







Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes

Evan Janzen ^{1,2}, Yuning Shen ^{2,3}, Alberto Vázquez-Salazar³, Ziwei Liu ⁴, Celia Blanco ³,
Josh Kenchel ^{1,2,3} & Irene A. Chen ^{1,2,3}✉

Systems of catalytic RNAs presumably gave rise to important evolutionary innovations, such as the genetic code. Such systems may exhibit particular tolerance to errors (error minimization) as well as coding specificity. While often assumed to result from natural selection, error minimization may instead be an emergent by-product. In an RNA world, a system of self-aminoacylating ribozymes could enforce the mapping of amino acids to anticodons. We measured the activity of thousands of ribozyme mutants on alternative substrates (activated analogs for tryptophan, phenylalanine, leucine, isoleucine, valine, and methionine). Related ribozymes exhibited shared preferences for substrates, indicating that adoption of additional amino acids by existing ribozymes would itself lead to error minimization. Furthermore, ribozyme activity was positively correlated with specificity, indicating that selection for increased activity would also lead to increased specificity. These results demonstrate that by-products of ribozyme evolution could lead to adaptive value in specificity and error tolerance.

¹Program in Biomolecular Science and Engineering, University of California, Santa Barbara, CA 93106, USA. ²Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA 93106, USA. ³Department of Chemical and Biomolecular Engineering, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA. ⁴MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. ✉email: ireneachen@ucla.edu

The origin of life is believed to have progressed through an RNA World in which ribozymes catalyzed critical biochemical reactions^{1,2}. In principle, ribozymes performing new functions could arise either by chance from a pool of random sequence molecules, or by adaptation of pre-existing ribozymes having promiscuous activities accessible through zero or a small number of mutations. Co-option of a pre-existing sequence (i.e., utilizing an existing sequence for a new reaction or function; also called exaptation) is a well-established mechanism for evolutionary innovation^{3–8}. Gene duplication coupled with co-option could lead to a more complex system as the ribozymes adopt additional substrates⁹. However, the degree to which the evolution of complex systems in the RNA World would rely on chance vs. co-option, and potential consequences of the co-option process, are unclear¹⁰.

Systems of ribozymes could form the basis for important aspects of prebiotic evolution, such as the early stages of the genetic code of protein translation, a ‘major evolutionary transition’¹¹. In modern biology, the mapping of specific codons to their cognate amino acids is assured through the aminoacylation of tRNAs by aminoacyl-tRNA synthetase (aaRS) proteins^{12–14}. However, a rudimentary form of these functions was presumably performed by ribozymes. Indeed, evolutionary analysis of the aaRS proteins indicates that these enzymes evolved after the establishment of a primitive genetic code^{15–19} and have heterogeneous genetic origins²⁰. Several ribozymes catalyzing aminoacylation reactions have been discovered by *in vitro* selection, including self-aminoacylating RNAs^{21–26}. Although these ribozymes do not necessarily mimic a precursor to the translation apparatus of modern biology (see Discussion), these ribozymes might still serve as a model for understanding emergent properties of such systems.

An important feature of evolved biochemical systems is robustness to errors. For example, in the context of the genetic code, non-synonymous point mutations tend to result in amino acid substitutions that conserve chemical properties^{27–31}. This ‘error minimization’ confers a clear selective advantage as it reduces the deleterious impact of mutations on the resultant protein^{32,33}. At the same time, the standard genetic code does not appear to be particularly optimal with respect to error minimization^{34–37}. This raises a fundamental open question about the origin of error minimization, namely, whether it is solely a direct product of natural selection to reduce the impact of errors, as opposed to a serendipitous by-product of evolution or emergence of the system (e.g., evolution favoring incorporation of additional amino acids to expand the genetic code)²⁸. In other words, while direct natural selection for error minimization is possible, it may be also possible that the process of developing a ribozyme system involves an evolutionary mechanism that happens to reduce the chemical consequences of errors, without direct natural selection to minimize the consequences of errors^{38,39}.

In this work, we evaluate the evolutionary potential of self-aminoacylating ribozymes to adopt new amino acid substrates. We previously used *in vitro* selection and high-throughput sequencing to exhaustively search RNA sequence space (21 nt) for self-aminoacylating ribozymes²⁴. These ribozymes were originally selected to react with biotinyl-Tyr(Me)-oxazolone (BYO), a chemically activated amino acid. The 5(4*H*)-oxazolones and related *N*-carboxyanhydrides can be made abiotically under prebiotically plausible conditions^{40–48}. Three distinct, evolutionarily unrelated catalytic motifs had been discovered from the exhaustive search. Here we determine the ability of these ribozymes to use different substrates, by measuring the activity of all single- and double-mutants of five ribozymes, representing the three catalytic motifs, for six alternative substrates, using a massively parallel assay

(*k*-Seq^{24,49}). This assay and related techniques leverage high-throughput sequencing to measure the activity of thousands of candidate sequences in a mixed pool^{50–53}. The six substrates (analogs of tryptophan, phenylalanine, leucine, isoleucine, valine, and methionine) represent a range of sizes and biochemical classes (aromatic, aliphatic, sulfur-containing), as well as amino acids thought to be early (Leu, Ile, Val) and late (Trp, Phe, Met) incorporations into the genetic code^{54–58}. Because of this span, the chosen amino acids should be considered model systems to study trends in rate enhancement, specificity, and proximity of ribozymes in sequence space, rather than as a detailed model of the early prebiotic emergence of the genetic code. Our findings indicate extensive opportunities for the ribozymes to incorporate new substrates into the system (co-option). In addition, we describe two major by-products of evolution of these ribozymes. First, a positive correlation between activity and specificity was observed, indicating that greater specificity would be a by-product of selection for greater activity. Second, related ribozymes react with chemically similar amino acids, suggesting that expansion of the code by co-option would incorporate a chemically similar amino acid into the system, with error minimization arising as a by-product. Such effects could favor the emergence of a complex biochemical system.

Results

Aminoacylation substrates and design of the ribozyme mutant pool.

To investigate whether ribozymes previously selected for aminoacylation with BYO (tyrosine analog) would react with substrates having other aminoacyl side chains, six additional biotinyl-aminoacyl oxazolones were synthesized for analysis (Fig. 1A): tryptophanyl (BWO), phenylalanyl (BFO), leucyl (BLO), isoleucyl (BIO), valyl (BVO), and methionyl (BMO). This set of substrates represents three chemical classes (small hydrophobic, aromatic, and sulfur-containing). Within the group of small hydrophobic side chains, both β -branched and -unbranched residues were included. The set includes side chains that are considered early as well as side chains that are considered late additions to the genetic code^{54–58}. In particular, aromatic residues, of which two were chosen to assess specificity within the class, are thought to have been added relatively late. The span over chemical space as well as putative prebiotic age of the substrates therefore probes general trends rather than a specific epoch during the emergence of translation. In order to assess the generality of any observed trend, five wild-type ribozymes were chosen for analysis, representing five different families containing three unrelated motifs (Supplementary Table 1).

Compounds were synthesized using previously described methods²⁴ and verified by NMR spectroscopy (see Methods). An initial test by a streptavidin gel shift assay at high substrate concentration (500 μ M) indicated that each oxazolone served as substrate for at least one ribozyme tested, although the two tested ribozymes (S-1A.1-a and S-2.1-a) differed in selectivity (Fig. 1B and Supplementary Fig. 11). For these ribozymes, reaction products were not observed when a single residue (G65 and G54, respectively) was chemically modified to 2'-O-methyl, indicating that reaction occurs at a single site (Supplementary Fig. 12A). This observation is consistent with previously reported results for these ribozymes reacting with BYO²⁴. This result was further confirmed by direct PAGE analysis of the reaction product (without streptavidin added) under acidic conditions, which shows a mobility difference between aminoacylated RNA and unreacted RNA (Supplementary Fig. 12B)⁵⁹.

To study the cross-reactivity of these ribozymes and their mutants systematically, pools of sequence variants were designed to explore the sequence space around five sequences representing

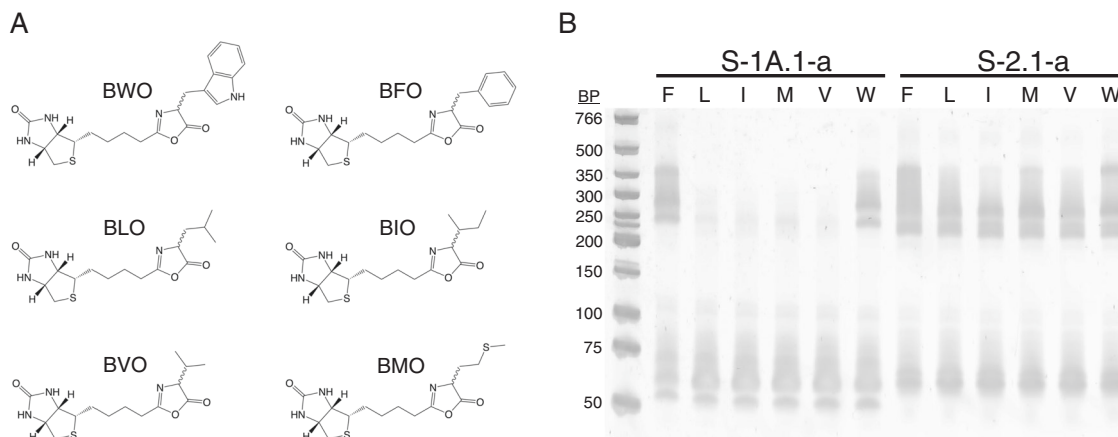


Fig. 1 Aminoacylation activity of two ribozymes with BXO substrates. **A** Biotinyl aminoacyl oxazolones (BXO) used in this study: tryptophanyl (BWO), phenylalanyl (BFO), leucyl (BLO), isoleucyl (BIO), valyl (BVO), and methionyl (BMO). **B** Aminoacylation activity of two ribozymes (S-1A.1-a, the center of Family 1A.1, and S-2.1-a, the center of Family 2.1) with BXO substrates analyzed by streptavidin gel shift on native PAGE (X = F, L, I, M, V, or W, as indicated; see Methods for details). Reactions were conducted for 90 min at 500 μ M BXO. The reacted RNA is detected by its slower migration through the gel due to complexation with streptavidin. This experiment was conducted once; also see Fig. S11. Since a single reactive site was identified (Fig. S12), multiple bands on native gels may be caused by the presence of multiple conformers or streptavidin oligomers. MW markers are the Low Molecular Weight Ladder (dsDNA) from NEB. Structures were drawn using ChemDraw 19.0. Source data are provided as a Source Data file.

each of the major ribozyme families obtained from the selection on BYO (Supplementary Table 1). The ribozyme families chosen for testing include all of the previously discovered motifs (Motifs 1, 2, and 3), specifically the two most abundant families containing Motif 1 (Family 1A.1 and 1B.1) and Motif 2 (Family 2.1 and 2.2), as well as the only family identified from Motif 3 (Family 3.1). These ribozyme families had been discovered during an exhaustive search of sequence space varying a central 21-mer region, and sequences from these motifs had comprised ~80% of the selected pool²⁴. Sequencing of the variant pool showed that it included 13.5% of the unique sequences from the originally selected pool (having abundance $\geq 10^{-6}$). Thus, the variant pool, based on these five ribozyme families, was designed to be representative of ribozymes having aminoacylation activity.

These ribozymes had been identified through selection with substrate BYO. To probe the sequence space for additional motifs, we also performed *in vitro* selections using substrates BFO and BLO, starting from libraries with completely random 21-mer variable regions. These selections followed a process identical to the original selection with the exception of the substrate compound. All families found in the BFO and BLO selections had already been identified in the earlier BYO selection (Supplementary Fig. 1). Interestingly, selection with BLO resulted predominantly in sequences containing Motif 2, consistent with the low activity of a Family 1A.1 ribozyme on BLO observed in the gel shift assay (Fig. 1B). While it is possible that cross-contamination of sequences from prior selection with BYO in the laboratory could bias the results of these selections, the failure to identify new motifs indicated a lack of new ribozymes having appreciably greater activity on BFO or BLO, suggesting that the designed pool of variants would likely include major motifs of the active sequence space.

Cross-reaction of self-aminoacylating ribozyme mutants with alternative side chains. Sequences in the ribozyme variant pool were assayed for activity on each alternative substrate in a massively parallel format by kinetic sequencing (*k*-Seq)^{24,49,60}. During *k*-Seq, a pool containing thousands of candidate ribozymes was reacted with a substrate at multiple concentrations. The reacted molecules, having been biotinylated through reaction, were isolated by streptavidin binding and then sequenced on the

Illumina platform. Quantitation of the reacted fraction was used to fit to a kinetic model to determine ribozyme activity. Data obtained from this method have been shown to correlate well with traditional biochemical assays with confidence intervals of the measurements obtained by experimental replicates and bootstrapping⁴⁹. In each *k*-Seq experiment, one of six BXO (X = W, F, L, I, V, or M) substrates was tested to measure reaction kinetics for sequences in the pool. Samples were exposed to substrate concentrations from 2 to 1250 μ M in triplicate. Reaction data were fit to a pseudo-first-order kinetic model ($F_s^{BXO} = A_s(1 - e^{-k_s[BXO]_{at}})$), with maximum reaction amplitude A_s and rate constant k_s for sequence s , where F_s^{BXO} is the fraction of RNA that is aminoacylated with substrate BXO, $[BXO]$ is the initial substrate concentration, t is the reaction time (90 min), and α is the coefficient accounting for substrate hydrolysis during the reaction. The product $k_s A_s$, reflecting ribozyme activity at non-saturating conditions, was accurately estimated across a wide range of activities^{24,49} (Supplementary Fig. 2). The data yielded $k_s A_s$ estimates for a total of 9,770 sequences, encompassing five family wild-type sequences and a complete set of both single and double mutants related to the five wild-type ribozymes (Supplementary Fig. 3).

k-Seq measures the combination of catalyzed and non-catalyzed (background) reactions. To measure the catalytic activity of the RNA, the nonspecific background reactivity of the substrate with RNA should be taken into account. In analogy with catalytic power used to characterize enzymes, we determined catalytic enhancement, i.e., the ratio of catalyzed to background reaction rates. Since RNA sequences were being compared against each other, it was natural to use the reactivity of the substrate with bulk, low-activity RNA from the pool as the background reaction rate. We measured the rate of the background reaction for BFO by gel shift assay with the randomized RNA library. The background rate was $0.55 \pm 0.18 \text{ M}^{-1} \text{ min}^{-1}$ ($\mu \pm \sigma$), which is within error to that measured previously for BYO ($0.65 \pm 0.28 \text{ M}^{-1} \text{ min}^{-1}$)²⁴. Comparing to the frequency distribution of $k_s A_s$ measured by *k*-Seq (Supplementary Fig. 4, Supplementary Table 2), the measured background rate was found to correspond to the center of a low-activity peak, indicating that this peak represented a background of catalytically inactive, or nearly inactive, mutants. This is consistent with observations that

individual Motif 1 ribozymes display little activity with some substrates at high concentration when analyzed by a gel-shift assay (Fig. 1B and Supplementary Fig. 11). The low-activity peak was therefore used as an internal control in *k*-Seq, and the effective background reaction rate (k_0A_0) of each substrate was estimated as the center of this peak. k_sA_s values for sequences reacted with each substrate were normalized by the corresponding k_0A_0 to obtain the catalytic enhancement above background, or r_s (defined as $r_s = k_sA_s/k_0A_0$ for each sequence *s*).

The r_s values obtained from the *k*-Seq experiments revealed that all tested families contained sequences which displayed some activity on a new substrate or on multiple new substrates (Supplementary Fig. 5). Details of the frequency distribution of catalytic enhancement depended on both the aminoacyl side chain of the substrate as well as the ribozyme family. The distribution of sequences in Families 1A.1, 1B.1, and 3.1 could be characterized as containing a peak centered around background activity accompanied by a long, high-activity tail, particularly with BWO and BFO (Supplementary Fig. 5). In contrast, the distributions of Families 2.1 and 2.2 displayed distinct peaks at

higher activity, with bimodality apparent in some cases (especially for Family 2.1). This indicated a higher tolerance for mutations in Families 2.1 and 2.2 than in 1A.1, 1B.1, and 3.1, as mutant sequences were less likely to exhibit substantial detrimental effects.

Ribozyme families distinguish different chemical features of substrate side chains. To assess the activity and specificity of individual ribozyme mutants for each substrate, catalytic enhancement values for different substrates were compared in a pairwise fashion (Fig. 2 and Supplementary Fig. 6). All families displayed a high degree of correlation among activities for non-aromatic amino acid analogs (BLO (Leu), BIO (Ile), BVO (Val), and BMO (Met)) and also between activities for the two aromatic analogs (BWO (Trp) and BFO (Phe)) (Fig. 3A). The high correlations indicated that few sequences exhibit large activity differences between amino acids within the same biochemical class.

However, when comparing amino acids of different classes (i.e., aromatic vs. non-aromatic), strong correlations were only

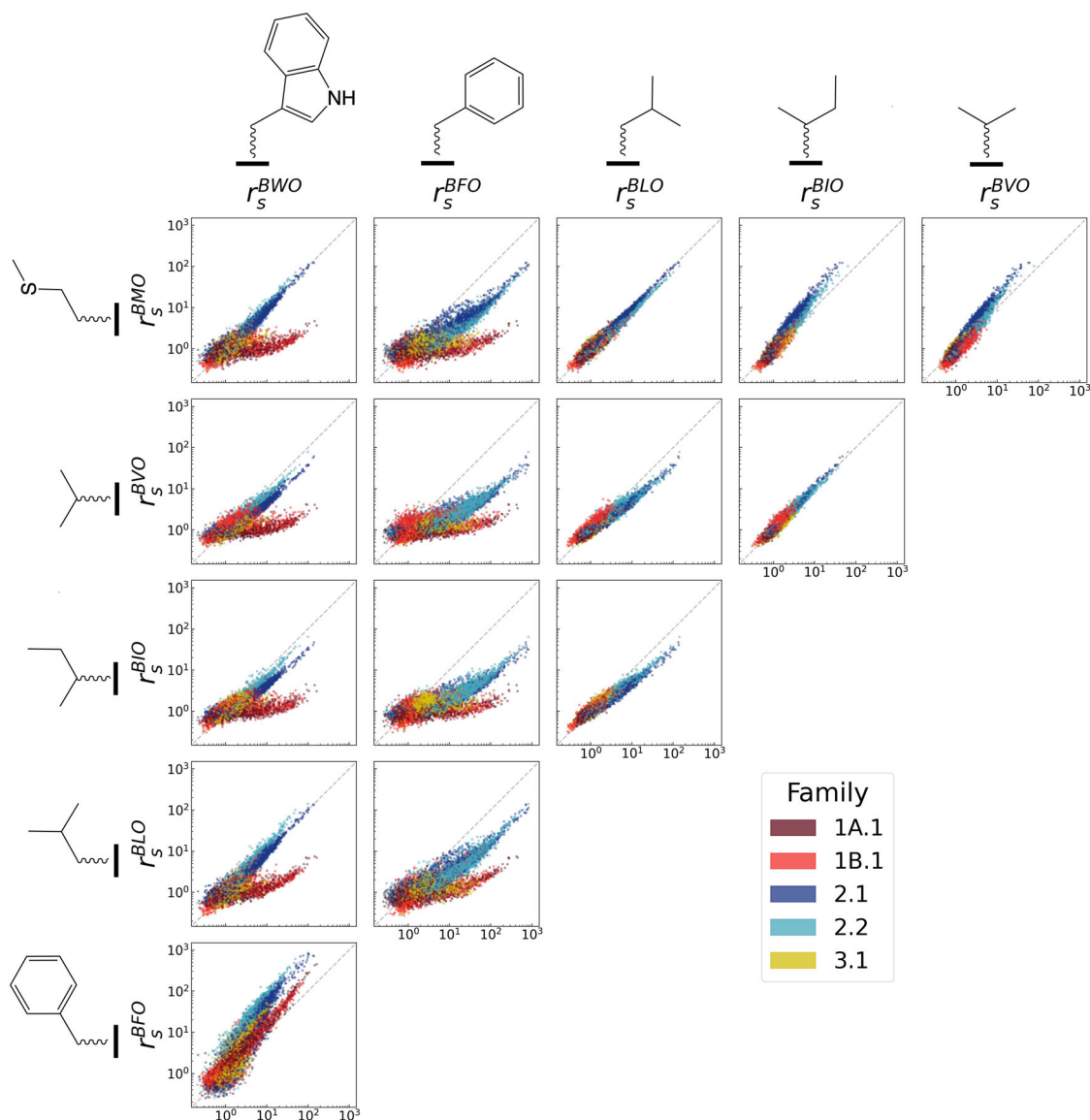


Fig. 2 Pairwise comparisons of activity on different substrates. Pairwise comparisons of catalytic enhancement (r_s) for individual sequences with each BXO substrate. Dashed gray line indicates the identity line. Substrates are ordered by hydrophilicity⁸⁹. See Supplementary Fig. 6 for error bars and mutant order for each family. Source data are provided as a Source Data file.

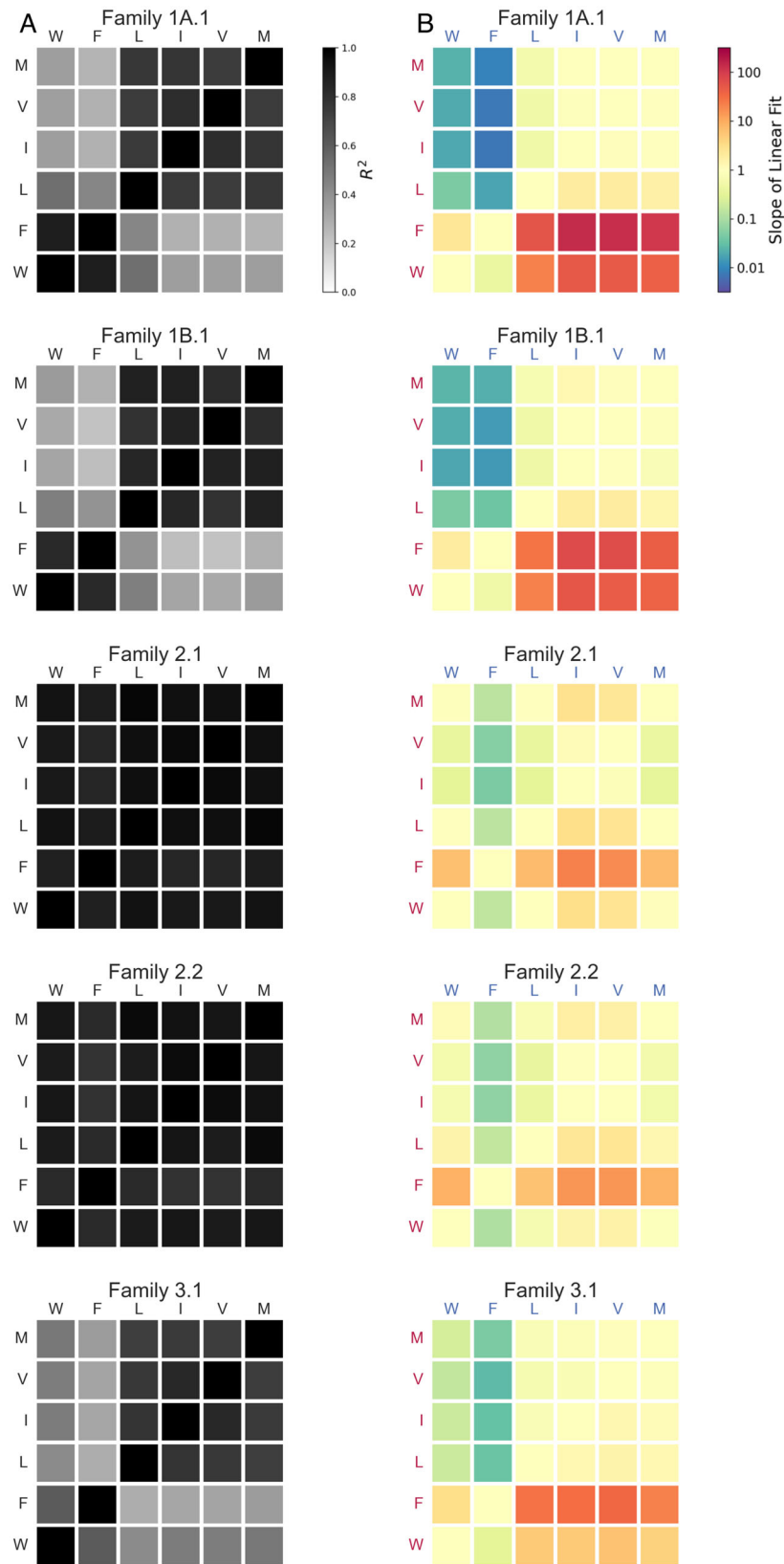


Fig. 3 Substrate preferences and correlations of activity. **A** Heat maps of coefficient of determination (R^2) for pairwise comparisons in Fig. 2. **B** Heat maps for slopes of linear regression fits for pairwise comparisons in Fig. 2. Slope > 1 indicates a preference for the substrate on the y-axis; slope < 1 indicates a preference for the substrate on the x-axis. Source data are provided as a Source Data file.

observed for Families 2.1 and 2.2, indicating that the effects of mutations in Motif 2 sequences tend to be relatively independent of the side chain. In contrast, Families 1A.1, 1B.1, and 3.1 showed substantially lower activity with non-aromatic side chains (Fig. 2), resulting in lower correlations between activity on aromatic and non-aromatic side chains (Fig. 3A). These preferences were also captured by the slopes on the correlation plots (Fig. 3B), which confirm that Motif 1 sequences strongly favor aromatic side chains, while Motif 2 sequences demonstrate less pronounced preferences, and Motif 3 sequences display an intermediate strength of preference. While less pronounced than for Motif 1, some preferences were still observed for Motif 2 sequences, in which BFO was most preferred, BMO, BWO and BLO were weakly preferred, and BVO and BIO were disfavored. Interestingly, BVO and BIO, in contrast to the other side chains, are both branched at the β carbon position. For Family 3.1, BFO was preferred over BWO, and all non-aromatic substrates were similarly disfavored. The differences observed between trends characterizing the separate ribozyme motifs suggest differences in the recognition mechanisms among Motifs 1, 2, and 3. Nevertheless, all ribozyme families display some preferences that correspond to chemical features of the side chains.

Substrate specificity is positively correlated with activity. To probe the relationship between catalytic activity and substrate specificity, we used two measures of specificity. First, as a general measure of substrate specificity for each sequence, we adapted the ‘promiscuity index’⁶¹. Here, promiscuity refers to the ability of a sequence to react with multiple substrates at a similar level of activity. The promiscuity index ($I_s = -\frac{1}{\log N} \sum_{i=1}^N \frac{r_i}{\sum_{j=1}^N r_j} \log \frac{r_i}{\sum_{j=1}^N r_j}$) is a normalized entropy which describes the evenness of the distribution of rates across different substrates. The promiscuity index I_s ranges from 0 to 1, such that sequences that are completely promiscuous, having equal activity on all substrates, would have $I_s = 1$, and sequences completely specific to one substrate would have $I_s = 0$. Promiscuity was observed to decrease as overall activity increased for all families (Fig. 4 and Supplementary Fig. 7).

Second, since ribozymes in some families displayed preferential activity with aromatic amino acids compared to non-aromatic amino acids, we calculated the relative preference for aromatic substrates as $(r_s^{BWO} + r_s^{BFO}) / \sum_{X'} r_s^{BXO}$. This ‘aromatic preference’ ratio reflects the proportion of ribozyme products that would have aromatic side chains in a reaction containing all six substrates at equal, sub-saturating concentration (Supplementary Fig. 8). Both the aromatic preference and the promiscuity index showed that the total activity of a sequence was positively correlated with specificity (positively correlated with aromatic preference and negatively correlated with promiscuity index; Table 1).

Abundance of opportunities for co-option for alternative substrates. Sequences that can function with multiple substrates could potentially be co-opted to adopt new functions (i.e., react with new substrates). To quantify the frequency of sequences able to react with multiple substrates, we categorized sequences as active or inactive using a catalytic enhancement threshold r_t . Sequences below this threshold are considered to be nearly inactive, being close to the background rate (see above). An activity threshold of $r_t = 5$ was chosen for two reasons. First, this threshold is two-fold more than the estimated 95% range for background activity (Supplementary Fig. 4, Supplementary Table 2), so values of $r_s > 5$ are statistically significantly greater than the normalized background rate. Second, increasing the rate

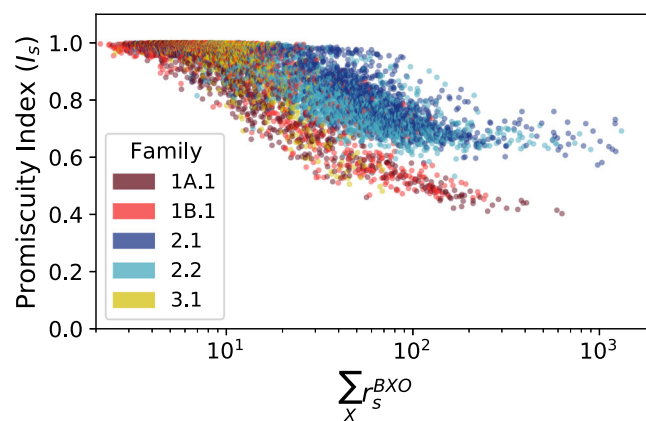


Fig. 4 Relationship between activity and promiscuity. Promiscuity index values for each sequence as a function of total activity (sum of activities with all tested substrates). The general trend indicates that specificity increases (promiscuity decreases) as overall activity increases. Source data are provided as a Source Data file.

Table 1 Correlations between overall catalytic activity and specificity for each ribozyme family (Pearson's R and Spearman's ρ ; $n = 1954$, p -values $< 10^{-95}$ in each case (two-sided)).

Family	Promiscuity Index		Aromatic Preference	
	R	ρ	R	ρ
1A.1	-0.696	-0.647	0.554	0.711
1B.1	-0.839	-0.502	0.738	0.477
2.1	-0.535	-0.888	0.452	0.911
2.2	-0.538	-0.866	0.445	0.865
3.1	-0.814	-0.462	0.749	0.513

of reaction by a factor of 5 is potentially significant in a prebiotic context, as abundances are expected to depend exponentially on relative fitness. Using this threshold, ribozyme mutants that were active on more than one substrate were considered capable of co-option (i.e., potentially able to adopt a new substrate).

Consistent with the observation that sequences in Families 2.1 and 2.2 displayed a high level of correlation of activities among all tested substrates, these families also had the most sequences being active with at least two substrates (1029 sequences in Family 2.1; 853 sequences in Family 2.2), and many were active with all six tested substrates (Fig. 5). Such sequences would be capable of co-option for new substrates. In contrast, Families 1A.1, 1B.1, and 3.1, which contain more inactive sequences and generally preferred aromatic amino acids, had most sequences accepting only one (or zero) substrates. Of sequences accepting multiple substrates in Families 1A.1, 1B.1, and 3.1, most were only active with two substrates. Nevertheless, even in these families, >2% of sequences accepted 2 or more substrates (254 sequences in Family 1A.1, 278 sequences in Family 1B.1, and 43 sequences in Family 3.1).

Increase of co-opted activity on the fitness landscape. The sequences identified as presenting opportunities for co-option are active on two (or more) substrates, but may not be optimally active on either. To determine how readily co-option might lead to a sequence with increased activity (i.e., to the optimally active sequence on a given substrate within the sequence space explored here) through evolution over the fitness landscape,

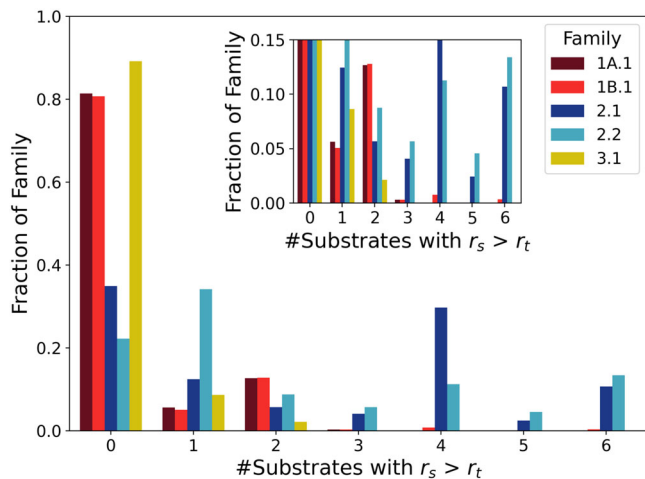


Fig. 5 Activity on multiple substrates and co-option potential. The frequency distribution of the fraction of unique sequences in each family (y-axis) that is active on a given number of substrates (x-axis). Activity on 2 or more substrates indicates potential for co-option. While Motif 2 sequences (Families 2.1 and 2.2) show a higher abundance of sequences active on more substrates, all families possess some sequences with activity on multiple substrates. Inset shows an enlargement of the low y-value region of the plot. Source data are provided as a Source Data file.

we investigated the connectivity of optimal sequences (i.e., fitness peaks) for each substrate within the fitness landscape defined by each substrate, for each ribozyme family. With the exception of Family 3.1, the substrate peaks (highest r_s) for each family were accessible to one another by evolutionary pathways proceeding through single mutations, while maintaining some activity (i.e., maintaining $\sum_X r_s^{BXO} > 30$, in analogy to $r_t = 5$ for 6 substrates) (Fig. 6). Family 3.1 was unique among families, in that the few co-optable sequences active on non-aromatic substrates were isolated in sequence space from the larger number of aromatic-prefering ribozyme mutants. While aromatic substrates were generally preferred, substantial increases in the activity on non-aromatic substrates could be obtained through 1-2 mutations. This analysis indicates that the number of mutations from wild-type required to improve activity on a new substrate can be relatively small.

Discussion

A system of self-aminoacylating ribozymes is an ideal platform for studying co-option in ribozyme evolution, as aminoacylations by the 20 biogenic amino acids represent naturally distinct functions in the context of a genetic code. Here we determined the activities of multiple self-aminoacylating ribozyme families with several activated amino acid substrates. While several examples of ribozymes accepting multiple small molecule substrates have been previously described^{10,22,62}, the high-throughput analysis described here allows quantification of trends in substrate preference, promiscuity and activity. These ribozymes were originally discovered by exhaustive in vitro selection over sequence space (21 nt random region flanked by constant regions)²⁴. Each tested family contained dozens or hundreds of sequences that could utilize multiple substrates, often with high correlations in activity between substrates. In addition, the optimally active sequences with each substrate were closely connected in sequence space in four of the five families, demonstrating high evolvability and optimization potential between functions. This highlights the potential for ribozymes with activity for a selected substrate to adopt other amino acid substrates. In an RNA World scenario, this process could be

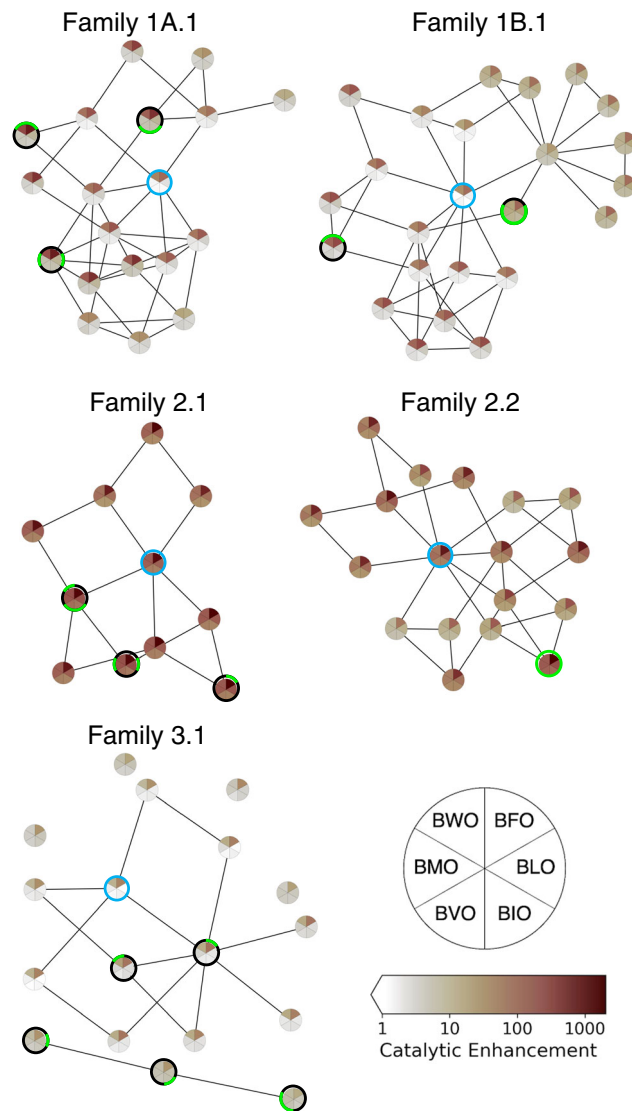


Fig. 6 Evolutionary pathways for increasing activity on different substrates. Each circular 'pie' represents a single sequence, whose catalytic enhancement for each substrate is shown by sector shading according to the heat map legend. For each family, the wild-type and the ribozymes having the six highest catalytic enhancements for each substrate are included. The wild-type sequence in each family is highlighted by a blue circle; the most active sequence for each substrate is indicated by a green sector outline for the substrate. Among the set of high-activity sequences, every pair of sequences for which Hamming distance $d = 2$ was examined to identify intervening sequences ($d = 1$ to both sequences of the pair) having substantial overall activity ($\sum_X r_s^{BXO} > 30$). The intervening sequences are also shown in the plot. Lines connect sequences where $d = 1$. Sequences and catalytic enhancement values are given in Supplementary Table 3. Source data are provided as a Source Data file.

beneficial for expanding metabolic chemical space and incorporating new compounds into increasingly complex systems.

While all families displayed substantial potential for adopting new substrates through co-option, ribozyme families differed in substrate preference and overall activity and extent of co-option potential. Namely, Families 1A.1, 1B.1, and 3.1 contained relatively few active ribozymes, and these tended to display strong preference for aromatic amino acid side chains, although some sequences in these families were more promiscuous. The families in Motif 1 followed the general preference order of

F,W > M,L,I,V, and the Motif 3 family followed the general preference order of F > W > M,L,I,V. Thus, these ribozymes appear to distinguish aromatic and non-aromatic side chains. On the other hand, Families 2.1 and 2.2 contained many sequences with high activity on all tested substrates, and also tended to prefer BFO. The families in Motif 2 followed the general preference order of F > M,W,L > I,V. This preference order suggests that Motif 2 ribozymes prefer the aromatic side chains, and are also subject to steric constraints, as they prefer F over W and also prefer L (non-branched β -carbon) over I and V (branched β -carbon). Given that these ribozymes were not selected for specificity (i.e., no counter-selections or negative selections), these preferences reflect inherent chemical and structural features of the RNA interactions with different side chains.

The evolution of error minimization in the standard genetic code has been a subject of extensive theoretical and analytical study stemming from the realization that the code is unusually conservative in light of mutations. Since error minimization has adaptive value, a prevalent and intuitive view is that this property arose through natural selection^{30,31,63}. However, an alternative view is that this trait emerged as a by-product during the initial expansion of the genetic code^{36,37,39}. For example, it has been suggested that duplication of aminoacyl-tRNA synthetases would lead to emergence of a conservative pairing, as the tRNA and amino acid substrates would be similar to the ancestral versions⁶⁴. Since the catalytic elements of the earliest protein translation machinery were presumably composed of RNA, and indeed, phylogenetic evidence suggests that the genetic code predates aminoacyl-tRNA synthetases, a similar logic suggests that code expansion in the RNA World would have a tendency to conserve biophysical features of the substrate^{38,39}. However, this expectation has not been previously tested experimentally.

Using our system of self-aminoacylating ribozymes, we found that all ribozymes showed preferences for certain biophysical features, being particularly sensitive to aromaticity and branching in the side chain. Thus, co-option of these ribozymes for adopting alternative substrates would produce an association between these biophysical features and the RNA sequence, possibly including the primitive anticodon region. A previous computational analysis of hypothetical alternative genetic codes showed little association between error-minimizing properties and the possible over-representation of nucleotide triplet codons or anticodons in binding sites of amino acid aptamers⁶⁵, concluding that error minimization would arise independently of a stereochemical origin of the genetic code. An important difference is that our present study does not test a mechanism for how the very first codon assignments were made. Instead, we address code expansion. Our results suggest that introduction of a new amino acid into the code could occur through co-option and optimization of a ribozyme already tasked with reaction with a chemically similar amino acid. If so, error minimization could arise as a by-product of code expansion, as new amino acids were adopted into the code. It is not necessary for the ribozyme's substrate recognition site to overlap its decoding site (e.g., an anticodon) for this process to occur; instead it is only necessary that the ribozymes are related by descent, which itself would result in a correlation between anticodon sequence and substrate recognition. No relationship was observed between codon or anticodon sequences and amino acid preferences for the ribozymes used in this study (Supplementary Fig. 14), although the number of sequences is small. Prior studies have reported examples of related RNA sequences that recognize similar substrates^{66,67}. In the present work, this principle was shown to apply to the case of aminoacylation ribozymes, setting up error minimization for the genetic code, and the large-scale analysis of many sequences allowed quantitation of substrate preferences, promiscuity, and correlations. Systematic, quantitative analysis is useful since

understanding the origin of life requires understanding whether certain properties of life are general vs. exceptional.

While other aminoacylation ribozymes are being developed to create alternative protein translation systems⁶⁸, it should be noted that the reactions studied here deviate from a precursor genetic code in at least three respects. First, the presence of biotin, while experimentally convenient, is unlikely to be prebiotically plausible, and some interaction between the ribozymes and substrate may be attributable to the biotin group; however, the observation that the aminoacyl side chain influences reactivity suggests that the ribozyme does interact with the side chain. Second, the product of reaction lacks a free amine for additional condensation. Third, the ribozymes are modified *in cis* at an internal site when reacting with BYO⁶⁹, which differs from the charging *in trans* of tRNA at the 3' end. Nevertheless, while the self-aminoacylating ribozymes studied here are a model system, and do not directly mimic a precursor genetic code, these results demonstrate the general principle that ribozyme co-option to incorporate new substrates could lead to tolerance of errors, as a by-product of system expansion.

Substrate preferences were amplified with increasing activity, resulting in a positive correlation between activity and substrate specificity. Previous research on the relationship between activity and specificity has noted intuitively appealing trade-offs between these two properties in some systems^{70–76}, as may be caused by ground-state discrimination in enzymes. In contrast, the results seen here indicate a positive correlation between catalytic activity and substrate specificity, instead reminiscent of enzymes that employ transition-state discrimination^{75,77}. The correlation observed would depend on the particular system under study and the relevant binding or stabilization mechanisms⁷⁸. Regardless, for this case, the evolutionary consequence of the positive activity-specificity correlation would be that natural selection for greater activity would also lead to greater substrate specificity, as a by-product. At the same time, given the prevalence of promiscuous sequences and the short evolutionary pathways among optimal sequences for different substrates, new substrate specificities would still be accessible even from highly active, specialized sequences. Such properties of overlapping fitness landscapes could facilitate the expansion from a weakly active, promiscuous ribozyme to an elaborated system of ribozyme-substrate pairs.

While the order in which amino acids were incorporated into the genetic code is a subject of debate, the amino acid substrates tested here include those that are generally believed to be early (L, I, V) and late (W, F, M) additions to the code^{54–58}. The aromatic residues were generally preferred by all ribozyme families. Such a preference is not surprising based on considerations for intermolecular interactions (e.g., π - π stacking) and is supported by an analysis of amino acid preferences among RNA aptamers evolved *in vitro*⁷⁹. Thus, in a plausible scenario, self-aminoacylating RNAs that react with 'early' amino acid substrates would have promiscuous activity on 'late' substrates, allowing co-option of these ribozymes to incorporate new substrates once they become available. During code expansion, any natural selection for increased activity would also lead to increased substrate specificity, and error minimization would emerge due to the biophysical and structural preferences of the ribozymes. These evolutionary by-products, in turn, would further improve the ability of a primitive genetic code to faithfully convert genetic information into peptide sequences with defined biophysical properties.

Emergent phenomena have been argued to be critical complements to natural selection in prebiotic evolution, including the origin of translation^{80,81} and replicase ribozymes⁸². Like the spandrels of St. Mark's Cathedral, architectural by-products that later acquired important esthetic value⁸³, error minimization and specificity may originate as mechanistic by-products of prebiotic evolution, to later become invaluable features of the complex system.

Methods

General synthesis methods. Reagents and solvents were obtained from Sigma-Aldrich or Fisher Scientific and were used without purification, unless otherwise noted. All ^1H NMR spectra were recorded using a Varian Unity Inova AS600 (600 MHz) with samples dissolved in DMSO-*d*₆; chemical shifts δH are reported in ppm with reference to residual internal DMSO ($\delta\text{H} = 2.50$ ppm). Spectra were analyzed using MNova 14 software.

Preparation of biotinyl-amino acids. Biotinylation reactions were performed in 10 mL anhydrous pyridine under nitrogen. Typical reactions contained L-amino acid methyl ester hydrochloride (1 mmol), biotin (1 mmol), N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride (EDC, 2 mmol), and 4-(dimethylamino)pyridine (0.1 mmol). The mixture was allowed to react at room temperature with stirring overnight, after which the solvent was evaporated under reduced pressure. The residue was then dissolved in dichloromethane (DCM) and washed with equal volumes of distilled water, saturated sodium bisulfate solution (twice), and saturated sodium bicarbonate solution (twice). The solution was dried with sodium sulfate, filtered, and the solvent was evaporated with reduced pressure to yield a clear, yellow solid (^1H NMR chemical shifts reported in Supplementary Table 4).

The recovered compound was dissolved by sonication in iPrOH:H₂O (2:1 v/v) (15 mL), to which 1 mL of 3 M NaOH was added. This solution was stirred overnight at room temperature, after which the isopropyl alcohol was evaporated under reduced pressure and the product was precipitated from the remaining solution by the addition of 1 M HCl to produce a white solid. This compound was recovered by filtration, washed with water, and dried *in vacuo* (Supplementary Table 4).

Preparation of biotinyl-aminoacyl oxazolones. Oxazolone formation was performed by reacting biotinyl-amino acids (0.1 mmol) with EDC (0.12 mmol) in anhydrous DCM and stirred at 4 °C overnight. The organic phase was then washed with distilled water (twice), saturated sodium bicarbonate solution, and saturated sodium chloride solution and dried with sodium sulfate. The solution was then filtered and the solvent was evaporated under reduced pressure to yield a solid product, which was stored at -20 °C (Supplementary Table 4 and Supplementary Fig. 9). NMR characterization was performed as described above. Mass spectra were obtained to verify compound synthesis (Supplementary Table 5). DART-MS spectra were collected on a Thermo Exactive Plus MSD (Thermo Scientific) equipped with an ID-CUBE ion source and a Vapur Interface (IonSense). Both the source and MSD were controlled by Excalibur v. 3.0. The analyte was spotted onto OpenSpot sampling cards (IonSense) using acetonitrile as the solvent. Ionization was accomplished using He plasma with no additional ionization agents. Mass calibration was carried out using Pierce LTQ Velos ESI (+) and (-) Ion calibration solutions (Thermo Fisher Scientific). The mass spectra are reported in Supplementary Fig. 16.

Substrate solutions were prepared by weighing biotinyl-aminoacyl-oxazolone (BXO, where X = W (Trp), F (Phe), L (Leu), I (Ile), V (Val), or M (Met)) and dissolving in acetonitrile with sonication to a final concentration of 25 mM. Fresh solutions were prepared daily for each set of experiments. As a secondary means of verifying BXO concentrations in prepared solutions, a HABA biotin quantification kit (AnaSpec) was used to measure the biotin concentrations of each solution. Average measured biotin concentration and standard deviation of triplicates are shown in Supplementary Table 6 (expected BXO concentration for all samples is 25 mM). While biotin quantitation measurements indicate systematically lower BXO concentrations than by weight by a factor of ~2, BXO concentrations were similar across different compounds. The low-activity background peaks also provide internal normalization to account for differences between compounds (see Results).

Kinetic sequencing (k-Seq). DNA libraries for kinetic sequencing experiments were designed based on prior work⁴⁹. Libraries were obtained from Integrated DNA Technologies (IDT) or Keck Biotechnology Laboratory with the sequence 5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-[central variable region, length 21]-TTCAGTGCAGACTTGACGAAGCTG-3' (nucleotides upstream of the transcription start site are underlined). The variable region was designed to contain one of the five wild-type sequences of interest (Supplementary Table 1) with variability at each position corresponding to 91% wild-type base and 3% each substitution. RNA was transcribed using HiScribe T7 RNA polymerase (New England Biolabs) and purified by denaturing polyacrylamide gel electrophoresis (PAGE). Reaction pools were prepared as an equimolar mixture of each purified RNA pool and quantified by Qubit 3 Fluorometer (Invitrogen).

Kinetic sequencing experiments were performed^{24,49}. Reactions were performed in 50 μL aqueous solutions containing selection buffer (100 mM HEPES, 100 mM NaCl, 100 mM KCl, 5 mM MgCl₂, 5 mM CaCl₂) and 5% acetonitrile at a pH between 6.9 and 7.0. Reactions contained 0.43 μM RNA and BXO at 1250, 250, 50, 10, or 2 μM . Acetonitrile carryover from preparation of BXO substrates was not observed to have an effect on ribozyme activity at this concentration (Supplementary Fig. 13). Reactions were incubated at room temperature with rotation for 90 minutes and stopped by desalting using Micro Bio-Spin Columns with Bio-Gel P-30 (Bio-Rad Laboratories). Reacted sequences were isolated with 100 μL

Streptavidin MagneSphere paramagnetic beads (Promega) per sample. Beads were washed three times with PBS + 0.01% Triton X-100 and sequences were eluted into 50 μL water by heating to 70 °C for 1 minute. Samples were reverse transcribed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). Following reverse transcription of k-Seq samples, qPCR reactions were performed in triplicate for each sample, including input RNA, using SsoAdvanced Universal SYBR Green Supermix (Bio-Rad Laboratories) with 2 μL of cDNA following the manufacturer's protocol and containing 500 nM forward and reverse primers 5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGA-3' and 5'-CAGCTTCGTCAA GTCTGCAGTGAA-3'. Serial dilutions of random library ssDNA were prepared in triplicate from 5 \times 10⁻⁵ to 5 \times 10² pg/ μL alongside each experiment for generating standard curves (Supplementary Fig. 10)⁸⁴. Samples were analyzed using Bio-Rad CFX96 Touch system. The remaining cDNA was amplified by PCR with Phusion DNA Polymerase (Thermo Fisher Scientific) using the same forward and reverse primers as used for qPCR above. Samples were adapted for sequencing using the Nextera XT DNA Library Preparation Kit (Illumina), pooled, and sequenced by Illumina NovaSeqS4 PE150 (Novogene).

Aminoacylation ribozyme selections. Selections for self-aminoacylating ribozymes with BFO and BLO were conducted in analogy to BYO aminoacylation²⁴. Libraries were obtained from IDT with the sequence 5'-GATAATACGACTCAC TATAGGGAATGGATCCACATCTACGAATTC-N₂₁-TTCAGTGCAGACTTGA CGAAGCTG-3' (T7 promoter sequence underlined), where N is an equimolar mixture of A, G, C, and T. For the first round of selection, 145 pmol of library DNA was transcribed using HiScribe T7 polymerase (New England Biolabs) and RNA was purified by gel electrophoresis. For the first round of selection, reactions contained 3.2 μM RNA and 50 μM BFO or BLO in 1 mL of selection buffer with 0.2% acetonitrile. Reactions were incubated at room temperature with rotation for 90 minutes and stopped by desalting using Micro Bio-Spin Columns with Bio-Gel P-30 (Bio-Rad Laboratories). Reacted sequences were isolated by addition of one sample volume of Streptavidin MagneSphere paramagnetic beads (Promega) per sample. Beads were washed bead buffer (PBS + 0.01% Triton X-100), 20 mM NaOH, and once more with bead buffer, then eluted by heating to 65 °C for 10 minutes in 95% formamide with 10 mM EDTA. Samples were reverse transcribed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) and amplified with Phusion DNA Polymerase (Thermo Fisher Scientific). For subsequent rounds of selection, 7.2 pmol (round 2) or 3.6 pmol (rounds 3-5) of recovered DNA was transcribed and RNA was used at 2.2 μM in 200 μL reactions. Selections were performed for five rounds in duplicate. Samples were prepared for sequencing using the Nextera XT DNA Library Preparation Kit (Illumina), pooled, and sequenced by Illumina NextSeq 500 (Biological Nanostructures Laboratory, California NanoSystems Institute at UCSB).

Electrophoretic mobility shift assay and determination of BFO uncatalyzed reaction rate. Gel shift assays were performed²⁴. Gel shift assays for observation of reactivity were performed with 500 μM BXO per sample unless otherwise noted. Aminoacylated RNA was incubated with 95 nmol streptavidin and run on an 8% native polyacrylamide gel with 0.5X TBE. For determining the uncatalyzed reaction rate with BFO, aminoacylation reactions were performed in 50 μL selection buffer with 5% acetonitrile containing BFO at 1250, 250, 50, 10, or 2 μM and 0.43 μM random library RNA which was fluorescently labeled using 5' EndTag Nucleic Acid Labeling System (Vector Laboratories) and fluorescein maleimide (TCI Chemicals). Reactions were incubated at room temperature for 90 minutes with rotation and stopped by desalting using Micro Bio-Spin Columns with Bio-Gel P-30 (Bio-Rad Laboratories). 95 nmol of streptavidin (New England Biolabs) was added to each sample, which were then incubated for 15 minutes with rotation at room temperature, run on an 8% polyacrylamide gel, imaged on an Amersham Typhoon 5 Biomolecular Imager, and analyzed using ImageQuant 8.1 software. For uncatalyzed reaction rate determination, all high molecular weight bands were grouped and compared to total RNA quantified in the lane to calculate the fraction reacted at each concentration, which was fit to the kinetic model.

Acid gel aminoacylation assay. 500 ng of RNA were reacted with BXO as described above. Samples were then analyzed on acid PAGE (8% polyacrylamide, acid buffer: 100 mM NaOAc pH 5.2, 7.5 M urea) at 4 °C and 10 W. Gels were stained with SYBR[®] Gold (Thermo Fisher Scientific), and then scanned using an Amersham Typhoon 5 Biomolecular Imager.

Computational analyses of k-Seq data. Sequencing reads were processed using trimmomatic SE CROP:90 to facilitate joining⁸⁵, and then paired-end reads were joined and unique sequences were enumerated using EasyDIVER⁸⁶. Joining was performed using the following PANDaseq⁸⁷ flags: -a -1 1 -A pear -C completely_miss_the_point:0. These flags strip primers after assembly rather than before (-a), require sequences to have a minimum length of 1 after removing primers (-1 1), set the assembly algorithm to PEAR⁸⁸ (-A pear), and exclude sequences with mismatches in overlapping paired-end regions (completely_miss_the_point:0). Primer sequences were extracted using CTACGAATTC as the forward primer and CTGCAGTGAA as the reverse primer.

k-Seq analyses were performed using the 'k-seq' package⁴⁹. Briefly, the absolute quantity (ng) of a sequence in a sample was calculated as the fraction of the sequence's read count over the total number of reads in the sample, multiplied by the mean total RNA (ng) from triplicated qPCR measurements. The input amount (ng) for a sequence was determined by the median sequence amount across 6 replicates for the unreacted pool. The fraction reacted (F_s) was calculated as the reacted amount in the sample divided by the input amount. Sequences that contain ambiguous nucleotides ('N'), that were not 21 nucleotides long, or that were more than two substitutions from a center sequence were excluded in downstream fitting. For each sequence, the fractions reacted in samples were fit to the pseudo-first order kinetic model $F_s^{\text{BXO}} = A_s(1 - e^{-k_s \alpha [\text{BXO}]t})$, where F_s^{BXO} is the fraction reacted for sequence *s* with substrate BXO, A_s is the maximum reaction amplitude, k_s is the rate constant, and $[\text{BXO}]$ is the initial concentration of BXO. α is the coefficient accounting for the hydrolysis of substrate BXO during the reaction time ($t = 90$ min), and a fixed value (0.479, measured for BYO²⁴) was used for all substrates. Note that the effect of α on estimated k_s cancels out when calculating the catalytic enhancement ratio r_s . To quantify the estimation uncertainty of kinetic model parameters (k_s , A_s) for each sequence, samples (fractions reacted) were bootstrapped (resampling with replacement to the original size) for 1000 times and each bootstrapped sample set was fit into the model for k_s and A_s . Statistics (e.g., median, standard deviation, 2.5-percentile, 97.5-percentile) were calculated from bootstrapped results. The median of product $k_s A_s$ was used to represent the activity of each sequence.

Background reaction rate estimation. Histograms (100 bins) of \log_{10} -transformed kA values for sequences from all families were fit to a bimodal Gaussian distribution (Supplementary Fig. 4 and Supplementary Table 2). The mean of the low-activity peak (μ_l) was used as the estimated uncatalyzed rate ($k_0 A_0$) and the standard deviation of the fit (σ_l) was used to inform the choice of catalytic enhancement threshold. Additionally, the uncatalyzed reaction rate was calculated for BFO by gel shift assay as described previously for BYO²⁴ (see above).

Clustering analysis of sequences from selections. Sequences were clustered into families based on sequence similarity, using a custom Python script (see Data Availability). The script *ClusterBOSS.py* uses the enumerated read output files generated from the EasyDIVER package⁸⁶. In general, first, all sequences were sorted according to their read count values. Then, the most abundant sequence was chosen as a candidate 'center' sequence to start a family, as long as its read count value was at least 10 ($c_{\min} = 10$). The Levenshtein edit distance (number of substitutions, insertions, or deletions) from this candidate sequence to every other sequence in the distribution was computed (no restriction on minimum number of counts; $a_{\min} = 1$). If the distance was less than a cutoff ($d_{\text{cutoff}} = 3$ mutations from the center sequence), the sequence was considered to be part of the same family as the initially chosen center sequence. No restriction was applied to the number of sequences required to define a family ($n_{\min} = 1$), which includes the center sequence and any sequences found to cluster with it. Once assigned to a family, sequences were not allowed to be clustered into another family. To find the rest of the family clusters, we followed the same procedure until all sequences had been explored.

Promiscuity indices. Promiscuity indices were calculated using the calculator available at <http://hetaira.herokuapp.com/>. Due to the single-turnover nature of the aminoacylation ribozymes studied here, promiscuity indices are calculated using catalytic enhancement values instead of the catalytic efficiency as originally described by Nath and Atkins⁶¹.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data from high-throughput sequencing and *k*-Seq analysis (Figs. 2–6) have been deposited in the Dryad Digital Repository under DOI 10.25349/D92C9C (<https://doi.org/10.25349/D92C9C>). Source data are provided with this paper. The processed data are available in the Source Data and Supplementary Information file. Source data are provided with this paper.

Code availability

Scripts not reported elsewhere are available at <https://github.com/ichen-lab-ucsb/ClusterBOSS> (ClusterBOSS: Cluster Based On Sequence Similarity) and https://github.com/ichen-lab-ucsb/WFLIVM_k-Seq (scripts used to generate figures in this manuscript). Previously published tools are available at <https://github.com/ichen-lab-ucsb/EasyDIVER> and <https://github.com/ichen-lab-ucsb/k-seq>.

Received: 5 September 2021; Accepted: 16 June 2022;

Published online: 25 June 2022

References

- Pressman, A., Blanco, C. & Chen, I. A. The RNA world as a model system to study the origin of life. *Curr. Biol.* **25**, R953–R963 (2015).
- Joyce, G. F. & Szostak, J. W. Protocells and RNA self-replication. *Cold Spring Harb. Perspect. Biol.* **10**, <https://doi.org/10.1101/cshperspect.a034801> (2018).
- Gould, S. J. & Vrba, E. S. Exaptation - a missing term in the science of form. *Paleobiology* **8**, 4–15 (1982).
- Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
- Ycas, M. On earlier states of the biochemical system. *J. Theor. Biol.* **44**, 145–160 (1974).
- Aharoni, A. et al. The 'Evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).
- Espinosa-Cantu, A., Ascencio, D., Barona-Gomez, F. & DeLuna, A. Gene duplication and the evolution of moonlighting proteins. *Front. Genet.* **6**, 227 (2015).
- Peracchi, A. The limits of enzyme specificity and the evolution of metabolism. *Trends Biochem. Sci.* **43**, 984–996 (2018).
- Voros, D., Konnyu, B. & Czarán, T. Catalytic promiscuity in the RNA World may have aided the evolution of prebiotic metabolism. *PLoS Comput Biol.* **17**, e1008634 (2021).
- Janzen, E., Blanco, C., Peng, H., Kenchel, J. & Chen, I. A. Promiscuous ribozymes and their proposed role in prebiotic evolution. *Chem. Rev.* **120**, 4879 (2020).
- Szathmáry, E. & Smith, J. M. The major evolutionary transitions. *Nature* **374**, 227–232 (1995).
- de Duve, C. Transfer RNAs: the second genetic code. *Nature* **333**, 117–118 (1988).
- Perona, J. J. & Hadd, A. Structural diversity and protein engineering of the aminoacyl-tRNA synthetases. *Biochemistry* **51**, 8705–8729 (2012).
- Tawfik, D. S. & Gruic-Sovulj, I. How evolution shapes enzyme selectivity - lessons from aminoacyl-tRNA synthetases and other amino acid utilizing enzymes. *FEBS J.* **287**, 1284–1305 (2020).
- Artymiuk, P. J., Rice, D. W., Poirrette, A. R. & Willet, P. A tale of two synthetases. *Nat. Struct. Biol.* **1**, 758–760 (1994).
- Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* **30**, 1427–1464 (2002).
- Aravind, L., Anantharaman, V. & Koonin, E. V. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETPP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. *Proteins* **48**, 1–14 (2002).
- Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12**, 392–399 (2002).
- Fournier, G. P., Andam, C. P., Alm, E. J. & Gogarten, J. P. Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. *Orig. Life Evol. Biosph.* **41**, 621–632 (2011).
- Fournier, G. P., Andam, C. P. & Gogarten, J. P. Ancient horizontal gene transfer and the last common ancestors. *BMC Evol. Biol.* **15**, 70 (2015).
- Illangasekare, M., Sanchez, G., Nickles, T. & Yarus, M. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* **267**, 643–647 (1995).
- Illangasekare, M. & Yarus, M. Specific, rapid synthesis of Phe-RNA by RNA. *Proc. Natl Acad. Sci. USA* **96**, 5470–5475 (1999).
- Li, N. & Huang, F. Ribozyme-catalyzed aminoacylation from CoA thioesters. *Biochemistry* **44**, 4582–4590 (2005).
- Pressman, A. D. et al. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J. Am. Chem. Soc.* **141**, 6213–6223 (2019).
- Saito, H., Kourouklis, D. & Suga, H. An in vitro evolved precursor tRNA with aminoacylation activity. *EMBO J.* **20**, 1797–1806 (2001).
- Murakami, H., Ohta, A., Ashigai, H. & Suga, H. A highly flexible tRNA acylation method for non-natural polypeptide synthesis. *Nat. Methods* **3**, 357–359 (2006).
- Woese, C. R. On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 1546–1552 (1965).
- Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
- Haig, D. & Hurst, L. D. A quantitative measure of error minimization in the genetic-code. *J. Mol. Evol.* **33**, 412–417 (1991).

30. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
31. Goodarzi, H., Nejad, H. A. & Torabi, N. On the optimality of the genetic code, with the consideration of termination codons. *Biosystems* **77**, 163–173 (2004).
32. Zhu, W. & Freeland, S. The standard genetic code enhances adaptive evolution of proteins. *J. Theor. Biol.* **239**, 63–70 (2006).
33. Firnberg, E. & Ostermeier, M. The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res.* **41**, 7420–7428 (2013).
34. Archetti, M. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J. Mol. Evol.* **59**, 258–266 (2004).
35. Novozhilov, A. S., Wolf, Y. I. & Koonin, E. V. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2**, 24 (2007).
36. Massey, S. E. The neutral emergence of error minimized genetic codes superior to the standard genetic code. *J. Theor. Biol.* **408**, 237–242 (2016).
37. Attie, O., Sulkow, B., Di, C. & Qiu, W. G. Genetic codes optimized as a traveling salesman problem. *PLoS ONE* **14**, e0224552 (2019).
38. Wolf, Y. I. & Koonin, E. V. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol. Direct* **2**, 14 (2007).
39. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the universal genetic code. *Annu. Rev. Genet.* **51**, 45–62 (2017).
40. Leman, L., Orgel, L. & Ghadiri, M. R. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science* **306**, 283–286 (2004).
41. Biron, J. P., Parkes, A. L., Pascal, R. & Sutherland, J. D. Expedient, potentially primordial, aminoacylation of nucleotides. *Angew. Chem. Int. Ed. Engl.* **44**, 6731–6734 (2005).
42. Danger, G., Boiteau, L., Cottet, H. & Pascal, R. The peptide formation mediated by cyanate revisited. N-carboxyanhydrides as accessible intermediates in the decomposition of N-carbamoylamino acids. *J. Am. Chem. Soc.* **128**, 7412–7413 (2006).
43. Danger, G., Plasson, R. & Pascal, R. Pathways for the formation and evolution of peptides in prebiotic environments. *Chem. Soc. Rev.* **41**, 5416–5429 (2012).
44. Danger, G. et al. 5(4H)-oxazolones as intermediates in the carbodiimide- and cyanamide-promoted peptide activations in aqueous solution. *Angew. Chem. Int. Ed. Engl.* **52**, 611–614 (2013).
45. Liu, Z., Beauflis, D., Rossi, J. C. & Pascal, R. Evolutionary importance of the intramolecular pathways of hydrolysis of phosphate ester mixed anhydrides with amino acids and peptides. *Sci. Rep.* **4**, 7440 (2014).
46. Liu, Z., Rigger, L., Rossi, J. C., Sutherland, J. D. & Pascal, R. Mixed anhydride intermediates in the reaction of 5(4H)-oxazolones with phosphate esters and nucleotides. *Chemistry* **22**, 14940–14949 (2016).
47. Liu, Z. W. et al. 5(4H)-Oxazolones as effective aminoacylation reagents for the 3'-terminus of RNA. *Synlett* **28**, 73–77 (2017).
48. Liu, Z. et al. Harnessing chemical energy for the activation and joining of prebiotic building blocks. *Nat. Chem.* **12**, 1023–1028 (2020).
49. Shen, Y., Pressman, A., Janzen, E. & Chen, I. Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkab199> (2021).
50. Yokobayashi, Y. High-throughput analysis and engineering of ribozymes and deoxyribozymes by sequencing. *Acc. Chem. Res.* **53**, 2903–2912 (2020).
51. Kobori, S. & Yokobayashi, Y. High-throughput mutational analysis of a twister ribozyme. *Angew. Chem. Int. Ed. Engl.* **55**, 10354–10357 (2016).
52. Kobori, S., Nomura, Y., Miu, A. & Yokobayashi, Y. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res.* **43**, e85 (2015).
53. Jalali-Yazdi, F., Lai, L. H., Takahashi, T. T. & Roberts, R. W. High-throughput measurement of binding kinetics by mRNA display and next-generation sequencing. *Angew. Chem. Int. Ed. Engl.* **55**, 4007–4010 (2016).
54. Trifonov, E. N. The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22**, 1–11 (2004).
55. Zaia, D. A., Zaia, C. T. & De Santana, H. Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.* **38**, 469–488 (2008).
56. Higgs, P. G. & Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).
57. Cleaves, H. J. 2nd The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498 (2010).
58. Longo, L. M. & Blaber, M. Protein design at the interface of the pre-biotic and biotic worlds. *Arch. Biochem. Biophys.* **526**, 16–21 (2012).
59. Walker, S. E. & Fredrick, K. Preparation and evaluation of acylated tRNAs. *Methods* **44**, 81–86 (2008).
60. Lai, Y. C., Liu, Z. & Chen, I. A. Encapsulation of ribozymes inside model protocells leads to faster evolutionary adaptation. *Proc. Natl. Acad. Sci. USA* **118**, <https://doi.org/10.1073/pnas.2025054118> (2021).
61. Nath, A. & Atkins, W. M. A quantitative index of substrate promiscuity. *Biochemistry* **47**, 157–166 (2008).
62. Stuhlmann, F. & Jäschke, A. Characterization of an RNA active site: interactions between a Diels-Alderase ribozyme and its substrates and products. *J. Am. Chem. Soc.* **124**, 3238–3244 (2002).
63. Archetti, M. Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* **59**, 400–415 (2004).
64. Pak, D., Kim, Y. & Burton, Z. F. Aminoacyl-tRNA synthetase evolution and sectoring of the genetic code. *Transcription* **9**, 205–224 (2018).
65. Yarus, M., Widmann, J. J. & Knight, R. RNA-amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.* **69**, 406–429 (2009).
66. Yang, Y., Kochoyan, M., Burgstaller, P., Westhof, E. & Famulok, M. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* **272**, 1343–1347 (1996).
67. Batey, R. T. Structure and mechanism of purine-binding riboswitches. *Q. Rev. Biophys.* **45**, 345–381 (2012).
68. Chen, J., Chen, M. & Zhu, T. Translating protein enzymes without aminoacyl-tRNA synthetases. *Chem*, 786–798. <https://doi.org/10.1016/j.chempr.2021.01.017> (2021).
69. Pressman, A. D. et al. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J. Am. Chem. Soc.* **141**, 6213–6223 (2019).
70. Mayr, H. & Ofial, A. R. The reactivity-selectivity principle: an imperishable myth in organic chemistry. *Angew. Chem. Int. Ed. Engl.* **45**, 1844–1854 (2006).
71. Khersonsky, O. & Tawfik, D. S. in *Comprehensive Natural Products II* (eds Hung-Wen Liu & Lew Mander) 47–88 (Elsevier, 2010).
72. Savir, Y., Noor, E., Milo, R. & Tlusty, T. Cross-species analysis traces adaptation of rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. USA* **107**, 3475–3480 (2010).
73. Larson, M. H. et al. Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc. Natl. Acad. Sci. USA* **109**, 6555–6560 (2012).
74. Johansson, M., Zhang, J. & Ehrenberg, M. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proc. Natl. Acad. Sci. USA* **109**, 131–136 (2012).
75. Tawfik, D. S. Accuracy-rate tradeoffs: how do enzymes meet demands of selectivity and catalytic efficiency? *Curr. Opin. Chem. Biol.* **21**, 73–80 (2014).
76. Flamholz, A. I. et al. Revisiting trade-offs between rubisco kinetic parameters. *Biochemistry* **58**, 3365–3376 (2019).
77. Beard, W. A., Shock, D. D., Vande Berg, B. J. & Wilson, S. H. Efficiency of correct nucleotide insertion governs DNA polymerase fidelity. *J. Biol. Chem.* **277**, 47393–47398 (2002).
78. Carothers, J. M., Oestreich, S. C. & Szostak, J. W. Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J. Am. Chem. Soc.* **128**, 7929–7937 (2006).
79. Blanco, C., Bayas, M., Yan, F. & Chen, I. A. Analysis of evolutionarily independent protein-RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Curr. Biol.* **28**, 526–537 (2018).
80. Lanier, K. A. & Williams, L. D. The origin of life: models and data. *J. Mol. Evol.* **84**, 85–92 (2017).
81. Lanier, K. A., Petrov, A. S. & Williams, L. D. The central symbiosis of molecular biology: molecules in mutualism. *J. Mol. Evol.* **85**, 8–13 (2017).
82. Attwater, J., Raguram, A., Morgunov, A. S., Gianni, E. & Holliger, P. Ribozyme-catalysed RNA synthesis using triplet building blocks. *Elife* **7**, <https://doi.org/10.7554/eLife.35255> (2018).
83. Gould, S. J. & Lewontin, R. C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B Biol. Sci.* **205**, 581–598 (1979).
84. Lai, Y.-C., Liu, Z. & Chen, I. A. Encapsulation of ribozymes inside model protocells leads to faster evolutionary adaptation. *Proc. Natl. Acad. Sci. USA* **118**, e2025054118 (2021).
85. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
86. Blanco, C., Verbanic, S., Seelig, B. & Chen, I. A. EasyDIVER: a pipeline for assembling and counting high-throughput sequencing data from in vitro evolution of nucleic acids or peptides. *J. Mol. Evol.* **88**, 477–481 (2020).
87. Masella, A. P., Bartram, A. K., Trzaskowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
88. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
89. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78**, 3824–3828 (1981).

Acknowledgements

The authors thank John Sutherland for support for organic synthesis and discussion of exaptation, Jen Smith for advice on sequencing, Huan Peng, Yei-Chen Lai, and Abe Pressman for technical advice, Robert Pascal for advice on organic synthesis, William Atkins and Abhinav Nath for advice on the promiscuity index, and Greg Khitrov for assistance with mass spectrometry. The authors acknowledge the use of the Biological Nanostructures Laboratory and research facilities within the California NanoSystems Institute, supported by the University of California, Santa Barbara and the University of California, Office of the President, and the UCLA Molecular Instrumentation Center. NMR was performed in the MRL Shared Experimental Facilities, which are supported by the MRSEC Program of the NSF under Award No. DMR 1720256; a member of the NSF-funded Materials Research Facilities Network (www.mrfn.org). Use was made of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara. Funding from the Simons Foundation Collaboration on the Origin of Life (290356FY18; IAC), NASA (NNX16AJ32G, 80NSSC21K0595; IAC), National Institute of General Medical Sciences (DP2GM123457; IAC), NSF (1935372; IAC) and the Camille Dreyfus Teacher-Scholar Program (IAC) is acknowledged.

Author contributions

E.J. and I.A.C. designed the project; E.J., A.V.-S. and J.K. conducted experiments and analyzed data; E.J., Z.L., and J.K. synthesized oxazolones; Y.S. analyzed *k*-Seq data; C.B. wrote the clustering script; E.J. and I.A.C. interpreted data; E.J. and I.A.C. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-31387-0>.

Correspondence and requests for materials should be addressed to Irene A. Chen.

Peer review information *Nature Communications* thanks the other anonymous Reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022