Fitness Landscapes and Evolution of Catalytic RNA

Ranajay Saha[# 1], Alberto Vázquez-Salazar[# 1], Aditya Nandy*[1,2,3], Irene A. Chen*[1,4]

# equal contribution

1. Department of Chemical and Biomolecular Engineering, University of California, Los Angeles
2. Department of Chemistry, The University of Chicago, Chicago, IL, 60637
3. The James Franck Institute, The University of Chicago, Chicago, IL, 60637
4. Department of Chemistry and Biochemistry, University of California, Los Angeles

*Correspondence to AN and IAC: aditya.nandy@ucla.edu, ireneachen@ucla.edu

Abstract:

The relationship between genotype and phenotype, or the fitness landscape, is the foundation of genetic engineering and evolution. However, mapping fitness landscapes poses a major technical challenge due to the amount of quantifiable data that is required. Catalytic RNA is a special topic in the study of fitness landscapes, due to its relatively small sequence space combined with its importance in synthetic biology. The combination of in vitro selection and high-throughput sequencing has recently provided empirical maps of both complete and local RNA fitness landscapes, but the astronomical size of sequence space limits purely experimental investigations. Next steps are likely to involve data-driven interpolation and extrapolation over sequence space using various machine learning techniques. We discuss recent progress in understanding RNA fitness landscapes, particularly with respect to protocells and machine representations of RNA. The confluence of technical advances may significantly impact synthetic biology in the near future.

Introduction

Molecular evolution is often likened to a hill-climbing process in sequence space. In this conceptualization, molecular sequences (e.g., DNA, RNA or protein) inhabit a potential space of all possible sequences. A landscape may be drawn in this space, such that each point represents a possible genotype and the height of that point represents the fitness associated with that genotype (Figure 1a). The 'fitness landscape' describes the relationship between genotypes (the genetic makeup of an individual) and its reproductive success, or fitness (25, 109). Valleys represent genotypes with low fitness, while peaks represent genotypes with high fitness. The fitness of similar genotypes may be highly correlated (smooth landscape) or poorly correlated (rugged landscape) (14). The features of the fitness landscape determine the potential for adaptation and evolvability (36, 86) given genetic variation as the raw material (43, 109). Organisms can move through the landscape via mutations, and natural selection tends to push populations uphill, towards higher fitness (14). It is important to recognize that a full sequence space is extremely high-dimensional, since the number of dimensions equals the number of variable sites (i.e., the sequence length). The large number of dimensions allows for a very large number of potential evolutionary pathways, making exploration of these pathways quite challenging.

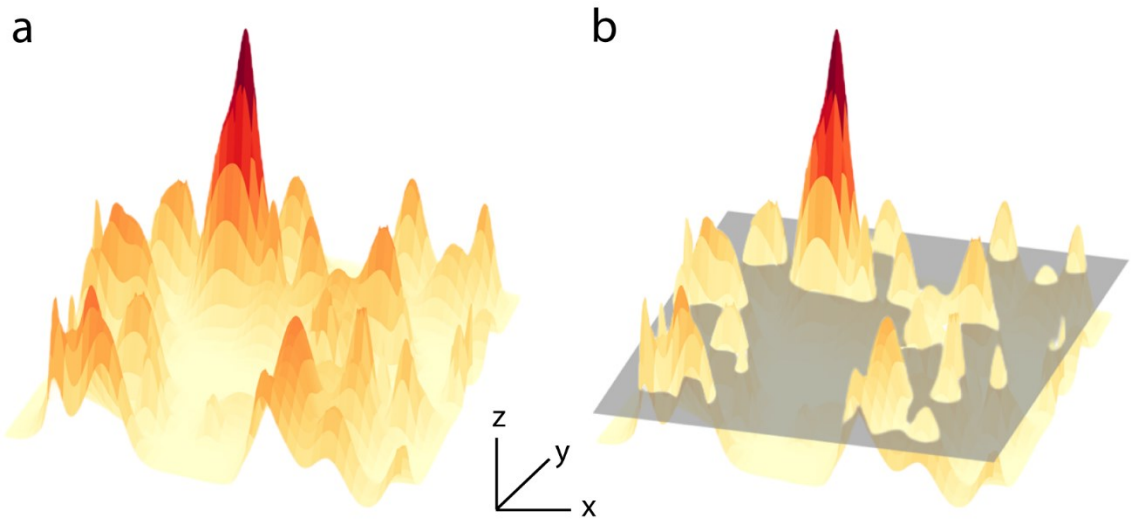

Figure 1. Conceptual drawings of a fitness landscape. (a) The x- and y-axis represent sequence space (which is *L*-dimensional in reality), and the z-axis shows fitness (also illustrated by color). Peaks represent high-fitness families. (b) In vitro selection establishes a threshold activity (a 'sea level', shown in gray) required for survival, eliminating sequences of low fitness. Only features of the fitness landscape above this threshold can be experimentally determined.

Although a fitness landscape 'map' may be inadequate for describing evolution in complex organisms evolving dynamically with significant interplay and feedback between genetic and environmental factors, the fitness landscape is a powerful metaphor for molecules. In particular, molecular activity (e.g., catalytic power or binding affinity) may be equated to 'fitness', the environments are typically kept constant, and the activity of a sequence is essentially determined only by its genotype. Molecular fitness landscapes are an extensive topic, including much theoretical work, and have been reviewed elsewhere (for examples, see (25, 82)). The simplest exploration of the fitness landscape is a biased random walk, where a genotype may move to a neighboring sequence (e.g., a single mutation) with a probability that is related to the fitness of that sequence. This process models a low mutation rate, in which new copies have only a single mutation. Additional mechanisms (e.g., high mutation rates, recombination, rearrangements)

would generate greater diversity and thus more expansive exploration of the fitness landscape. However, biological mechanisms generally explore only a very small fraction of potential sequence space. Here, we discuss recent and emerging data-driven techniques to map fitness landscapes of catalytic RNA. The RNA molecule presents a unique duality in biochemistry: it can both store genetic information, as in the case of RNA viruses, and catalyze chemical reactions, in the case of ribozymes, through complex three-dimensional structures. RNA molecules with catalytic activity, often selected in vitro (10, 19), are of special interest to study both the functional diversity of RNA and the RNA world hypothesis (11, 54, 91), which posits that RNA would comprise the metabolic architecture of the earliest cells. RNA is therefore a particularly intriguing molecule for studying the fitness landscape, and thus understanding how life could evolve in sequence space.

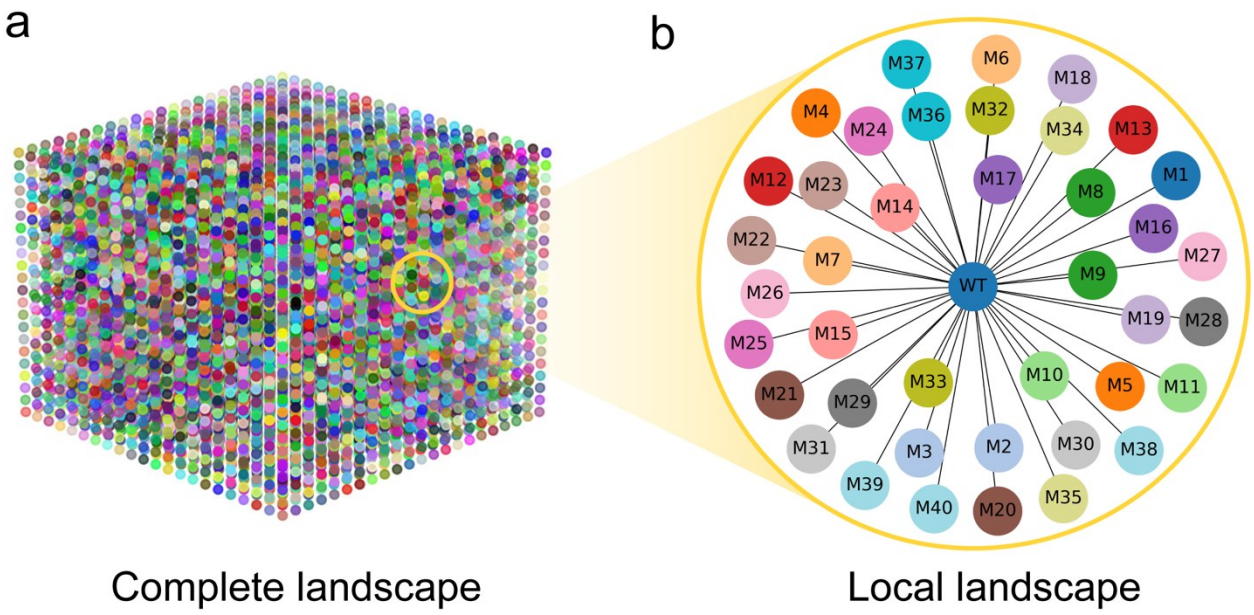

Figure 2. Conceptual drawings of possible experimentally explored sequence space. (a) Complete coverage of sequence space, limited to the regime of approximately $L < 26$ for RNA. The space is finite but very large, with L dimensions; this drawing is an abstract visualization. Virtually all points in sequence space at this L can be synthesized in a random pool. (b) Local

coverage of sequence space near a wild-type (WT) sequence, usually a known ribozyme. Many mutants (M) of the ribozyme are synthesized, exploring the local volume of sequence space immediately surrounding the WT.

Fitness Landscape Cartography

Although the size of sequence space increases exponentially with length (4^L sequences of length L), there exists an intermediate regime for L in which L is small enough that it is possible to create a library containing almost all possible sequences through random synthesis, and at the same time L is large enough to have sequence-specified chemical activities (Figure 2a). In this regime (approximately $L = 5$, based on the function exhibited by very small ribozymes (102), up to perhaps 26, due to experimental laboratory scales), one may take advantage of the fact that the vast majority of sequence space lacks activity in a given environment. Therefore, a two-step process can be employed to map the fitness landscape: 1) isolating the small fraction of active sequences from a random library covering sequence space, effectively separating the 'wheat from the chaff' (Figure 3a), and then 2) assaying those sequences in a high-throughput method (Figure 3b). In this two-step process (also called SCAPE, for *s*equencing to measure *c*atalytic *a*ctivity *p*aired with in vitro *e*volution (92)), in vitro selection is used for the first step of enriching active sequences. For example, active variants can be captured on solid phase support to select based on binding affinity (for aptamers) or chemical conjugation to a ligand (for ribozymes). In vitro selection thus reduces the complexity of the RNA library from 4^L down to an assayable number, perhaps in the thousands. This strategy is not limited to down-selection of random libraries, but can also be applied to designed libraries, such as a library based on a ribozyme where only specific positions have been randomized (77) (also called "mutate-select-and-sequence" (112)), or a library of recombinants derived from known ribozyme sequences

(20), with the caveat that the sequence space being mapped is a specially biased subset of full sequence space (Figure 2b).

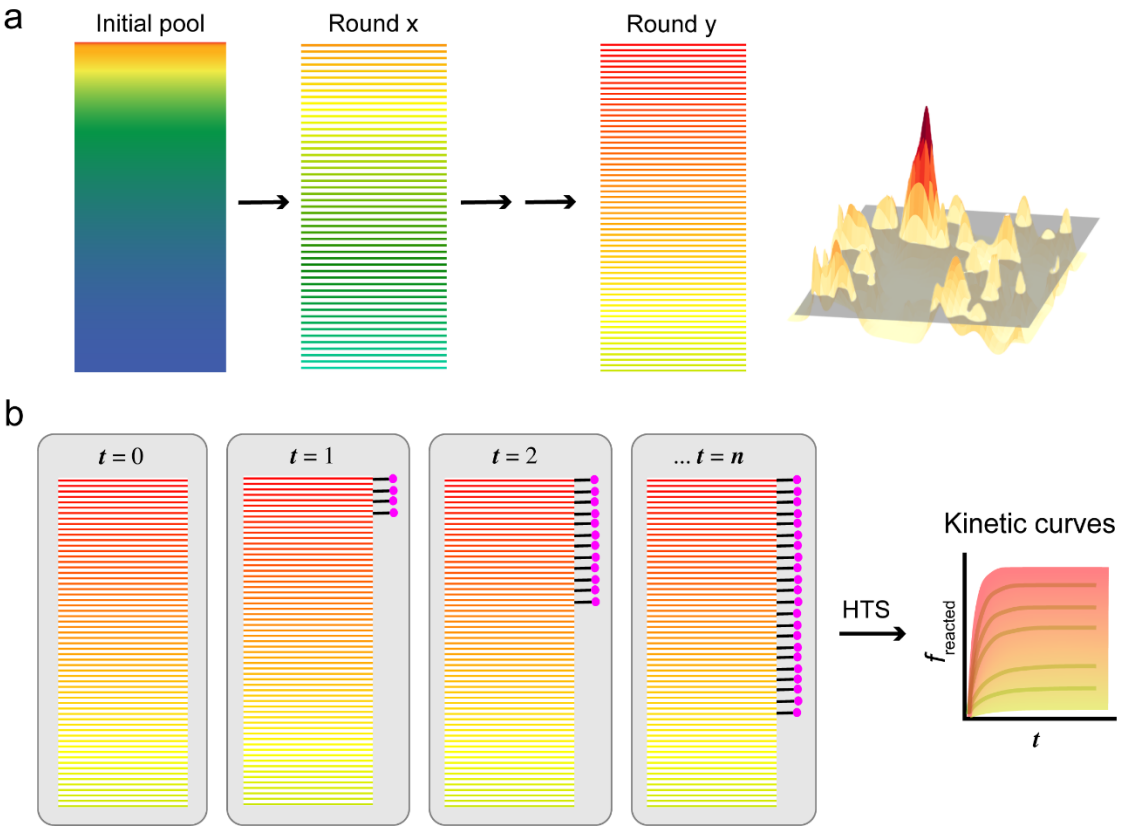

Figure 3. Two-part procedure for mapping fitness landscapes by SCAPE. (a) First, an initial pool of sequences is designed to cover a defined sequence space. Fitness of different sequences is shown by color, with red being highest fitness and blue being low fitness. Most sequences in the initial pool are likely to have low fitness. During rounds of in vitro selection, the low fitness sequences are eliminated while high fitness sequences are enriched. At some point (Round y), the complexity of the pool has been reduced to an assayable number of sequences having fitness above the threshold value (gray). (b) Second, the selected pool (Round y) is assayed with a technique such as high-throughput sequencing (HTS). For a ribozyme, kinetic characterization of the pool is performed with reactions at multiple time points (or substrate concentrations, etc.), to assay all of the sequences in parallel. Reacted molecules (indicated by the purple marker) are

separated biochemically from unreacted molecules and sequenced. For each unique sequence, data on the number of reads is translated into the reacted fraction across time points and fit to a kinetic model.

Once the selected variants are isolated, a high-throughput assay is then employed for the second step of determining activity values for the selected sequences (Figure 3). High-throughput sequencing (HTS) technology provides the ability to collect a large amount of data (~10^8 sequence reads). By using the number of reads as a means to quantify sequences before and after a reaction, the relative activities of different sequences within a pool can be determined in a massively parallel assay. What is required from the experimenter is a method to separate reacted and unreacted RNAs, such as by affinity chromatography or gel electrophoresis, resulting in a sample of reactive RNAs that can be sequenced and counted through reads. With appropriate normalization standards, sequence read abundances can also be translated into absolute activity values (e.g., rate constants or binding affinities). HTS was first applied by Pitt and Ferré-D'Amaré to a heavily mutagenized library of the class II RNA ligase ribozyme (mutation rate of 21% per position in 45 nt) subjected to in vitro selection (90). The study yielded the local fitness landscape of this laboratory-evolved ribozyme, with relative abundance of reads, which correlated with ribozyme activity, used as a proxy for fitness for individual sequences. An advance on HTS assay techniques is to perform the reaction while sampling at multiple time points or substrate concentrations to obtain data for fitting to a kinetic model (as opposed to a single reaction point) (28, 92). These techniques can be extended beyond nucleic acid sequences, as long as the molecules are uniquely labeled by nucleic acids. For example, binding kinetics were measured for over twenty thousand peptide sequences in parallel, by using mRNA display to read out the identity of bound peptides (48). HTS assays also have utility outside of the goal of

mapping fitness landscapes, and can be applied to a designed library without selection ("mutate-and-sequence") (58-60, 112). Additional technical variations could be envisioned, such as characterization of inhibitors and substrate specificity (49). These methods have emerged from the synergy between HTS and the tradition of using nucleic acids as 'barcodes' while screening synthetic chemical (or biomolecular) libraries (31).

How many sequences can be assayed by these methods depends on the depth of sequencing as well as the evenness of the pool being assayed. A general rule of thumb from our experience is that at least 100 (and preferably 1000) sequence reads are needed to obtain a reliable estimate of activity (99). If the sequence variants are evenly represented in the pool, then a sequencing depth of 10^8 reads would give 100 reads of 10^6 sequences. Thus, in principle, up to about a million sequences might be assayed at this depth, but any deviation from an even distribution would reduce this number. In particular, pools resulting from in vitro selection are likely to be highly uneven, such that the assayable number of sequences is more likely to be on the order of tens of thousands. As with any other experimental protocol, HTS experiments should be conducted with replicates to evaluate reproducibility. Overall, for a limited scope, SCAPE can experimentally determine comprehensive or local maps of fitness (i.e., activity) landscapes for functional RNAs.

The Chemical Environment

Environmental conditions can greatly affect fitness landscapes. For instance, a mutation in bacteria that confers resistance to an antibiotic would increase fitness in the presence of the antibiotic, but might decrease fitness in the absence of the antibiotic, due to the added metabolic load (22). Environmental factors that can influence molecular RNA fitness include pH, ion

concentration, temperature, availability of substrates and nutrients, interactions with other biopolymers, and more. While a full discussion of the variety of environmental inputs is beyond the scope of this review, a few examples are given here to illustrate their importance.

The pH of an environment plays a pivotal role in influencing the structure, function, and stability of RNA molecules by altering the protonation states of heteroatoms in the nucleobases, hydroxyls, and phosphates (66). Under acidic conditions, the protonation of G:C base pairs may reduce hydrogen bonding, decreasing the stability of the duplex, and noncanonical A:C and C:C base pairings may occur (12). Basic pH promotes hydrolysis of the RNA backbone through deprotonation of the 2′-OH group, among other effects. These changes can significantly disrupt base-pairing and tertiary interactions, in addition to impacting the interactions of RNA with metal ions and organic molecules. Cations, particularly divalent cations such as magnesium (Mg^{2+}), can interact with the negatively charged phosphate backbone of RNA, allowing the RNA molecule to fold into a compact structure and sometimes participating in the catalytic mechanism. One study on mutational variants of the *Azoarcus* group I self-splicing intron over a $MgCl_2$ concentration series (1-48 mM) showed that fitness in the library increased with increasing magnesium concentration (88). Likewise, changes in temperature can significantly influence the folding rate and equilibrium ensemble of RNA structures. High temperatures generally denature nucleic acids, but some catalytic nucleic acids, such as a G-rich peroxidase DNAzyme that forms a four-stranded guanine-quadruplex, can retain activity at temperatures as high as 95 degrees C (40). At the other extreme, temperatures nearing the freezing point may be advantageous in reducing the rate of hydrolysis while concentrating reactants as water

crystallizes (8). These effects are not always predictable, and experimental study is therefore required.

In addition to the effects of ions and small molecules, the interior of cells is widely appreciated to be 'crowded' with macromolecules (Figure 4a). Thus, the volume available to a macromolecule is nonspecifically constrained by the volumes of other macromolecules present, affecting dynamics such as diffusion, binding, and conformational changes (i.e., excluded volume effects) (93, 96). For example, the highly stable G-quadruplex motif in RNA is known to play essential roles in biological reactions such as translational regulation and alternative splicing (34, 35, 41, 62). The model crowding agent polyethylene glycol (PEG200) stabilized G-quadruplexes having 3-4 G-quartets (71). Crowding also appears to compensate for the loss of other ribozyme interactions. For example, the addition of PEG enhances the activity of Azoarcus group I ribozyme and reduces its Mg^{2+} requirement (27, 56, 57), and PEG also enhances the activity of a ligase ribozyme in disfavorable conditions for folding and activity (e.g., high urea concentration, alkaline pH) (23, 68). At the same time, crowding agents may also have effects aside from the excluded volume effect, e.g., due to chemical interactions with the RNA, so mechanistic studies are often important (21, 57, 103).

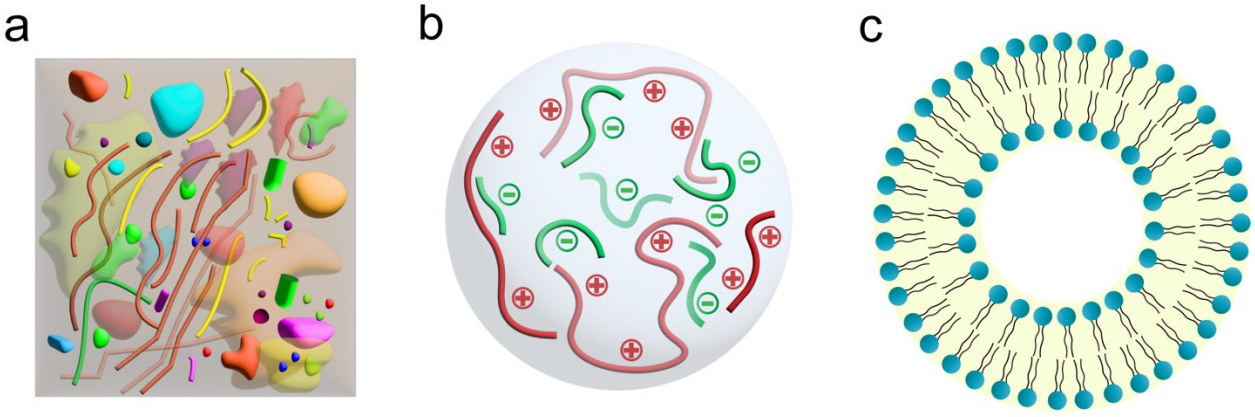

Figure 4. Conceptual illustrations (not to scale) showing (a) crowded environment of biological cells, (b) a complex coacervate droplet formed due to electrostatic interaction between oppositely charged molecules, and (c) a bilayer membrane vesicle.

Given the fluctuations of natural environments and the dependence of fitness on the environment, fitness landscapes are inherently dynamic. Such shifts could facilitate the exploration and selection of novel motifs, which might remain undiscovered under standard or constant conditions. Many conditions, particularly pH, temperature and ionic conditions, affect both catalytic activity and hydrolytic stability, illustrating that molecular fitness includes multiple components. For an RNA polymerase ribozyme, increasing Mg^{2+} concentration results in greater polymerization activity but also greater hydrolysis, ultimately limiting the extent of polymerization (65, 74). This example illustrates that different aspects of fitness can experience opposing effects in a changing environment. The fitness landscape is a complex, dynamic object.

Fitness Landscapes for Protocells

While fitness is usually considered to be a property of an individual organism, natural selection actually occurs at multiple levels, often simultaneously (39, 69). For example, selection at the level of the gene within the genome may lead to the evolution of apparently parasitic sequences, such as transposons (107), while higher-level selection may drive the evolution of cooperative traits (81). Given the complex integration of distinct molecules that is required for metabolism, levels of selection above the molecular must have been essential to the emergence of cellular life. Thus, understanding the fitness of the earliest molecular ensembles, including how RNA fitness

is affected by other components of the ensemble, is important for understanding the origin of life.

Protocells are experimental models of primitive cells (13, 63, 101). Regardless of their physical form, protocells are essentially a nonrandom grouping of individual molecules, creating a potential level of selection above the ribozyme (3). Cooperative phenotypes, which may be disadvantageous (i.e., selected against) at the level of individual molecules, can evolve in protocells (9, 101). Protocells that create physical compartments for RNA have gained particular attention for the origin of life (45, 89). Understanding how protocells affect the functional behavior of RNAs and mapping the fitness landscape of protocells themselves are current goals in this area. Two types of protocells, coacervates and membrane vesicles, are considered here.

The formation of coacervates is now a well-recognized process in cell biology (44). Complex coacervates form when a charged macromolecule interacts with oppositely charged molecules, creating a separate, still liquid, macromolecule-rich phase (Figure 4b). The interactions are generally nonspecific, highly dynamic and of moderate strength, with the associations being driven by electrostatic interactions and the entropic gain from counterion release (106). Combinations of neutral molecules, such as dextran and PEG, may also form aqueous two-phase systems (ATPS) with the liquid phases having distinct polymer compositions. Although the equilibrium state is bulk separation of the phases, individual droplets can form as a metastable state. In the context of a cell, these droplets may act as membraneless organelles. It has long been postulated that coacervate droplets could organize prebiotic molecules together to form a

protocell (83), a hypothesis that has recently gained experimental interest following advances in coacervate cell biology.

Within a coacervate, the altered microenvironment increases local concentrations and may change relative energy levels for ground and transition states, affecting the rate of chemical reactions (76). An early demonstration showed that compartmentalization in an ATPS increased the reaction rate of the hammerhead ribozyme by nearly two orders of magnitude, largely due to the increased RNA concentration (100). Complex coacervates made of negatively charged RNA and positively charged peptides may be particularly interesting from the standpoint of RNA-peptide coevolution (30, 73). Recently, a complex coacervate made of poly-L-lysine and the hairpin ribozyme was found to enhance ribozyme activity by 1-2 orders of magnitude. The hairpin ribozyme catalyzes both cleavage and ligation, given the appropriate substrates. Interestingly, the coacervate environment was found to shift the equilibrium toward ligation, likely due to the high RNA concentration, suggesting that this effect might be harnessed in protocells to create longer RNAs from shorter oligonucleotides (67). Furthermore, the ligation products altered the physical properties of the droplets, including reduced rates of RNA release from the droplets (104). Coacervates containing out-of-equilibrium chemical systems exhibit phenotypes at the droplet level, such as changes in growth and fusion rates (75). Droplet properties are also influenced by the sequences of the peptide component, with one study illustrating that charged-interspaced heteropeptides (Arg-Gly-Gly repeats compared to poly-Arg) favored the liquid rather than gel phase and also showed better sequestration of Mg^{2+}, enabling ligase ribozyme activity (46). These studies lay the groundwork for potential genotype-phenotype coupling that could lead to natural selection at the level of the protocell. Along with

this progress, important challenges also remain in the field of coacervate protocells, including the colloidal stability of droplets as compartments for multiple generations of selection and replication (2, 26), and exchange of biopolymers among droplets that may interfere with individuation of compartments (50).

Mimicking the bilayers of contemporary biological cells, membrane vesicles have become established as an experimental model for protocells (Figure 4c). In particular, fatty acids with eight or more carbons form bilayer vesicles at a pH near their pK_a in the membrane (38), and are envisioned as a transition stage for protocells before the appearance of the more robust diacylphospholipid membranes found in modern cells (51). Fatty acids can be synthesized under simulated prebiotic conditions (70, 95), can grow and divide (1, 42, 105, 115), and allow permeation of RNA building blocks such as nucleoside phosphorimidazoles and cations such as Mg^{2+} (2). Vesicles also appear to form a kinetically stable microenvironment suitable for multiple generations of selection and replication (50). The lability of fatty acid vesicles to high Mg^{+2} makes ribozyme compatibility challenging, but this sensitivity may be mitigated by partial chelation (2) or with ribozymes having low Mg^{2+} requirement (17, 24). Thus, RNA fitness within fatty acid protocells would involve tolerance to low Mg^{2+} or chelating conditions.

In additional to chemical compatibility, another effect of protocells on RNA fitness landscape occurs due to an excluded volume effect, specifically due to the physical confinement presented by the membrane itself, which restricts the volume of the encapsulated macromolecules. By altering the energies of different conformations (114), confinement stabilizes compact structures in RNA. This results in effects such as higher ligand binding affinity for an RNA aptamer (97),

enhanced RNA-RNA association (both intermolecular and intramolecular), and increased docking interactions for the hairpin ribozyme, leading to the increased catalytic activity (87). Encapsulation restored activity for folding-deficient mutants of the hairpin ribozyme (87), similar to an effect of macromolecular crowding (85), suggesting significant impacts to the ribozyme fitness landscape. These findings motivated evaluation of how encapsulation inside vesicles altered the local fitness landscapes of several self-aminoacylating ribozymes (64, 92). In a high-throughput study, thousands of ribozyme sequences showed consistently higher activity when encapsulated. At the same time, epistatic effects (i.e., "ruggedness" on the landscape) were amplified. Interestingly, encapsulation also increased the variance of fitness, such that the RNA population adapted more quickly during in vitro evolution, in accordance with Fisher's Fundamental Theorem of Natural Selection (37). This study illustrated how protocells could alter the fitness landscape, and its exploration, in significant ways. Combined with other studies demonstrating mechanisms for genotype-phenotype coupling at the protocell level (1, 16, 33), these lines of work show how membrane vesicles yield complex system-level behaviors. While vertical transmission of genetic information could occur through growth and division, horizontal transfer has also been implemented. Giant unilamellar vesicles (GUVs) encapsulating RNA and subjected to a freeze-thaw process showed mixing of contents among individuals (98). Understanding the higher-level fitness landscape of vesicle protocells, in terms of both its fundamental structure and how populations explore the landscape through genetic transmission, is an important future challenge.

Considerations for Machine Learning

Catalytic RNA fitness landscapes are amenable to data-driven discovery using machine learning (ML) due to the large quantities of available sequencing data that enable analysis of underlying trends. Early kinetic sequencing studies (78) investigated the kinetics of catalytic RNA substrate specificity, evaluating a pool of ~10^3 sequences. Fluorescence-based measurements then enlarged the possible space to ~10^4 sequences (6). Since then, deep sequencing (28, 60, 79, 113), including *k*-Seq (49, 92, 99), has unlocked large (10^3-10^6) RNA sequence spaces to identify high fitness sequences for different types of reactivity. Analysis of the active sequences could unveil underlying motif commonalities that likely give rise to catalytic behavior (92). The presence of these relationships indicates that ML could be used for quantitative sequence-function mapping. In vitro evolution, or sequential evolution of ligands by exponential enrichment (108) (SELEX), over large sequence spaces, used in many sequencing studies (4, 5), represents a frontier where ML-acceleration for RNA discovery remains largely untapped. The data from in vitro evolution can be used to construct quantitative ML models that map sequence and conditions to function. This data can be used to identify common characteristics in high-fitness motifs and ideal conditions (61) (i.e. temperature, pH, ion identity) for selection during one round. Correspondingly, subsequent rounds of in vitro evolution may use improved conditions and thus reduce the number of rounds needed to identify the best-performing sequences.

A critical challenge for ML-accelerated discovery is representing RNA sequences in a form that ML models are able to learn. The way that an RNA sequence is represented to a model is called a "representation" or "featurization" and influences the types of models that are used, the way in which a model learns the underlying trends, and the interpretability of model predictions. In the context of linear regression, this representation is the "x" variable in y =

$W*x+b$, with "W" being the weights and "b" being the biases. In ML models, instead of a linear mapping between the "x" and "y" variables, a nonlinear mapping is performed instead, adding flexibility to the fitting function (f) that maps x and y ($y = f(x)$). Often, representation choice influences whether or not a model is able to generalize to new chemical spaces beyond the training data. Whether constructing models that harness labeled data (i.e. supervised models), those that cluster data without labels (i.e. unsupervised models), or those that use a subset of labeled points (i.e. semi-supervised models), representation choice is crucial for successful ML model development. When selecting a representation, it is important to consider the data, the problem at hand, and the objectives of the ML model. Some essential considerations include: 1) whether the sequences are of consistent length, 2) whether modified nucleotides (i.e. methylated or xeno-nucleic acids) are used, and 3) whether future objectives are inherently extrapolative or interpolative based on the input data. Although there have been advances in ribozyme structure determination, many catalytic RNA fitness-landscape problems are naturally posed as sequence-function mappings. With the exception of ML-accelerated directed evolution (111), this sequence-function approach is in contrast to the sequence-structure and subsequent structure-function mappings that are typically used to understand enzymatic catalysis. This difference might be attributed to factors such as: 1) few RNA structures display quarternary structures (52) that commonly occur in proteins, and 2) the use of high-throughput sequencing (14) and mutational analysis (60, 113) for RNA provides direct insights into sequence-function maps, enabling structures to be forgone.

One-hot encoding is the simplest representation for RNA sequences. One-hot encoding is a binary vector representation where "1" is used to represent the presence of a value ('hot' if present, 'cold' if absent with value set to "0"). In this representation, sequences of length N are

represented by an *N* by 4 sized bit vector (32). One-hot encodings do not readily generalize to larger or shorter sequence lengths; the length *N* must be set by the largest sequence, with remaining positions filled with 0 (in a process called zero padding) if no nucleotide exists. Each of the four columns represents the four RNA nucleotides (e.g. A, U, G, C) and contains a 1 if a nucleotide is present at that sequence position or a 0 if it is not (84) (Figure 5). Due to its nature, one-hot encoding is sparse and encodes the presence of certain nucleotides in specific positions, but not their relationships or positional dependencies relative to other nucleotides. Correspondingly, this primitive representation will fail to generalize effectively for nearly all sequence-property relationships, which depend on motifs of many nucleotides as opposed to a single nucleotide in a specific sequence position. Because they are sparse, one-hot representations require more data to learn from.

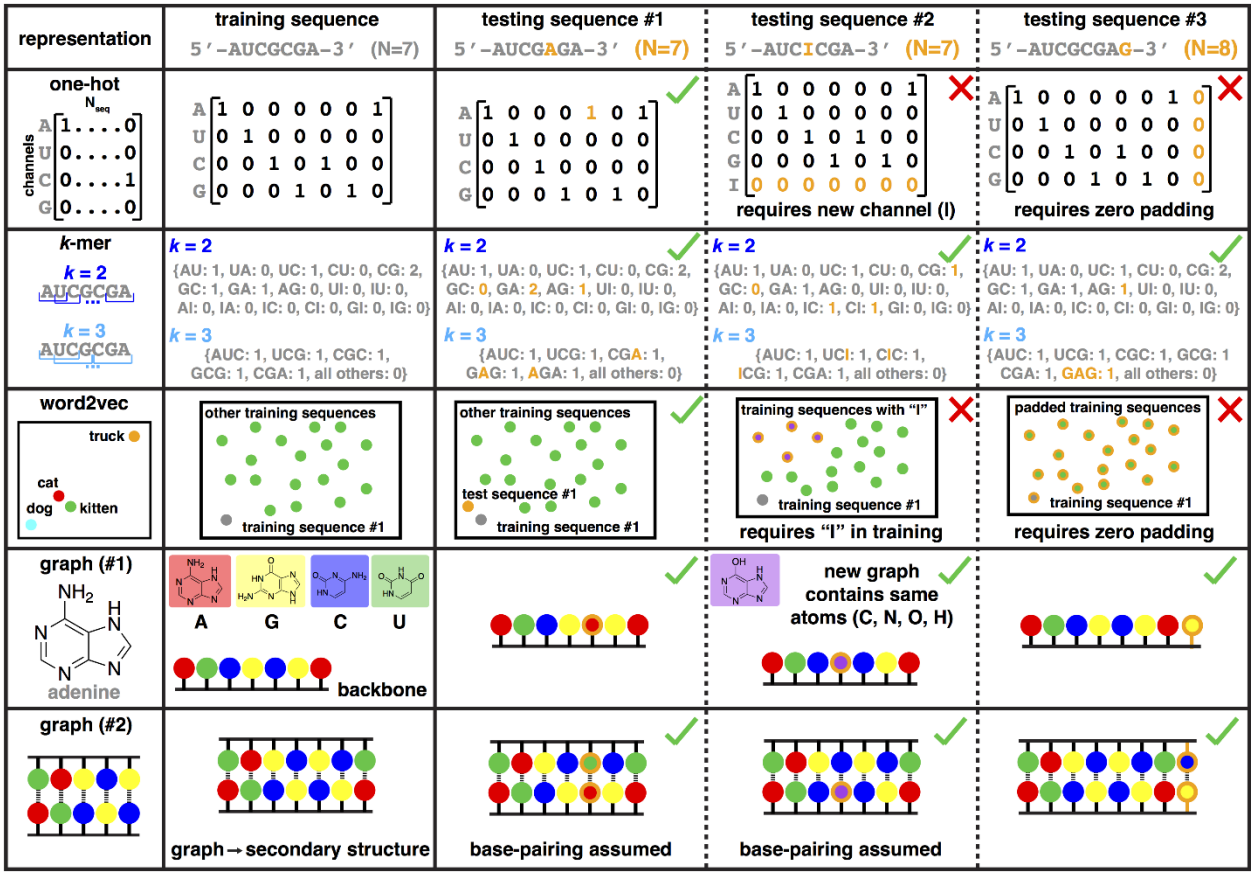

Figure 5. Various representations for RNA sequences for use in machine learning models with a 7-nucleotide long training sequence given as an example: 1) one-hot encoding, using "1" to indicate the presence of a given nucleotide at a given position, 2) *k*-mer encoding, developing a histogram of subsequence counts based on a given sliding window, 3) word2vec encoding, learning a distribution of RNA sequences, 4) a molecular graph encoding, incorporating information from different bonds and atoms into a connectivity graph, and 5) an abstracted molecular graph encoding, incorporating secondary structure between RNA strands. For each encoding, a representative training sequence and corresponding encoding is denoted. Then, three test sequences (with differences noted in orange in the top row) are presented. A green check mark indicates if the training data are natively able to handle the test sequence and how that test sequence would be encoded, with changes to the representation in orange. A red cross indicates that the representation and given training sequence would not readily generalize to the presented test set sequence. In this case, the changes required to the training data representation are denoted in orange, along with a caption that identifies the corresponding challenges.

The *k*-mer representation encodes relationships to adjacent nucleotides by defining a window size (i.e. *k*) and sliding the window along the sequence, leading to a histogram of subsequence fragments (7). As an example, for a sequence “AUCGCGA” to be represented as a 2-mer (*k*=2), we represent the sequence as follows: [AU: 1, UC: 1, CG: 2, GC: 1, GA: 1]. In this histogram, “CG” appears twice in the sequence and is thus counted twice (Figure 5). The *k*-mer representation is highly sensitive to window size: high *k*-values generate a more global representation relative to small *k*-values, which generate a local representation. As the *k*-values increase, the number of features also increases, thus generating an increasingly sparse representation and presenting similar challenges to one-hot encodings. Due to its window-size, the *k*-mer representation enables extraction of trends that depend on subsets of sequences, although the histogram-like nature removes positional dependence.

Although one-hot and *k*-mer representations convert text-based sequences into numerical representations for use in ML models, these representations suffer from a curse of dimensionality that can make them too sparse (e.g. too many features). Similarly, the *k*-mer representation can fail to encode critical relationships that are not readily captured by a sliding window. Here, ideas from natural language processing (NLP) facilitate improved text-based representations. Word2vec (72) is a numerical NLP representation that is learned from a distribution of words in a corpus (i.e. a collection of text). In this representation, words are obtained from a corpus and stored with their neighboring words (Figure 5). A model then learns to predict the next word given a set of words and generates a set of probabilities for what the next word will be. This unsupervised approach learns patterns from the distribution of words in the corpus. From these distributions, these word2vec models encode context because they predict the next word given a set of words. For RNA, the "words" are the nucleotides and similar sequences are grouped together. The word2vec representation can then be fine-tuned for downstream tasks such as sequence-function mapping. This representation can be particularly valuable for encoding long-range dependencies that are more challenging to encode in a *k*-mer representation.

When non-natural nucleotides or chemically functionalized nucleotides are used in sequencing experiments, they produce a challenge to the one-hot, *k*-mer, or word2vec representations if not handled carefully. Here, the modified nucleobases are chemically distinct from the canonical nucleobases, but share similarities. Therefore, we suggest that encoding chemical information about the nucleobases via a molecular graph (i.e. the atoms and bonds of the nucleobase) can provide a novel strategy to learn the underlying trends with quantitative sequence-function maps. With this strategy, similarity to other nucleobases (i.e. adenine and hypoxanthine, the nucleobase in inosine, have similar structural characteristics) could be utilized

while distinguishing the two compounds (Figure 5). We anticipate that this strategy will be challenged by the large connectivity graph sizes for long RNA sequences that have hundreds or thousands of atoms. More abstract graph-based representations (where nodes represent nucleotides) are essential for encoding secondary structure such as base-pairing (80, 110) (Figure 5). These base-pairing interactions, essential for forming structures such as hairpins, may be essential for catalytic activity; a graph-based representation allows a model to harness this information.

A recent study applying ML to an F1*U ligase ribozyme has successfully introduced deep learning (i.e. deep neural networks) into RNA sequence exploration to find peaks on the fitness landscape, as quantified by relative ligation activity (94) . In particular, this study used in silico selection, recombination, and mutation to find paths that are free of epistasis, which challenge ML models. These data were then successfully used to train a deep neural network to predict and identify functional mutational variants that have comparable activity to the wild type. These models enable evaluation of the paths between the fitness peaks, enabling a comparison between genotype (sequence) and corresponding phenotype (fitness). Due to the rarity of high fitness regions and the presence of epistatic (i.e. nonadditive) effects that can lead to activity cliffs (15), supervised ML models are challenged by data bias, since the majority of data will be from low- or moderate-fitness regions of the fitness landscape. By incorporating information from rounds of in vitro selection, the underlying data is not biased solely towards deleterious mutants, enabling predictive ML model training. Aside from ribozymes, similar approaches have been used for DNA sequences that bind carbon nanotubes (55). Here, ML models were trained to identify the binding response of a sequence to serotonin. These models were subsequently used to identify low- and high-fitness sequences for serotonin binding. This ML-driven approach led

to the discovery of five new DNA-carbon nanotube conjugates that had higher intensity response to serotonin than the best previously identified combinations.

Generative models have seen recent use for RNA sequence discovery due to their ability to learn the underlying trends and generate novel sequences for testing. In particular, the recent use of Restricted Boltzmann Machines (29) (RBMs) and Hidden Markov Models (47) (HMM) in neural networks has seen increased use due to interpretability and sequence suggestions. Here, architectures used are similar to traditional variational autoencoders (VAEs), which are unsupervised models (i.e. models that do not require data labels) that encode sequences into an information-rich, low-dimensional "latent space" (i.e. an abstract vector space of arbitrary dimension that positions chemically similar sequences near each other) and then decode the latent space back into a sequence. If the latent space accurately maps the fitness landscape, decoding peaks on the landscape can lead to the discovery of high fitness regions. While training semi-supervised (i.e. with a subset of sequences labeled with properties) variants of VAEs known as a conditional VAEs (CVAEs), this approach has led to the discovery of new catalytic sequences. In recent work (18), CVAEs successfully identified novel RNA-like polymers called highly functionalized nucleic acid polymers (HFNAPs) by using the binding affinity to daunomycin as a proxy for fitness. As generative models improve, they can improve sampling of RNA sequence space to accelerate the discovery of highly active catalytic ribozymes.

Concluding Remarks

Studies of fitness landscapes of RNA molecules were recently revolutionized by high-throughput sequencing, which enabled quantitative assays of large numbers of sequences. When coupled with in vitro selection, significant insights can be gained about these fitness landscapes.

Important areas for future study in this area include probing how the environment, especially dynamic environments, change the fitness landscape, and understanding the structure of protocellular fitness landscapes and their relationship with molecular fitness landscapes. Given the absolute need for interpolation and extrapolation to map fitness landscapes for molecules with greater than a couple dozen nucleotides, machine learning, like HTS before it, may be poised to enable major discoveries in this area. New representations, such as molecular graphs, could help advance these methods by encoding relevant features. Along with close attention to opportunities to gain not only predictive power but also scientific understanding, the field may soon realize the molecular cartographer's dream (53): maps of fitness landscapes to guide synthetic biology.

Disclosure Statement

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Acknowledgments

The authors thank Ulrich Müller for insights on ribozyme fitness landscapes. Funding from the Simons Foundation Collaboration on the Origin of Life (290356FY18), NASA (80NSSC21K0595), NSF (EF-1935372, EF-1935087), Sloan Foundation (G-2022-19518), and Moore Foundation (11479) is acknowledged. AN gratefully acknowledges the support of the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program. AVS's research was supported by an appointment to the NASA Postdoctoral Program, administered by Oak Ridge Associated Universities under contract with NASA.

References

1. Adamala K, Szostak JW. 2013. Competition between model protocells driven by an encapsulated catalyst. *Nat. Chem.* 5: 495-501
2. Adamala K, Szostak JW. 2013. Nonenzymatic template-directed RNA synthesis inside model protocells. *Science* 342: 1098-100
3. Adamski P, Eleveld M, Sood A, Kun Á, Szilágyi A, et al. 2020. From self-replication to replicator systems en route to de novo life. *Nat. Rev. Chem.* 4: 386-403
4. Agresti JJ, Kelly BT, Jäschke A, Griffiths AD. 2005. Selection of ribozymes that catalyse multiple-turnover Diels–Alder cycloadditions by using in vitro compartmentalization. *Proc. Natl. Acad. Sci. U.S.A.* 102: 16170-75
5. Ameta S, Winz M-L, Previti C, Jäschke A. 2014. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Res.* 42: 1303-10
6. Andreasson JOL, Savinov A, Block SM, Greenleaf WJ. 2020. Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the GLMS ribozyme. *Nat. Commun.* 11: 1663
7. Angenent-Mari NM, Garruss AS, Soenksen LR, Church G, Collins JJ. 2020. A Deep learning approach to programmable RNA switches. *Nat. Commun.* 11: 5057
8. Attwater J, Wochner A, Pinheiro VB, Coulson A, Holliger P. 2010. Ice as a protocellular medium for RNA replication. *Nat. Commun.* 1: 76
9. Bansho Y, Furubayashi T, Ichihashi N, Yomo T. 2016. Host–parasite oscillation dynamics and evolution in a compartmentalized RNA replication system. *Proc. Natl. Acad. Sci. U. S. A.* 113: 4045-50
10. Benner SA. 2023. Rethinking nucleic acids from their origins to their applications. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 378: 20220027
11. Bernhardt HS. 2012. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biol. Direct* 7: 23
12. Bernhardt HS, Tate WP. 2012. Primordial soup or vinaigrette: did the RNA world evolve at acidic pH? *Biol. Direct* 7: 4
13. Blain JC, Szostak JW. 2014. Progress toward synthetic cells. *Annu. Rev. Biochem.* 83: 615-40
14. Blanco C, Janzen E, Pressman A, Saha R, Chen IA. 2019. Molecular fitness landscapes from high-coverage sequence profiling. *Annu. Rev. Biophys.* 48: 1-18
15. Charest N, Shen Y, Lai Y-C, Chen IA, Shea J-E. 2023. Discovering pathways through ribozyme fitness landscapes using information theoretic quantification of epistasis. *bioRxiv*: 10.1101/2023.05.22.541765
16. Chen IA, Roberts RW, Szostak JW. 2004. The emergence of competition between model protocells. *Science* 305: 1474-76
17. Chen IA, Salehi-Ashtiani K, Szostak JW. 2005. RNA catalysis in model protocell vesicles. *J. Am. Chem. Soc.* 127: 13213-19
18. Chen JC, Chen JP, Shen MW, Wornow M, Bae M, et al. 2022. Generating experimentally unrelated target molecule-binding highly functionalized nucleic-acid polymers using machine learning. *Nat. Commun.* 13: 4541
19. Curtis EA. 2022. Pushing the limits of nucleic acid function. *Chem. Eur. J.* 28: e202201737

20. Curtis EA, Bartel DP. 2013. Synthetic shuffling and in vitro selection reveal the rugged adaptive fitness landscape of a kinase ribozyme. *RNA* 19: 1116-28
21. Daher M, Widom JR, Tay W, Walter NG. 2018. Soft interactions with model crowders and non-canonical interactions with cellular proteins stabilize RNA folding. *J. Mol. Biol.* 430: 509-23
22. Das SG, Direito SOL, Waclaw B, Allen RJ, Krug J. 2020. Predictable properties of fitness landscapes induced by adaptational tradeoffs. *eLife* 9: e55155
23. DasGupta S, Zhang S, Szostak JW. 2023. Molecular crowding facilitates ribozyme-catalyzed RNA assembly. *bioRxiv*: 2023.04.30.538884
24. DasGupta S, Zhang SJ, Smela MP, Szostak JW. 2023. RNA-catalyzed RNA ligation within prebiotically plausible model protocells. *Chem.Eur. J.*: e202301376
25. de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* 15: 480-90
26. Deck C, Jauker M, Richert C. 2011. Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA. *Nat. Chem.* 3: 603-08
27. Desai R, Kilburn D, Lee H-T, Woodson SA. 2014. Increased ribozyme activity in crowded solutions. *J. Biol. Chem.* 289: 2972-77
28. Dhamodharan V, Kobori S, Yokobayashi Y. 2017. Large scale mutational and kinetic analysis of a self-hydrolyzing deoxyribozyme. *ACS Chem. Biol.* 12: 2940-45
29. Di Gioacchino A, Procyk J, Molari M, Schreck JS, Zhou Y, et al. 2022. Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLoS Comput. Biol.* 18: e1010561
30. Di Giulio M. 1997. On the RNA world: evidence in favor of an early ribonucleopeptide world. *J. Mol. Evol.* 45: 571-78
31. Dockerill M, Winssinger N. 2023. DNA-encoded libraries: towards harnessing their full power with Darwinian evolution. *Angew. Chem., Int. Ed.* 62: e202215542
32. El Allali A, Elhamraoui Z, Daoud R. 2021. Machine learning applications in RNA modification sites prediction. *Comput. Struct. Biotechnol. J* 19: 5510-24
33. Engelhart AE, Adamala KP, Szostak JW. 2016. A simple physical mechanism enables homeostasis in primitive cells. *Nat. Chem.* 8: 448-53
34. Fay MM, Lyons SM, Ivanov P. 2017. RNA G-quadruplexes in biology: principles and molecular mechanisms. *J. Mol. Biol.* 429: 2127-47
35. Fisette J-F, Montagna DR, Mihailescu M-R, Wolfe MS. 2012. A G-Rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing. *J. Neurochem.* 121: 763-73
36. Fragata I, Blanckaert A, Dias Louro MA, Liberles DA, Bank C. 2019. Evolution in the light of fitness landscape theory. *Trends Ecol. Evol.* 34: 69-82
37. Frank SA, Slatkin M. 1992. Fisher's fundamental theorem of natural selection. *Trends Ecol. Evol.* 7: 92-95
38. Gebicki JM, Hicks M. 1976. Preparation and properties of vesicles enclosed by fatty acid membranes. *Chem. Phys. Lipids* 16: 142-60
39. Gould SJ. 2002. *The structure of evolutionary theory*. Cambridge, MA: Harvard University Press
40. Guo Y, Chen J, Cheng M, Monchaud D, Zhou J, Ju H. 2017. A thermophilic tetramolecular G-quadruplex/hemin dnazyme. *Angew. Chem., Int. Ed.* 56: 16636-40

41. Halder K, Wieland M, Hartig JS. 2009. Predictable suppression of gene expression by 5′-UTR-based RNA quadruplexes. *Nucleic Acids Res.* 37: 6811-17
42. Hanczyc MM, Fujikawa SM, Szostak JW. 2003. Experimental models of primitive cellular compartments: Encapsulation, growth, and division. *Science* 302: 618-22
43. Hershberg R. 2015. Mutation—the engine of evolution: studying mutation and its role in the evolution of bacteria. *Cold Spring Harb. Perspect. Biol.* 7: a018077
44. Hyman AA, Weber CA, Jülicher F. 2014. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* 30: 39-58
45. Ichihashi N, Usui K, Kazuta Y, Sunami T, Matsuura T, Yomo T. 2013. Darwinian evolution in a translation-coupled RNA replication system within a cell-like compartment. *Nat. Commun.* 4: 2494
46. Iglesias-Artola JM, Drobot B, Kar M, Fritsch AW, Mutschler H, et al. 2022. Charge-density reduction promotes ribozyme activity in RNA–peptide coacervates via RNA fluidization and magnesium partitioning. *Nat. Chem.* 14: 407-16
47. Iwano N, Adachi T, Aoki K, Nakamura Y, Hamada M. 2022. Generative aptamer discovery using RaptGen. *Nat. Comput. Sci.* 2: 378-86
48. Jalali-Yazdi F, Huong Lai L, Takahashi TT, Roberts RW. 2016. High-throughput measurement of binding kinetics by mRNA display and next-generation sequencing. *Angew. Chem., Int. Ed.* 55: 4007-10
49. Janzen E, Shen Y, Vázquez-Salazar A, Liu Z, Blanco C, et al. 2022. Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes. *Nat. Commun.* 13: 3631
50. Jia TZ, Hentrich C, Szostak JW. 2014. Rapid RNA exchange in aqueous two-phase system and coacervate droplets. *Orig. Life Evol. Biosph.* 44: 1-12
51. Jin L, Kamat NP, Jena S, Szostak JW. 2018. Fatty acid/phospholipid blended membranes: a potential intermediate state in protocellular evolution. *Small* 14: 1704077
52. Jones CP, Ferré-D'Amaré AR. 2015. RNA quaternary structure and global symmetry. *Trends Biochem. Sci.* 40: 211-20
53. Joyce GF, Orgel LE. 1993. Prospects for understanding the origin of the RNA world. In *The RNA world*, ed. RF Gesteland, JF Atkins, pp. 1-25. New York: Cold Spring Harbor Laboratory Press
54. Joyce GF, Szostak JW. 2018. Protocells and RNA Self-replication. *Cold Spring Harb. Perspect Biol.* 10
55. Kelich P, Jeong S, Navarro N, Adams J, Sun X, et al. 2021. Discovery of DNA–carbon nanotube sensors for serotonin with machine learning and near-infrared fluorescence spectroscopy. *ACS Nano* 16: 736-45
56. Kilburn D, Roh JH, Behrouzi R, Briber RM, Woodson SA. 2013. Crowders perturb the entropy of RNA energy landscapes to favor folding. *J. Am. Chem. Soc.* 135: 10055-63
57. Kilburn D, Roh JH, Guo L, Briber RM, Woodson SA. 2010. Molecular crowding stabilizes folded RNA structure by the excluded volume effect. *J. Am. Chem. Soc.* 132: 8690-96
58. Kobori S, Nomura Y, Miu A, Yokobayashi Y. 2015. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res.* 43: e85-e85
59. Kobori S, Takahashi K, Yokobayashi Y. 2017. Deep sequencing analysis of aptazyme variants based on a Pistol ribozyme. *ACS Synth. Biol.* 6: 1283-88

60. Kobori S, Yokobayashi Y. 2016. High-throughput mutational analysis of a twister ribozyme. *Angew. Chem. Int. Ed. Engl.* 55: 10354-7
61. Kohlberger M, Gadermaier G. 2021. SELEX: critical factors and optimization strategies for successful aptamer selection. *Biotechnol. Appl. Biochem.* 69: 1771-92
62. Kumari S, Bugaut A, Huppert JL, Balasubramanian S. 2007. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.* 3: 218-21
63. Lai Y-C, Chen IA. 2020. Protocells. *Curr. Biol.* 30: R482-R85
64. Lai Y-C, Liu Z, Chen IA. 2021. Encapsulation of ribozymes inside model protocells leads to faster evolutionary adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 118: e2025054118
65. Lawrence MS, Bartel DP. 2005. New ligase-derived RNA polymerase ribozymes. *RNA* 11: 1173-80
66. Le Vay K, Salibi E, Song EY, Mutschler H. 2020. Nucleic acid catalysis under potential prebiotic conditions. *Chem. Asian J.* 15: 214-30
67. Le Vay K, Song EY, Ghosh B, Tang T-YD, Mutschler H. 2021. Enhanced ribozyme-catalyzed recombination and oligonucleotide assembly in peptide-RNA condensates. *Angew. Chem., Int. Ed.* 60: 26096-104
68. Lee H-T, Kilburn D, Behrouzi R, Briber RM, Woodson SA. 2015. Molecular crowding overcomes the destabilizing effects of mutations in a bacterial ribozyme. *Nucleic Acids Res.* 43: 1170-76
69. Lewontin RC. 1970. The units of selection. *Annu. Rev. Ecol. Evol. Syst.* 1: 1-18
70. Mansy SS. 2010. Membrane transport in primitive cells. *Cold Spring Harb. Perspect. Biol.* 2: a002188
71. Matsumoto S, Tateishi-Karimata H, Takahashi S, Ohyama T, Sugimoto N. 2020. Effect of molecular crowding on the stability of RNA G-quadruplexes with various numbers of quartets and lengths of loops. *Biochemistry* 59: 2640-49
72. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. *Distributed representations of words and phrases and their compositionality*. Presented at Adv. Neural Inf. Process.
73. Müller F, Escobar L, Xu F, Węgrzyn E, Nainytė M, et al. 2022. A prebiotically plausible scenario of an RNA–peptide world. *Nature* 605: 279-84
74. Müller UF, Bartel DP. 2008. Improved polymerase ribozyme efficiency on hydrophobic assemblies. *RNA* 14: 552-62
75. Nakashima KK, van Haren MHI, André AAM, Robu I, Spruijt E. 2021. Active coacervate droplets are protocells that grow and resist Ostwald ripening. *Nat. Commun.* 12: 3819
76. Nakashima KK, Vibhute MA, Spruijt E. 2019. Biomolecular chemistry in liquid phase separated compartments. *Front. Mol. Biosci.* 6: 21
77. Nehdi A, Perreault J-P. 2006. Unbiased in vitro selection reveals the unique character of the self-cleaving antigenomic HDV RNA sequence. *Nucleic Acids Res.* 34: 584-92
78. Niland CN, Jankowsky E, Harris ME. 2016. Optimization of high-throughput sequencing kinetics for determining enzymatic rate constants of thousands of RNA substrates. *Anal. Biochem.* 510: 1-10
79. Nomura Y, Yokobayashi Y. 2019. Systematic minimization of RNA ligase ribozyme through large-scale design-synthesis-sequence cycles. *Nucleic Acids Res.* 47: 8950-60

80. Noviello TMR, Ceccarelli F, Ceccarelli M, Cerulo L. 2020. Deep learning predicts short non-coding RNA functions from only raw sequence data. *PLoS Comput. Biol.* 16: e1008415
81. Nowak MA, Highfield R. 2011. *SuperCooperators : altruism, evolution, and why we need each other to succeed*. New York: Free Press
82. Obolski U, Ram Y, Hadany L. 2018. Key issues review: evolution on rugged adaptive landscapes. *Rep. Prog. Phys.* 81: 012602
83. Oparin AI. 1938. *The origin of life (English Translation)*. New York: Macmillan
84. Pan X, Yang Y, Xia CQ, Mirza AH, Shen HB. 2019. Recent methodology progress of deep learning for RNA–protein interaction prediction. *WIREs RNA* 10
85. Paudel BP, Rueda D. 2014. Molecular crowding accelerates ribozyme docking and catalysis. *J. Am. Chem. Soc.* 136: 16700-03
86. Payne JL, Wagner A. 2019. The causes of evolvability and their evolution. *Nat. Rev. Genet.* 20: 24-38
87. Peng H, Lelievre A, Landenfeld K, Müller S, Chen IA. 2022. Vesicle encapsulation stabilizes intermolecular association and structure formation of functional RNA and DNA. *Curr. Biol.* 32: 86-96.e6
88. Peri G, Gibard C, Shults NH, Crossin K, Hayden EJ. 2022. Dynamic RNA fitness landscapes of a group I ribozyme during changes to the experimental environment. *Mol. Biol. Evol.* 39
89. Pinheiro VB, Arangundy-Franklin S, Holliger P. 2014. Compartmentalized self-tagging for in vitro-directed evolution of XNA polymerases. *Curr. Protoc. Nucleic Acid Chem.* 57: 9.9.1-9.9.18
90. Pitt JN, Ferré-D'Amaré AR. 2010. Rapid construction of empirical RNA fitness landscapes. *Science* 330: 376-79
91. Pressman A, Blanco C, Chen IA. 2015. The RNA world as a model system to study the origin of life. *Curr. Biol.* 25: R953-R63
92. Pressman AD, Liu Z, Janzen E, Blanco C, Müller UF, et al. 2019. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J. Am. Chem. Soc.* 141: 6213-23
93. Rivas G, Minton AP. 2016. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* 41: 970-81
94. Rotrattanadumrong R, Yokobayashi Y. 2022. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. *Nat. Commun.* 13: 4847
95. Rushdi A, Simoneit BT. 2001. Lipid formation by aqueous Fischer-Tropsch-type synthesis over a temperature range of 100 to 400 °C. *Orig. Life Evol. Biosph* 31: 103-18
96. Saha R, Pohorille A, Chen IA. 2015. Molecular crowding and early evolution. *Orig. Life Evol. Biosph.* 44: 319-24
97. Saha R, Verbanic S, Chen IA. 2018. Lipid vesicles chaperone an encapsulated RNA aptamer. *Nat. Commun.* 9: 2313
98. Salibi E, Peter B, Schwille P, Mutschler H. 2023. Periodic temperature changes drive the proliferation of self-replicating RNAs in vesicle populations. *Nat. Commun.* 14: 1222
99. Shen Y, Pressman A, Janzen E, Chen IA. 2021. Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Res.* 49: e67-e67

100. Strulson CA, Molden RC, Keating CD, Bevilacqua PC. 2012. RNA catalysis through compartmentalization. *Nat. Chem.* 4: 941-46
101. Szostak JW, Bartel DP, Luisi PL. 2001. Synthesizing life. *Nature* 409: 387-90
102. Turk RM, Chumachenko NV, Yarus M. 2010. Multiple translational products from a five-nucleotide ribozyme. *Proc. Natl. Acad. Sci. U.S.A.* 107: 4585-89
103. Tyrrell J, Weeks KM, Pielak GJ. 2015. Challenge of mimicking the influences of the cellular environment on RNA structure by PEG-induced macromolecular crowding. *Biochemistry* 54: 6447-53
104. Vay KL, Salibi E, Ghosh B, Tang T-YD, Mutschler H. 2022. Ribozyme-phenotype coupling in peptide-based coacervate protocells. *bioRxiv*: 2022.10.25.513667
105. Walde P, Wick R, Fresta M, Mangone A, Luisi PL. 1994. Autopoietic self-reproduction of fatty acid vesicles. *J. Am. Chem. Soc.* 116: 11649-54
106. Wang Q, Schlenoff JB. 2014. The polyelectrolyte complex/coacervate continuum. *Macromolecules* 47: 3108-16
107. Werren JH. 2011. Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl. Acad. Sci. U. S. A.* 108: 10863-70
108. Wilson DS, Szostak JW. 1999. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* 68: 611-47
109. Wright S. 1932. *The roles of mutation, inbreeding, crossbreeding and selection in evolution*. Presented at Proceedings of the sixth international congress of genetics
110. Yan Z, Hamilton WL, Blanchette M. 2020. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics* 36: 276-84
111. Yang KK, Wu Z, Arnold FH. 2019. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* 16: 687-94
112. Yokobayashi Y. 2019. Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods* 161: 41-45
113. Yokobayashi Y. 2020. High-throughput analysis and engineering of ribozymes and deoxyribozymes by sequencing. *Acc. Chem. Res.* 53: 2903-12
114. Zhou H-X, Dill KA. 2001. Stabilization of proteins in confined spaces. *Biochemistry* 40: 11289-93
115. Zhu TF, Szostak JW. 2009. Coupled growth and division of model protocell membranes. *J. Am. Chem. Soc.* 131: 5705-13

Reference Annotations

18. Chen et al. Semi-supervised generative models learned from experimental labels to discover novel daunomycin-binding sequences.

48. Jalali-Yazdi et al. Demonstrated how to use HTS to measure binding constants for many peptides in parallel.

55. Kelich et al. ML models learned from analytical responses (not just selection abundance) to improve DNA-nanotube biosensors.

64. Lai et al. Vesicle encapsulation increased fitness differences between ribozyme sequences, 'sharpening' the landscape and accelerating evolutionary adaptation.

67. Le Vay et al. Peptide/RNA coacervates enabled ribozyme catalysis, illustrating an RNA world function for peptides.

68. Lee et al. Macromolecular crowding increased mutational tolerance, 'flattening' the ribozyme fitness landscape.

69. Lewontin. Classic work explaining natural selection at multiple levels, from molecules and coacervates to populations.

88. Peri et al. Explores effect of environmental changes on a ribozyme fitness landscape.

92. Pressman et al. First complete map of a fitness landscape for ribozymes.

99. Shen et al. Addresses rigor and reproducibility of HTS assay technique and suggests best practices.