b RNA generation
ssDNA
Primer
DNA polymerase dNTPs
Flow cell surface
2nd-strand synthesis
Primer
RNA polymerase initiation
RNA polymerase
Transcription
+NTPs
RNA extension
Nascent RNA


c RNA polymerase stalling
RNA polymerase
Biotin
Primer
Streptavadin
dsDNA
Nascent RNA
Flow cell surface


d Cluster localization
RNA polymerase
Nascent RNA
dsDNA
Fluorescent oligonucleotide
Flow cell surface
Cluster locations


e Binding measurements
Bound complex
K_D
Fluorescently labelled binding partner


f Equilibrium binding
Concentration of fluorescent binding partner
1 nM
10 nM
100 nM
1,000 nM
Image quantification
Quantified fluorescence
1
10
100
1,000
Concentration of fluorescent binding partner (nM)
Measured K_D (nM)
1
500
20

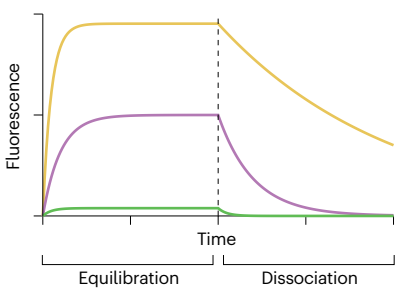

g Kinetic association measurements
Fluorescence
Time
Equilibration
Dissociation

Fig. 2 | Using next-generation sequencing chips for high-throughput biochemical measurements. a, Schematic of DNA sequencing using the Illumina platform, whereby the nucleotide sequence identity of all molecular variants in a DNA library is determined through sequencing by synthesis. **b**, Generation of RNA from a library of DNA clusters. The single-stranded DNA (ssDNA) on the array surface is made into double-stranded DNA (dsDNA) by annealing a primer and extending the complementary strand with a DNA polymerase. This is followed by transcription of nascent RNA from dsDNA by RNA polymerase. **c**, Stalling of the RNA polymerase by using a streptavidin 'roadblock', which binds a biotin present on the 5′ edge of the primer annealed in part **b**, is used to attach the RNA transcript to the flow cell surface. **d**, To assist in mapping between sequencing and binding data images, generated RNAs can be labelled with a fluorescent oligonucleotide complementary to a sequence that is present in all members of the RNA library. **e**, Binding measurements are carried out by introducing a fluorescently labelled binding partner into the flow cell and recording changes in fluorescence intensity at the RNA clusters. **f**, Equilibrium binding measurements for different RNA clusters can be obtained by equilibrating the flow cell with increasing concentrations of the binding partner to determine effective equilibrium dissociation constants (K_D). **g**, Kinetic association measurements can be obtained by measuring cluster intensities during equilibration (left of dashed line), and dissociation measurements can be obtained by flowing in buffer without binding partner after equilibration (right of dashed line). Line colours in parts **f** and **g** correspond to the DNA and RNA sequence variants shown in parts **a** and **d**. Parts **a**–**f** reprinted with permission from ref. [45], Cold Spring Harbor Laboratory Press.

stall the RNA polymerase and tether the RNA to the DNA cluster. For example, streptavidin, which binds a biotin present on the 5′ edge of the primer annealed in the previous step, can be introduced into the flow cell as the roadblock[28] (Fig. 2c). Alternatively, the roadblock can be the *Escherichia coli* replication terminator protein Tus[27].

The biochemical measurement is then carried out using a fluorescence microscope. In the early days of the technology, measurements were carried out directly on modified Illumina Genome Analyzer IIx instruments[27,28,33,34], which are now no longer supported. These have mostly been replaced by sequencing on Illumina MiSeq[30,31,35–42], followed by biochemical measurements on custom-built[30,31,35–42] and commercial[43,44] total internal reflection fluorescence microscopes, with custom-built fluidic adaptors to control flow and introduce medium into the flow cell. To assist in the mapping between sequencing and binding data image positions, generated RNAs can be labelled with a fluorescent oligonucleotide complementary to a sequence that is present in all members of the RNA library (Fig. 2d). Biophysical changes are recorded through changes in fluorescence intensity at another wavelength by, for example, flowing in a fluorescently labelled binding partner over the RNA clusters (Fig. 2e). The identities of the RNA clusters that bind the fluorescent partner are then assigned by mapping the location of the clusters from the binding measurements to those from the DNA sequencing using an image registration process such as hierarchical cross-correlation[45]. The fluorescence intensity can then be quantified by, for example, fitting and integrating a two-dimensional Gaussian distribution to each RNA cluster in the images[45]. Generally, unique members of the DNA library are represented multiple times on the array, both to minimize experimental noise and to quantify the technical errors of the measurement. The optimal number of clusters measured per molecular variant is, in practice, between 10 and 100 depending on the signal-to-noise ratio of the experimental set-up, whereby brighter fluorophores and stronger binders require fewer measurements per variant.

Measuring thermodynamic and kinetic parameters

Equilibrium measurements of fluorescent analytes binding to the DNA or RNA variants presented on the array can be carried out by flowing in different concentrations of the fluorescent binding partner, letting the binding equilibrate and recording images at each concentration. Hill equation binding curves fitted to the quantified fluorescence can be used to determine effective equilibrium dissociation constants (K_D) (Fig. 2f), and the equation $\Delta G = RT \log(K_D)$ gives effective binding free energies. (R is the universal gas constant; T is temperature (K)). Correspondingly, binding kinetics can be directly measured by continuously recording images while the binding reaction is equilibrating, which can be used to produce association curves by plotting the cluster-averaged fluorescence intensity for each molecular variant against time. Similarly, dissociation curves can be acquired by flowing in buffer without the binding partner over an already-equilibrated array (Fig. 2g). Effective association rate constants (k_a) and effective dissociation rate constants (k_d) can be determined either by the initial slopes of the binding curves or by fitting exponentials to the binding curves. This approach also gives a redundant measure of the effective equilibrium dissociation constant ($K_D = k_d/k_a$), which can be compared with the K_D obtained from the equilibrium measurement to check for consistency with an exponential binding model in which one target binds one probe.

Previous applications of RNA arrays

A grand challenge in molecular biophysics is the creation of predictive models that link RNA sequence to relevant thermodynamic and kinetics parameters expected from intermolecular or intramolecular binding (Fig. 3a). Such models would provide a means to predict the physical parameters of binding for molecular RNA variants that were previously unobserved experimentally. Furthermore, such models would ideally also provide deeper mechanistic insights into the sequence determinants of RNA binding by defining the intramolecular properties that affect the interaction (Fig. 3b). So far, models aiming to describe high-throughput datasets from RNA arrays have been constructed using a bottom-up approach, based on hypothesized expectations of how RNA interactions might work at the molecular level. In the following sections, we describe some important examples of these experimental applications and models.

RNA–RNA interactions

Many RNAs fold into complex tertiary structures that are essential for their biological functions[4,46–49]. Common methods for studying RNA tertiary structure, such as X-ray crystallography, NMR spectroscopy and cryo-electron microscopy, have limitations in terms of throughput and resolution; RNA array technology has the potential to address these methodological gaps. The formation of RNA tertiary structures has been studied by RNA array using tectoRNAs as a model system[30,35,41]. The tectoRNA model system consists of two folded RNAs with well-defined secondary structures — the variable 'chip piece' RNA that is presented on the array and a fluorescently labelled, constant 'flow piece' RNA that is flowed onto the sequencing chip — that bind each other to form a heterodimer (Fig. 4a). The binding free energy (ΔG) of a mutant library of chip piece RNAs for the constant flow piece RNA can then be determined.

A structure-based model was developed that randomly samples the conformation of each base pair step (two sequential base pairs) in a tectoRNA based on its relative occurrence in structured RNAs in the RNA crystal structure database[30] (Fig. 4b). The distance between the binding sites of the chip piece and flow piece tectoRNAs was then calculated for each sampled structure, and binding free energies were estimated using a fixed-distance cut-off to assign whether the RNAs were bound for each structure in the sampled library of chip piece RNAs (Fig. 4c). These predicted binding energies agreed well with the experimentally determined binding energies[30], demonstrating that accurate binding energies can be inferred when the variety of structural conformations is known. A future hope for RNA array technology is that the opposite inference can also be made — namely, that the RNA conformations present and their relative abundances can be predicted from binding energies using models trained on the RNA array data. This extension will almost certainly require several intramolecular properties and distance measures to be probed in these high-throughput experiments.

RNA–protein interactions

Physical interactions between RNA and RNA-binding proteins (RBPs) are crucial for post-transcriptional gene regulation[50], and these interactions are controlled by the sequence-specific binding energies of the RNA–RBP interactions. Several RBPs have been studied using RNA arrays[27,28,33,36,51]. Early studies looked at binding of green fluorescent protein (GFP) and negative elongation factor subunit E (NELF-E) to RNA aptamers[27,33], and of the coat protein of MS2 bacteriophage to RNA hairpins[28]. In these studies, an RNA mutant library is presented on the array and a fluorescently labelled RBP is flowed over the array. For NELF-E binding to RNA aptamers with two mutations, binding energies were accurately predicted as the sum of binding energies for the corresponding single mutants[27]. This was not the case for GFP binding to RNA, which showed a more complex contribution of individual RNA mutations to the total binding energy[27]. In the case of MS2 coat protein binding to RNA hairpins, binding data from an RNA array were used to train a model in which specific sequence features in the RNA hairpin — namely, base transversions, base transitions, loss of base pairing and non-canonical base pairing — had additive effects on binding energy. When trained on data from RNA hairpins having single mutations from the consensus sequence, the model could accurately predict the measured binding energies of RNA hairpins with two or three mutations[28]. Related modelling approaches have also proven successful for predicting the binding energies of the RNA-binding Pumilio proteins PUM1[36], PUM2[36] and PUF4[51], when trained and tested on RNA array data. In these studies, additive energy parameters were used for each RNA nucleotide at each position in the 9-base-long site that recognizes these proteins (additive consecutive model). The models gave more accurate predictions when energy terms were also added for 'flipping' individual bases out of the recognition site in RNA, as the protein does not always bind nine consecutive bases (additive nonconsecutive model), and also when coupling terms were added for a few neighbouring RNA positions that influence each other to contribute non-additively to the binding (additive nonconsecutive and coupling model)[36,51] (Fig. 5a). In sum, the RNA array technology has allowed for predictive mapping between RNA sequence and RBP binding free energy that might enable a better understanding of RBP biology and the engineering of new RNA–RBP binding pairs.

RNA-guided protein interactions

Nucleic acid-guided binding systems such as CRISPR–Cas have revolutionized genome measurement and manipulation in living organisms[52]. Accurate predictive models for RNA-guided protein interactions would thus enable the accurate prediction of on-target and off-target binding, allowing for improved sensitivity and specificity of these methods. To this end, several array studies have investigated proteins that bind other nucleic acid sequences through base pairing with a guide nucleic acid[32,38,43,44,53,54]. In these studies, single-stranded RNA, ssDNA or double-stranded DNA (dsDNA) was presented on the array, depending on the target of the protein of interest. Binding affinities have been measured for guide-loaded CRISPR proteins, including Cascade[43], Cas3[43] and different engineered variants of Cas9 and Cas12a[44], to off-target libraries of dsDNA on sequencing chips. Furthermore, binding and unbinding kinetics have been measured for catalytically inactive Cas9 (dCas9)[53]. As dCas9 has been observed to bind a greater number of off-target sites than are cleaved by Cas9[55], separating the DNA sequence determinants of binding from the determinants of cleavage across a large sequence space is a crucial challenge[44]. In this regard, one study compared the binding affinities of Cas9 measured on a sequencing array with the cleavage rates measured by NucleaSeq, a method that sequences the cleavage products of a DNA library at different time points to estimate the rate of cleavage of the library members[44]. Interestingly, this study found that the tested engineered Cas9s had much higher cleavage specificity than wild-type Cas9, but similar binding specificity. Modelling efforts for CRISPR proteins have included empirical models of binding affinity[43] and cleavage specificitity[44] involving additive effects of specific sequence features. Furthermore, mechanistic kinetic models

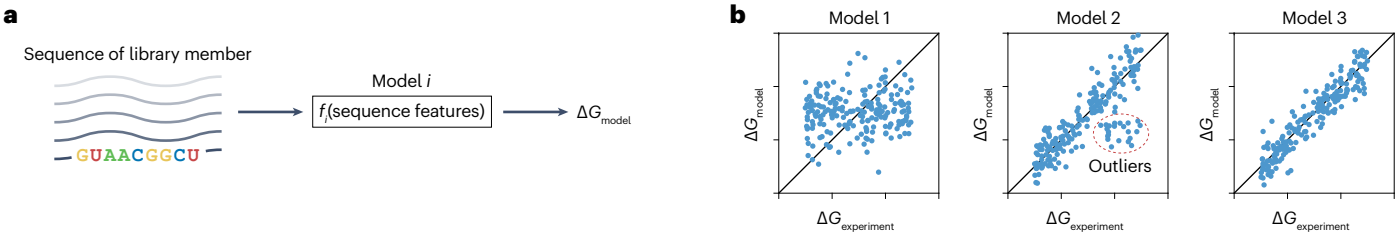


Fig. 3 | Predicting experimental parameters from RNA sequence information. a, Schematic of a model taking RNA sequence information as the input. After being trained on experimental data, the model outputs a predicted binding energy (ΔG_{model}) given a primary sequence of RNA. **b**, Correlation plots between experimental ($\Delta G_{experiment}$) and predicted (ΔG_{model}) binding energies for three hypothetical models with simulated data. The predicted binding energies of model 1 have poor correlation with the experimental data and this model can thus be rejected. Model 2 shows better correlation with the experimental data, with the exception of some outliers. Model 2 could be fine-tuned to get model 3, which shows the best agreement with the experimental data, before reporting the performance of model 3 on a test set of RNA sequences.

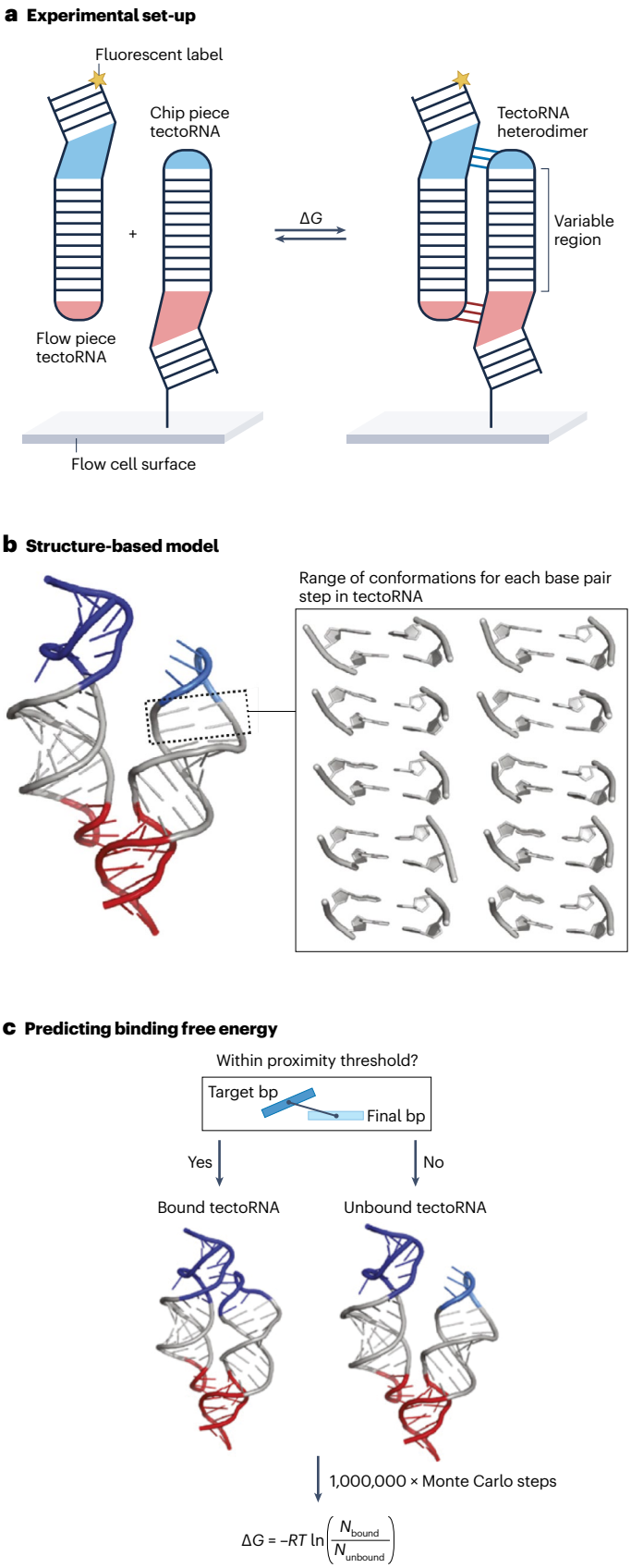


Fig. 4 | Structural modelling to predict RNA–RNA binding energies verified with RNA array binding data. a, Schematic of the formation of a tectoRNA heterodimeric complex on the RNA array. A variable 'chip piece' tectoRNA that is presented on the array binds a fluorescently labelled, constant 'flow piece' tectoRNA that is flowed onto the sequencing chip. **b**, A structure-based model randomly samples the conformation of each base pair step (two sequential base pairs) in a tectoRNA based on its relative occurrence in structured RNAs in the RNA crystal structure database. **c**, Binding free energies (ΔG) are predicted from this model, using a distance cut-off to assign whether sampled tectoRNA conformations are bound or unbound. The distance used for assigning binding is calculated between a target base pair (target bp) in the flow piece tectoRNA and the final base pair (final bp), on the molecule edge, in the tetra loop of the chip piece tectoRNA. N_{bound}, number of bound conformations; $N_{unbound}$, number of unbound conformations; R, universal gas constant; T, temperature (K). Adapted with permission from ref. [30], National Academy of Sciences.

have been developed that explicitly handle the base pair transitions that occur during guide RNA strand invasion of the target DNA[56,57] (Fig. 5b). These kinetic models have been used to determine in which order the nucleotides in the target DNA are bound during the formation of the specific Cas9-bound complex[56] and to deduce the free-energy landscape on the reaction path to the bound conformation[57] (Fig. 5c), and they can accurately predict experimental binding affinities and cleavage rates[57].

Binding interactions of two Argonaute-family proteins have been studied on sequencing chips: the RNA-loaded, RNA-binding eukaryotic protein Ago2[38] and the DNA-loaded, ssDNA-binding bacterial protein TtAgo[54]. In the case of Ago2, binding measurements on the RNA array were combined with high-throughput cleavage measurements using a method based on reverse transcription, PCR and sequencing. Binding energies and cleavage rates were modelled using two linear models with additive parameters for certain sequence position features. These models captured ~60% of the variability in binding energies and ~70% of the variability in cleavage rates of the test datasets[38]. Furthermore, the authors constructed a simple model of ordinary differential equations, parameterized with the in vitro-determined biochemical parameters, that could predict the degree of knockdown of RNA targets in live cells for RNAs with a constant sequence context around the target binding site[38]. In the case of TtAgo, DNA cleavage rates were measured directly on the sequencing chip after the binding measurement, whereby cleavage was detected as a loss of fluorescence when the end-labelled binding site of DNA was cleaved off[54]. Models for association rates and binding energies were constructed using secondary structure predictions and energy estimates for the guide DNA seed region using Nucleic Acid Package (NUPACK) software[58]. These models described and predicted the same large-scale variations in the data as were observed experimentally[54].

RNA–small molecule interactions

Structured RNAs can interact with small molecules to function as sensors, resulting in, for example, RNA cleavage[59] or transcription termination[60]. To this end, RNA arrays have also been used to study the binding and cleavage of RNA induced by small molecule ligands[40,42]. One study combined an RNA array with crowdsourced RNA design, whereby an internet community proposed RNA aptamers to bind certain small ligands (the 'inputs'), including flavin mononucleotide, theophylline and L-tryptophan. These RNAs were engineered to change their secondary structure upon detection of these input ligands, revealing binding

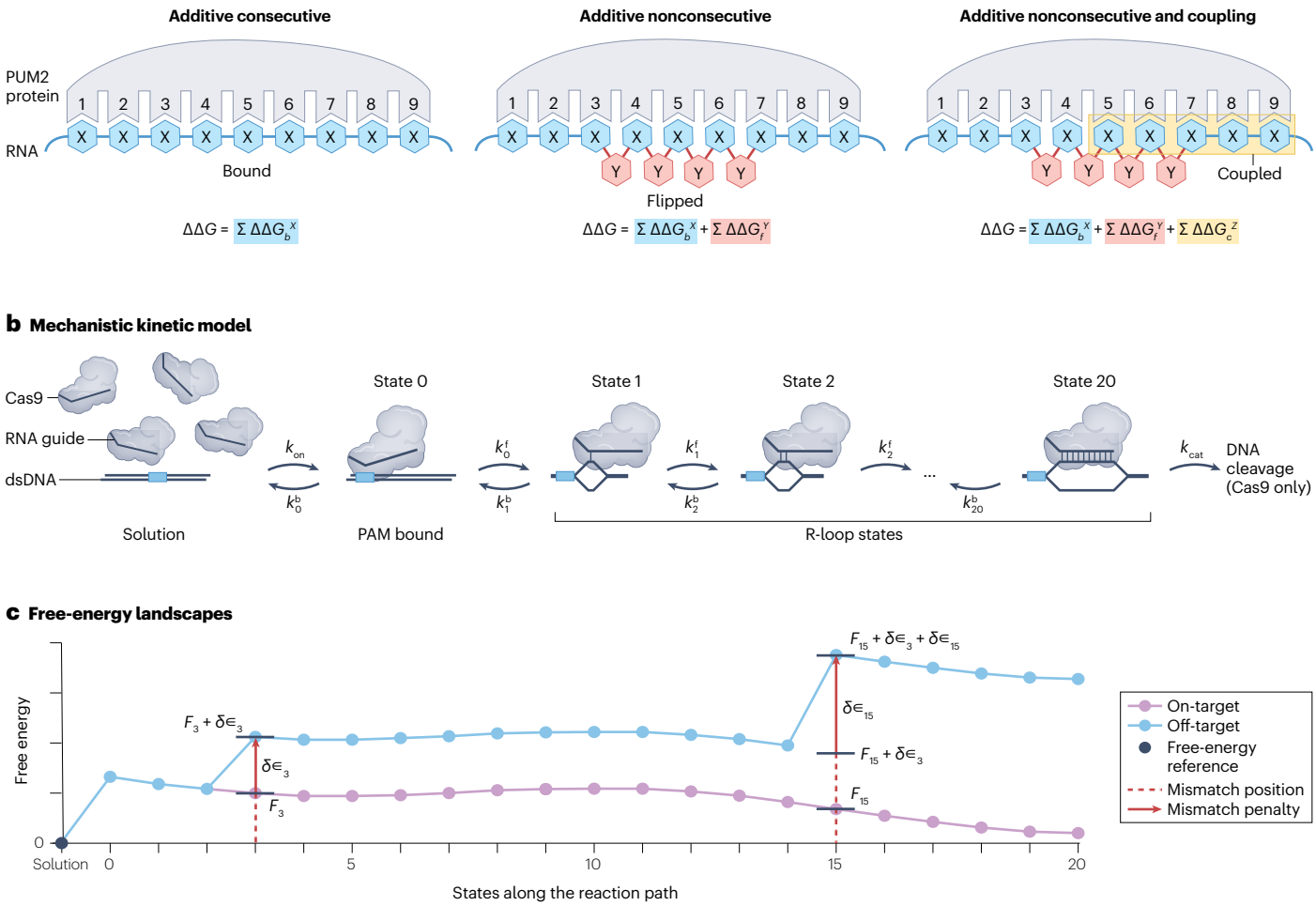


Fig. 5 | Thermodynamic and kinetic models for RNA–protein interaction and RNA-guided protein binding. a, Three additive energy models — in which specific RNA sequence features have additive effects on protein binding — have been used to predict the binding energy of the RNA-binding Pumilio protein PUM2[36]. Energy terms can be added for 'flipping' individual bases out of the recognition site in RNA, as the protein does not always bind nine consecutive bases, and coupling terms can be added for neighbouring RNA positions that influence each other to contribute non-additively to binding. From left to right, the models show improved predictions of experimental energies when trained and tested on binding data from RNA arrays. **b**, Schematic of a mechanistic kinetic model that describes DNA binding and cleavage by an RNA-guided Cas9 protein. The model consists of a free DNA-dissociated state, a protospacer adjacent motif (PAM)-bound state (state 0) and 20 nucleotide-bound states (states 1–20) corresponding to each base pair interaction between the RNA guide and the DNA target. The states in the model are connected sequentially, and the model can transition towards the fully bound state with rates k_i^f and towards the free DNA-dissociated state with rates k_i^b. Bimolecular association between Cas9 and the PAM is parameterized by the rate k_{on}. In the fully bound state, when all 20 base pairs of the R-loop are formed between the RNA guide and the DNA target, Cas9 can cleave DNA with rate k_{cat}. **c**, Free-energy landscapes implied by the fitted kinetic model are shown. The on-target (pink) free-energy landscape (F_n) of Cas9 is compared with an off-target (blue) free-energy landscape, where mismatches have been introduced at nucleotide positions 3 and 15 of the RNA guide. Each mismatch has an energetic cost ($\delta\epsilon_n$; the mismatch penalty) added onto the free energy of all later R-loop states. The kinetic model implies that all mismatch penalties are additive, so that, for example, the off-target free energy at nucleotide position 15 is equal to $F_{15} + \delta\epsilon_3 + \delta\epsilon_{15}$. dsDNA, double-stranded DNA; $\Delta\Delta G_b^X$, free energy terms for bound bases; $\Delta\Delta G_f^Y$, free energy terms for flipped-out bases; $\Delta\Delta G_c^Z$, free energy terms for coupled bases; X, bound base; Y, flipped-out base. Parts **b** and **c** reprinted from ref. [57], Springer Nature Limited.

sites for 'output' signalling molecules such as fluorescently labelled MS2 coat protein, malachite green or the GFP mimic DFHBI. On the RNA array, binding of these output signalling molecules could be detected as fluorescence (malachite green and DFHBI fluorescence become much brighter when bound to their RNA aptamers). For several of the input–output pairs, just one round of design and testing resulted in engineered RNA switches with a tenfold greater affinity for the output molecule when the input ligand was present. Iterative rounds of design and testing of the flavin mononucleotide–MS2 coat protein input–output pair produced RNAs that approached the thermodynamic optimum of a perfect molecular switch[42]. The glucosamine-6-phosphate (GlcN6P) riboswitch (glmS) ribozyme, an RNA molecule that undergoes a self-cleavage reaction upon binding of the specific ligand GlcN6P, has also been the subject of high-throughput investigation by RNA array[40]. GlcN6P-induced

self-cleavage of the ribozyme was detected as a loss of fluorescence of the labelled 5′ end of the ribozyme RNA. By fitting a Michaelis–Menten-type model to these cleavage data, the authors discovered that most point mutations of glmS only caused a small change in binding affinity for GlcN6P, whereas the variability in cleavage rates accounted for the majority of the overall catalytic differences observed. Naturally occurring ribozyme mutations generally maintain a high cleavage rate, which suggests that this biochemical parameter is conserved in evolution.

Further development of RNA arrays

In the current state of the art, several methodological challenges remain for high-throughput biophysical measurements in vitro. Further developments and improvements are possible and likely — namely, probing of intramolecular interactions at the single-molecule level, improving the scalability and throughput of the methods, and making the methods more widely available.

Probing intramolecular states by FRET

Fluorescence resonance energy transfer (FRET)[61,62] and fluorescence quenching are natural mechanisms of signal generation downstream of molecular conformational changes that would be useful on the RNA array. These techniques use an appropriately placed donor and acceptor fluorophore pair to report on end-to-end distance changes that might correspond to folded or unfolded states of an RNA molecule (Fig. 6a). This signalling mechanism may also allow for the measurement of melt curves for RNA structures by changing the temperature of the RNA array and measuring changes to the distance-dependent fluorescence signals as RNA structures melt. These energy transfer-based approaches can also be used to probe more than just two-state behaviour, giving a measure of the distance between two labelled points of a molecule[63–65]. For RNAs on an RNA array, a molecular ruler might be created by introducing donor and acceptor fluorophores at different locations in a molecule with known and stable structure, such as

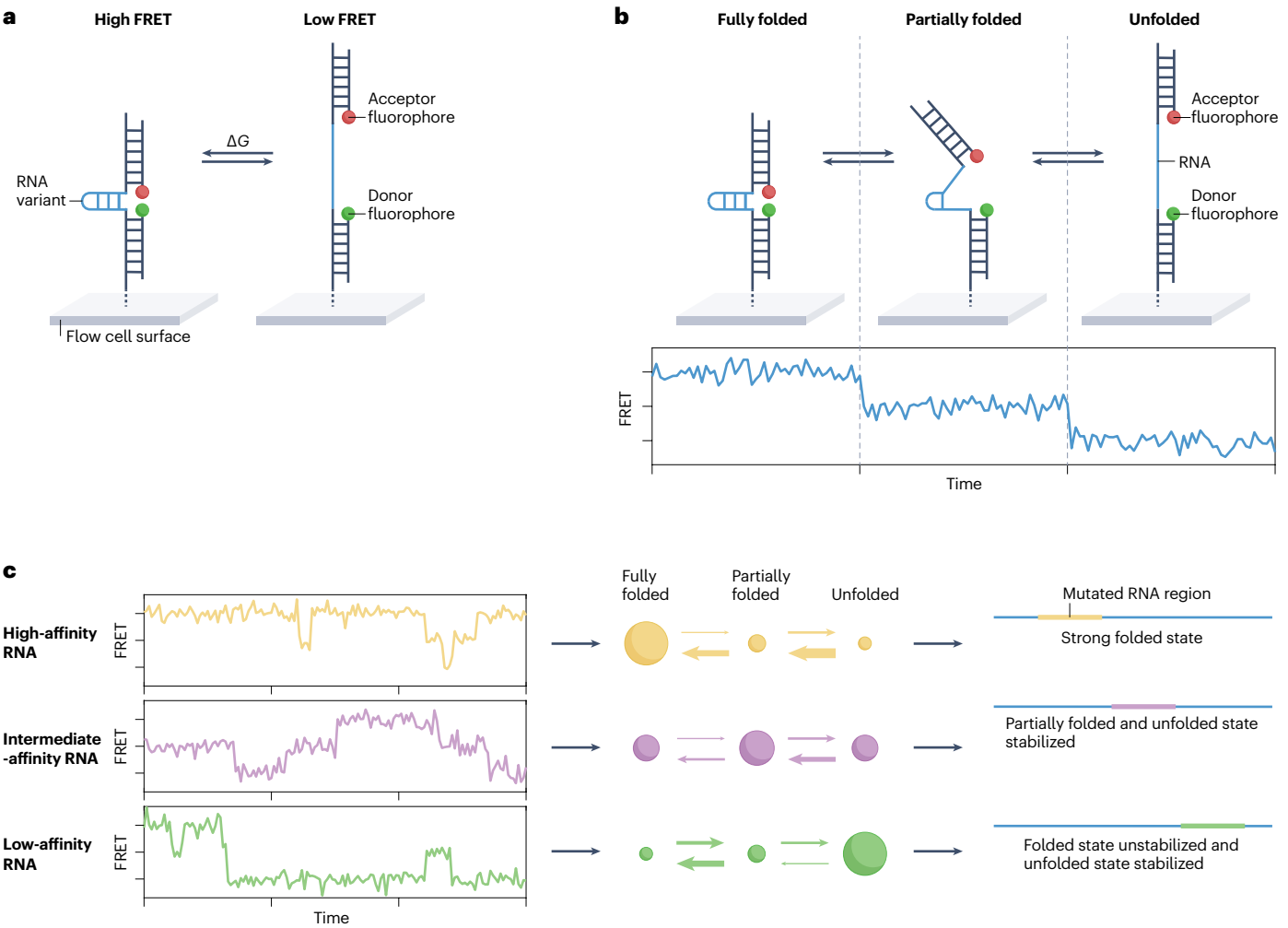


Fig. 6 | Single-molecule experiments carried out across a large sequence space. a, Schematic of how fluorescence resonance energy transfer (FRET) donor (green) and acceptor (red) fluorophores could be used to probe intramolecular binding of RNA library members (blue). **b**, Schematic of a hypothetical RNA with three dynamic conformational states, probed by FRET (top). A simulated FRET signal for a single RNA molecule going from the fully folded state to the partially folded state and then to the unfolded state is shown (bottom). **c**, Simulated FRET traces of different RNA sequences of differing intramolecular affinities in a hypothetical library (left) can be used to fit hidden Markov models for assignment of state occupancies and transition rates (middle). These fitted models can be used to gain sequence-dependent insights and build sequence-dependent models (right). Δ*G*, the change in free energy from the unfolded state to the folded state.

dsDNA, which can then be used to calibrate the fluorescence signal to average molecular distance. Site-specific labelling of RNA at variable nucleotide positions could be achieved on the array by the stepwise and site-specific stalling of transcription through omission of different nucleotides in the transcription reaction, changing which sets of nucleotides are omitted at the different steps and introducing an azide-modified nucleotide at the step corresponding to the RNA position that is to be labelled[66]. The azide-modified nucleotide can then be labelled with a fluorophore of choice using click chemistry[67]. Varying the donor and acceptor fluorophore locations in the molecular complex might provide further constraints on three-dimensional structure. Comparing the constraints from RNA array data with structures obtained by classical methods and results by molecular dynamic simulations are likely to help elucidate the intramolecular states of RNA.

Single-molecule measurements

Single-molecule measurements[63,68–70] have revolutionized our understanding of the molecular ensemble. By following signals from individual molecules over time, we can resolve the lifetimes and identities of underlying intramolecular states, as well as the transition probabilities between these states (Fig. 6b), which provides a full picture of the intramolecular kinetics of a molecular species (Fig. 6c). High-throughput, single-molecule sequencing by synthesis combined with fluorescence-based detection of intermolecular interactions has previously been used to study nucleosome modifications in a DNA library of the mouse genome[71]. However, biophysical measurements on commercial sequencing chips have largely, so far, been carried out in bulk mode, as all clusters on the array give an integrated signal aggregated from the binding to ~1,000 identical target molecules. In principle, single-molecule experiments might be implemented on a sequencing chip by carrying out the biophysical measurements before sequencing, such that the individual library molecules have not yet been made into clusters at the time of measurement, or by taking biophysical measurements after amplification and sequencing using a sufficiently low concentration of fluorescently labelled analyte to enable detection of single-molecule binding[72]. Challenges to implementing this strategy include the general challenges of nearly all single-molecule fluorescence experiments, including weak fluorescence signals emanating from single fluorophores that necessitate the use of high numerical aperture objectives and sensitive cameras. Challenges unique to implementation on a sequencing chip include the inherent optical properties of the flow cell (which cannot be easily modified) as well as the linkage of specific sequences to the locations of these molecular variants on the chip[72].

The time resolution of the dynamics of single-molecule RNA structural variation is fundamentally limited by the detectors used for acquiring fluorescence data. Current implementations of the RNA array technology involve relatively standard, laser-excited, wide-field imaging methods using non-amplified charge-coupled device cameras, which give a time resolution in the range of 10 milliseconds to seconds, at best. This limits the potential use of these single-molecule approaches for kinetic studies to the dynamics of RNA tertiary structures, whereas secondary structure dynamics that occur on faster time scales[2] would require faster detectors. Faster detection might be possible using scanning confocal microscopy followed by fluorescence correlation spectroscopic analysis of the data[73,74]. The detectors used can count the arrival of photons with nanosecond accuracy[74], in which case the time resolution of the measurement will no longer be limited by the detector but by the number of photons emitted by the fluorophore per unit of time, which currently is ~1 photon per 100 µs for commonly used organic fluorophores at high excitation power[75,76]. However, scanning an entire sequencing chip with a confocal microscope will take longer than wide-field imaging, for which imaging of different experimental conditions already takes several hours. This means that trade-offs will have to be made between time resolution, noise level in the signal output, the total time to carry out an experiment and the throughput of molecular variants.

Scaling-up the throughput of sequencing chips

The throughput of an RNA array is fundamentally limited by the number of DNA clusters that can be sequenced on any given sequencing technology platform. Most recent applications of array technology use Illumina MiSeq for sequencing, which has a throughput in the order of 10^7, giving a throughput of 10^5–10^6 unique library members per experiment when accounting for 10 to 100 repeat clusters per molecular variant. A natural progression of the technology is to use sequencers with increased throughput. Contemporary sequencing platforms can achieve higher throughputs — for example, from 10^9 for Illumina NextSeq to 2×10^{10} for Illumina NovaSeq.

Distributable software and hardware

For these high-throughput methods to have maximum impact in the fields of biochemistry and biophysics, they must be accessible to all laboratories that wish to use them. To achieve this, several outstanding challenges remain in terms of the distribution and automation of hardware and optics, software for control of hardware, know-how for library construction, protocols for executing experiments and analysis tools for interpretation of raw data. Ideally, all of the required hardware — such as the fluorescence microscope, fluidics system and computers controlling these units — would be assembled in a standardized way, and perhaps even sold as a single product together with the software necessary to control the hardware and run experiments. The instrument could integrate sequencing by synthesis or could take a flow cell from a separate sequencer as the input, a strategy that might be more straightforward to implement. The hope is that the experimental procedure can be made heavily automated and user friendly, similarly to what has already been achieved in commercial products for next-generation sequencing and for low-throughput binding measurements on biosensors.

In addition to standardized instrumentation, users of these methods would also need specific knowledge to construct the libraries and execute experimental protocols. In the simplest experimental scenario, library construction involves PCR amplification of an ordered DNA library with Illumina sequencing adaptors on the DNA ends. However, library construction is sometimes more complex and involves 'bottlenecking' as a means to couple each DNA sequence of interest to a unique barcode that is to be read out during sequencing[35]. In terms of the experimental protocol, variations are required to present either DNA or RNA on the array, or to carry out either equilibrium thermodynamics or kinetic measurements. All of this information should be compiled in an easily accessible protocol document, and common variations of the experimental protocol should be easily executable through a simple graphical user interface. Furthermore, streamlined tools for downstream image analysis should be provided. As cluster identification and fluorescence quantification are necessary steps in all variations of the experimental set-up, tools for these steps are already relatively standardized. These upstream analysis steps would ideally be automated and optimized to run in real-time on the hardware-control computer as it acquires data. However, as image acquisition is currently at least ten times faster than image analysis on one central processing

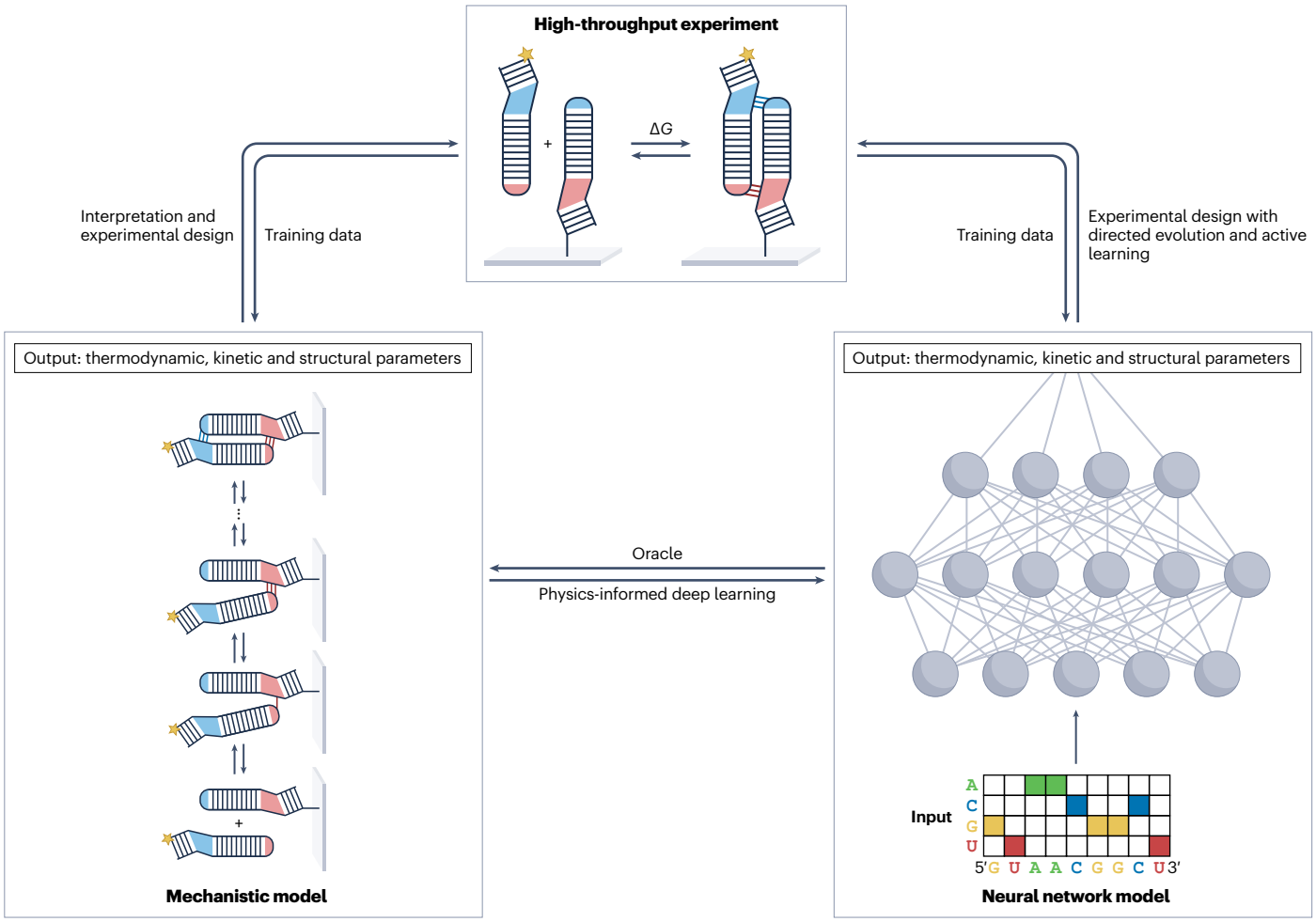


Fig. 7 | Feedback between experimental data, mechanistic modelling and deep learning methods. Schematic of how physical mechanistic models and neural network models can be used and combined to design, interpret and model experiments on RNA arrays. Arrows between the different boxes (nodes) represent the flow of information whereby, for example, the data coming from a high-throughput experiment on an RNA array (top node) can be used to train a neural network model (right node). The trained neural network model can in turn be used as an oracle to carry out experiments in silico with RNA sequences not present in the training data, and the outputs from the in silico experiments can be used to train and build a mechanistic model (left node). ΔG, binding free energy.

unit, images are generally uploaded to a centralized remote cluster for later analysis.

In sum, although the RNA array experimental procedure involves many steps and requires specific instruments, there is great potential for automation of the protocols and the development of a commercial or academically standardized product. When thinking about the future impact of this method, it is easy to draw parallels to the emergence of second-generation DNA sequencing techniques during the 1990s, which since then have replaced first-generation sequencing techniques for many applications. If the DNA, RNA and protein array technology can be made readily available, easy to use and inexpensive, it might provide a new standard for biochemical measurements.

Quantitative predictions from RNA arrays

Even though the data coming from RNA arrays are already being used to build and evaluate quantitative models of RNA structure and function[30,36,77], there is room for improvement in terms of making these models even more predictive and better at elucidating the underlying physical mechanisms of the molecular interactions. To achieve this, we anticipate that a combination of deep learning models, optimization algorithms such as directed evolution and active learning, and more detailed physical mechanistic models will be required.

Deep learning

Accurate prediction of protein structure enabled by deep learning algorithms such as AlphaFold[78] and RoseTTAFold[79] represents a major breakthrough in structural biology. Deep learning applied to RNA structure and function is also an active field of research in which progress has been made in the prediction of RNA secondary structure[80–86], RNA tertiary structure[87] and RNA–protein binding[88–91]. However, several problems remain to be solved in the field of RNA structure prediction. Current implementations of deep learning models for the prediction of RNA tertiary structure rely on high-resolution structural data as deposited in the Protein Data Bank[13,80,90]. The quality and applicability

of such models is thus limited both by the availability of data and by the fact that classically obtained structures do not capture the dynamic nature of RNA. We believe that a high-throughput method such as the RNA array is an exciting possible way to solve these issues, whereby deep learning methods could be trained on data coming from these high-throughput experiments. Deep learning models trained with a physics-informed machine learning approach[92], whereby the model is constrained to adhere to different underlying physical mechanisms or principles, seem to be particularly promising. Deep learning models can also be used as 'oracles' by carrying out in silico experiments with a trained deep learning model[93] to test different mechanistic hypotheses (Fig. 7). Finally, the DNA, RNA and protein array technologies can provide standardized large-scale data sets for the bioinformatics and machine learning community.

Directed evolution and active learning

Directed evolution is a method, most commonly used in protein engineering, to optimize the function (fitness) of a molecule by subjecting it to iterative rounds of mutagenesis and screening[94]. The best variants in each round are used as the starting point for the next round until the functional goal has been achieved. Examples of directed evolution include optimizing the activity, specificity and stability of enzymes[95,96] and, more recently, optimizing whole metabolic pathways and genomes[97,98]. Machine learning approaches have been applied to help navigate the fitness landscape of the molecule being optimized, thus reducing the experimental burden of screening variants[99]. These computational tools take the fitness of variants in the current round of evolution as input and from this input predict which region in the variant sequence space offers the greatest likelihood of increased fitness and thus should be screened in the next round[99]. To our knowledge, directed evolution has not yet been applied to high-throughput biochemical screening of variants on sequencing chips, but we believe that such an application is likely in the near future. For example, an RNA array could be used to optimize the binding affinity of an RNA of interest. To improve model predictions, active learning algorithms[100–102] could also use the previously available experimental data to iteratively suggest regions of RNA sequence space to further sample, such that areas of the sequence space that are likely to be more informative for model refinement are sampled more than those areas that are likely to be less informative. We envision that mechanistic physical modelling, deep learning models and optimization algorithms such as directed evolution and active learning can be integrated with feedback between the methodologies[92,99] to design, interpret and model experiments on RNA arrays (Fig. 7).

Future directions

The main outstanding challenges for carrying out high-throughput biochemical measurements on sequencing arrays include how to optimally design libraries to span the sampled sequence space to provide the most relevant information, and how to use the data coming from these experiments to build accurate predictive models. So far, most model implementations are semi-empirical, coarse-grained models that incorporate some physical basis in terms of how researchers think the interactions work at the molecular level, often with additive terms dependent on sequence features and sometimes extra empirical terms to account for non-additivity. When tested, these models can often describe the large-scale variations observed in the experimental data. However, we believe that much room for improvement exists in terms of building models that can learn underlying mechanistic principles, as well as predicting all of the nuances and details of the experimental data. As discussed in this Review, we believe that these challenges will be solved with a combination of deep learning modelling, directed evolution, kinetic mechanistic modelling, structural and atomistic simulations, and more informative experiments with a readout of intramolecular states through FRET and fluorescence quenching, ideally at the level of single molecules.

Determining the sequence–structure–function relationship for other molecular interactions (such as protein–DNA and protein–protein) is fundamentally similar to that for RNA. Thus, studying this problem for RNA, where the combinatorial sequence–structure–function space is smaller than that for proteins, is probably a good place to start, and we anticipate that the same approaches as discussed here should also apply to the more complex interactions of proteins.

Published online: 12 January 2023

References

1. Tinoco, I. Jr & Bustamante, C. How RNA folds. *J. Mol. Biol.* **293**, 271–281 (1999).
2. Ganser, L. R., Kelly, M. L., Herschlag, D. & Al-Hashimi, H. M. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.* **20**, 474–489 (2019).
   A comprehensive review that covers how the structural dynamics of RNA control cellular functions.
3. Al-Hashimi, H. M. & Walter, N. G. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321–329 (2008).
4. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
5. Mironov, A. S. et al. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–756 (2002).
6. Batey, R. T., Gilbert, S. D. & Montange, R. K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**, 411–415 (2004).
7. Flores, J. K. & Ataide, S. F. Structural changes of RNA in complex with proteins in the SRP. *Front. Mol. Biosci.* **5**, 7 (2018).
8. Shi, H. et al. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat. Commun.* **11**, 5531 (2020).
9. Vicens, Q. & Kieft, J. S. Thoughts on how to think (and talk) about RNA structure. *Proc. Natl Acad. Sci. USA* **119**, e2112677119 (2022).
10. Westhof, E. & Patel, D. J. Nucleic acids. From self-assembly to induced-fit recognition. *Curr. Opin. Struct. Biol.* **7**, 305–309 (1997).
11. Sussman, J. L., Holbrook, S. R., Warrant, R. W., Church, G. M. & Kim, S. H. Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *J. Mol. Biol.* **123**, 607–630 (1978).
12. Fürtig, B., Richter, C., Wöhnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem* **4**, 936–962 (2003).
13. Leontis, N. B. & Zirbel, C. L. in *RNA 3D Structure Analysis and Prediction* (eds Leontis, N. & Westhof, E.) 281–298 (Springer Berlin Heidelberg, 2012).
14. Holley, R. W. et al. Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
15. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proc. Natl Acad. Sci. USA* **77**, 4679–4682 (1980).
16. Wang, X. D. & Padgett, R. A. Hydroxyl radical 'footprinting' of RNA: application to pre-mRNA splicing complexes. *Proc. Natl Acad. Sci. USA* **86**, 7795–7799 (1989).
17. Latham, J. A. & Cech, T. R. Defining the inside and outside of a catalytic RNA molecule. *Science* **245**, 276–282 (1989).
18. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
19. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).
20. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.* **10**, 1643–1669 (2015).
21. Van Damme, R. et al. Chemical reversible crosslinking enables measurement of RNA 3D distances and alternative conformations in cells. *Nat. Commun.* **13**, 911 (2022).
22. Hafner, M. et al. CLIP and complementary methods. *Nat. Rev. Methods Prim.* **1**, 1–23 (2021).
23. Weidmann, C. A., Mustoe, A. M., Jariwala, P. B., Calabrese, J. M. & Weeks, K. M. Analysis of RNA–protein networks with RNP-MaP defines functional hubs on RNA. *Nat. Biotechnol.* **39**, 347–356 (2020).
24. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
25. Spitale, R. C. & Incarnato, D. Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-022-00546-w (2022).

26. Nutiu, R. et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
This paper reports the first implementation of a high-throughput biophysical measurement on a sequencing chip, involving binding of the yeast transcription factor GCn4 to a library of DNA sites.
27. Tome, J. M. et al. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods* **11**, 683–688 (2014).
This paper reports one of the first implementations of high-throughput biophysical measurements on sequencing chips for RNA, involving the binding of GFP and NELF-E to RNA aptamers.
28. Buenrostro, J. D. et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
This paper reports one of the first implementations of high-throughput biophysical measurements on sequencing chips for RNA, involving binding of the coat protein of MS2 bacteriophage to RNA hairpins.
29. Layton, C. J., McMahon, P. L. & Greenleaf, W. J. Large-scale, quantitative protein assays on a high-throughput DNA sequencing chip. *Mol. Cell* **73**, 1075–1082.e4 (2019).
30. Yesselman, J. D. et al. Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. *Proc. Natl Acad. Sci. USA* **116**, 16847–16855 (2019).
In this paper, the authors study RNA–RNA binding using tectoRNAs on the RNA array and construct a structure-based model that can predict experimental binding energies.
31. She, R. et al. Comprehensive and quantitative mapping of RNA–protein interactions across a transcribed eukaryotic genome. *Proc. Natl Acad. Sci. USA* **114**, 3619–3624 (2017).
32. Li, Z. et al. DNB-based on-chip motif finding: a high-throughput method to profile different types of protein-DNA interactions. *Sci. Adv.* **6**, eabb3350 (2020).
33. Ozer, A. et al. Quantitative assessment of RNA-protein interactions with high-throughput sequencing–RNA affinity profiling. *Nat. Protoc.* **10**, 1212–1233 (2015).
34. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).
35. Denny, S. K. et al. High-throughput investigation of diverse junction elements in RNA tertiary folding. *Cell* **174**, 377–390.e20 (2018).
36. Jarmoskaite, I. et al. A quantitative and predictive model for RNA binding by human Pumilio proteins. *Mol. Cell* **74**, 966–981.e18 (2019).
37. Wu, M. J., Andreasson, J. O. L., Kladwang, W., Greenleaf, W. & Das, R. Automated design of diverse stand-alone riboswitches. *ACS Synth. Biol.* **8**, 1838–1846 (2019).
38. Becker, W. R. et al. High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. *Mol. Cell* **75**, 741–755.e11 (2019).
39. Becker, W. R. et al. Quantitative high-throughput tests of ubiquitous RNA secondary structure prediction algorithms via RNA/protein binding. Preprint at *bioRxiv* https://doi.org/10.1101/571588 (2019).
40. Andreasson, J. O. L., Savinov, A., Block, S. M. & Greenleaf, W. J. Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. *Nat. Commun.* **11**, 1663 (2020).
41. Bonilla, S. L. et al. High-throughput dissection of the thermodynamic and conformational properties of a ubiquitous class of RNA tertiary contact motifs. *Proc. Natl Acad. Sci. USA* **118**, e2109085118 (2021).
42. Andreasson, J. O. L. et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. *Proc. Natl Acad. Sci. USA* **119**, e2112979119 (2022).
43. Jung, C. et al. Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips. *Cell* **170**, 35–47.e13 (2017).
44. Jones, S. K. Jr et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat. Biotechnol.* **39**, 84–93 (2021).
45. Denny, S. K. & Greenleaf, W. J. Linking RNA sequence, structure, and function on massively parallel high-throughput sequencers. *Cold Spring Harb. Perspect. Biol.* **11**, a032300 (2019).
46. Bartel, D. P. Metazoan microRNAs. *Cell* **173**, 20–51 (2018).
47. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
48. Cate, J. H. et al. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273**, 1678–1685 (1996).
49. Serganov, A. & Patel, D. J. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.* **8**, 776–790 (2007).
50. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
51. Sadée, C. et al. A comprehensive thermodynamic model for RNA binding by the *Saccharomyces cerevisiae* Pumilio protein PUF4. *Nat. Commun.* **13**, 4522 (2022).
52. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
53. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl Acad. Sci. USA* **114**, 5461–5466 (2017).
54. Ober-Reynolds, B. et al. High-throughput biochemical profiling reveals functional adaptation of a bacterial Argonaute. *Mol. Cell* **82**, 1329–1342.e8 (2022).
55. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
56. Marklund, E. et al. Sequence specificity in DNA binding is mainly governed by association. *Science* **375**, 442–445 (2022).
57. Eslami-Mossallam, B. et al. A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nat. Commun.* **13**, 1367 (2022).
References 56 and 57 (Marklund et al. and Eslami-Mossallam et al.) show how high-throughput data on binding, unbinding and cleavage of DNA by Cas9 can be used to gain microscopic mechanistic insights and build kinetic mechanistic models.
58. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
59. Soukup, G. A. & Breaker, R. R. Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA* **96**, 3584–3589 (1999).
60. Suess, B., Fink, B., Berens, C., Stentz, R. & Hillen, W. A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Res.* **32**, 1610–1614 (2004).
61. Förster, T. Zwischenmolekulare energiewanderung und fluoreszenz. *Ann. Phys.* **437**, 55–75 (1948).
62. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc. Natl Acad. Sci. USA* **58**, 719–726 (1967).
63. Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods* **25**, 78–86 (2001).
64. Muschielok, A. et al. A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5**, 965–971 (2008).
65. Lerner, E. et al. Toward dynamic structural biology: two decades of single-molecule Förster resonance energy transfer. *Science* **359**, eaan1133 (2018).
66. Chauvier, A. et al. Monitoring RNA dynamics in native transcriptional complexes. *Proc. Natl Acad. Sci. USA* **118**, e2106564118 (2021).
67. Winz, M.-L., Samanta, A., Benzinger, D. & Jäschke, A. Site-specific terminal and internal labeling of RNA by poly(A) polymerase tailing and copper-catalyzed or copper-free strain-promoted click chemistry. *Nucleic Acids Res.* **40**, e78 (2012).
68. Betzig, E. & Chichester, R. J. Single molecules observed by near-field scanning optical microscopy. *Science* **262**, 1422–1425 (1993).
69. Ha, T. et al. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* **93**, 6264–6268 (1996).
70. Zhuang, X. et al. A single-molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051 (2000).
71. Shema, E. et al. Single-molecule decoding of combinatorially modified nucleosomes. *Science* **352**, 717–721 (2016).
This paper shows the first implementation of high-throughput, single-molecule sequencing by synthesis combined with screening of binding, which is used to study nucleosome modifications in a DNA library of the mouse genome.
72. Severins, I., Joo, C. & van Noort, J. Exploring molecular biology in sequence space: the road to next-generation single-molecule biophysics. *Mol. Cell* **82**, 1788–1805 (2022).
This review summarizes the previous applications of high-throughput biophysical measurements on sequencing chips, and discusses in detail how the technology can be extended to carry out single-molecule experiments.
73. Magde, D., Elson, E. & Webb, W. W. Thermodynamic fluctuations in a reacting system — measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* **29**, 705 (1972).
74. Yu, L. et al. A comprehensive review of fluorescence correlation spectroscopy. *Front. Phys.* **9**, 644450 (2021).
75. Zheng, Q. et al. Ultra-stable organic fluorophores for single-molecule research. *Chem. Soc. Rev.* **43**, 1044–1056 (2014).
76. Marklund, E. et al. DNA surface exploration and operator bypassing during target search. *Nature* **583**, 858–861 (2020).
77. Wayment-Steele, H. K. et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat. Methods* **19**, 1234–1242 (2022).
78. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
79. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
80. Yu, H., Qi, Y. & Ding, Y. Deep learning in RNA structure studies. *Front. Mol. Biosci.* **9**, 869601 (2022).
81. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 5407 (2019).
82. Zhang, H. et al. A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Front. Genet.* **10**, 467 (2019).
83. Wang, L. et al. DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front. Genet.* **10**, 143 (2019).
84. Calonaci, N., Jones, A., Cuturello, F., Sattler, M. & Bussi, G. Machine learning a model for RNA structure prediction. *Nar. Genom. Bioinform.* **2**, lqaa090 (2020).
85. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).
86. Fu, L. et al. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14 (2022).

87. Townshend, R. J. L. et al. Geometric deep learning of RNA structure. *Science* **373**, 1047–1051 (2021).
In this paper, the authors apply deep learning to build a model that can predict the tertiary structure of RNAs after being trained on high-resolution structural data.
88. Wei, J., Chen, S., Zong, L., Gao, X. & Li, Y. Protein–RNA interaction prediction with deep learning: structure matters. *Brief. Bioinform.* **23**, bbab540 (2021).
89. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
90. Lam, J. H. et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat. Commun.* **10**, 4941 (2019).
91. Trabelsi, A., Chaabane, M. & Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* **35**, i269–i277 (2019).
92. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
93. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
94. Arnold, F. H. Design by directed evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
95. Arnold, F. H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253–257 (2001).
96. Zhao, H., Chockalingam, K. & Chen, Z. Directed evolution of enzymes and pathways for industrial biocatalysis. *Curr. Opin. Biotechnol.* **13**, 104–110 (2002).
97. Wang, Y., Yu, X. & Zhao, H. Biosystems design by directed evolution. *AIChE J.* **66**, e16716 (2020).
98. Tan, Z. L. et al. In vivo continuous evolution of metabolic pathways for chemical production. *Microb. Cell Fact.* **18**, 82 (2019).
99. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).
This review covers how machine learning has been applied to assist in the navigation of large sequence spaces during directed evolution.
100. Settles, B. Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **6**, 1–114 (2012).
101. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
102. Sverchkov, Y. & Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Comput. Biol.* **13**, e1005466 (2017).
103. Ennifar, E., Walter, P., Ehresmann, B., Ehresmann, C. & Dumas, P. Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site. *Nat. Struct. Biol.* **8**, 1064–1068 (2001).
104. Okada, K. et al. Solution structure of a GAAG tetraloop in helix 6 of SRP RNA from *Pyrococcus furiosus*. *Nucleosides Nucleotides Nucleic Acids* **25**, 383–395 (2006).
105. Kim, N.-K. et al. Solution structure and dynamics of the wild-type pseudoknot of human telomerase RNA. *J. Mol. Biol.* **384**, 1249–1261 (2008).
106. Kuglstatter, A., Oubridge, C. & Nagai, K. Induced structural changes of 7SL RNA during the assembly of human signal recognition particle. *Nat. Struct. Biol.* **9**, 740–744 (2002).
107. Stoddard, C. D. et al. Free state conformational sampling of the SAM-I riboswitch aptamer domain. *Structure* **18**, 787–797 (2010).
108. Collie, G. W., Haider, S. M., Neidle, S. & Parkinson, G. N. A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex. *Nucleic Acids Res.* **38**, 5569–5580 (2010).

Acknowledgements
The authors thank E. Sharma for discussions. This work was supported in part by NIH grants R01GM111990, P50HG007735, R01HG009909, P01GM066275, UM1HG009436 and R01GM121487 to W.J.G. W.J.G. acknowledges support as a Chan Zuckerberg Investigator. E.M. was supported by the Swedish Research Council grant 2020-06459.

Author contributions
All authors researched, discussed, wrote and edited the manuscript.

Competing interests
W.J.G. is a consultant and equity holder for 10x Genomics, Guardant Health, Quantapore and Ultima Genomics, and cofounder of Protillion Biosciences. The other authors declare no competing interests.

Additional information

Correspondence should be addressed to William J. Greenleaf.

Peer review information *Nature Reviews Genetics* thanks M. Depken and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.