University of California
Santa Barbara

Evolutionary Emergence, Optimization, and Co-Option of Aminoacylation Ribozymes

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Biochemistry and Molecular Biology

by

Evan Janzen

Committee in charge:

Professor Irene A. Chen, Chair
Professor Luc Jaeger
Professor Kevin W. Plaxco
Professor Omar A. Saleh

June 2021

The Dissertation of Evan Janzen is approved.

Professor Luc Jaeger

Professor Kevin W. Plaxco

Professor Omar A. Saleh

Professor Irene A. Chen, Committee Chair

May 2021

Evolutionary Emergence, Optimization, and Co-Option of Aminoacylation Ribozymes

Copyright © 2021

by

Evan Janzen

Acknowledgements

The successful completion of my doctoral work was the result of the support and assistance of many, only a fraction of whom can be acknowledged in this space.

First and foremost, I must thank my advisor, Irene Chen. She provided me with countless opportunities to succeed, continuously challenged me and pushed me to learn and grow, and provided me with the support and freedom to achieve to my fullest potential.

I am also very thankful for my many fantastic labmates for creating such a fun and collaborative environment: Abe Pressman, Yuning Shen, Celia Blanco, Ranajay Saha, Huan Peng, Josh Kenchel, Alberto Vazquez-Salazar, Yei-Chen Lai, Sam Verbanic, Claire Tran, Greg Campbell, Jen Mobberley, Ray Borg, Damayanti Bagchi, Baoqing Zhou, and Steven Yang.

This work was the result of many successful collaborations, from which I learned a great deal, including valuable input from Robert Pascal, Jerry Joyce, Uli Müller, and John Sutherland. A special thank you also to Ziwei Liu for being an incredibly knowledgeable, helpful, and patient teacher.

I thank my committee of Luc Jaeger, Kevin Plaxco, and Omar Saleh for their guidance and support, as well as the support staff at UCSB for their assistance: Stella Hahn, Nicole Becker, Lauren Baker, Jen Smith, Shamon Walker, Cabe Fletcher, Adrian Shelor, Trevor Bellefeuille, and Rob Callaway.

Lastly, I'd like to thank my many friends and family, in particular my wife, Stephanie Khairallah, who have provided endless support and encouragement.

Curriculum Vitæ

Evan Janzen

Education

2021	**Ph.D. in Biochemistry and Molecular Biology**, University of California, Santa Barbara
2013	**M.S. in Molecular and Integrative Physiology**, University of Kansas
2010	**B.S. in Biochemistry and Molecular Biology, Minor in History**, Emporia State University

Employment History

2015 - 2021	**Graduate Research Assistant**, University of California, Santa Barbara
2013 - 2015	**Microbiologist**, Bayer HealthCare
2011 - 2013	**Graduate Research Assistant**, Stowers Institute for Medical Research
2008 - 2010	**Pharmacy Technician**, Walmart

Teaching Experience

2021	**Introductory Biology I Teaching Assistant**, University of California, Santa Barbara
2020	**Introductory Biology II - Ecology and Evolution Teaching Assistant**, University of California, Santa Barbara
2015, 2021	**Introductory Biology I Laboratory Teaching Assistant**, University of California, Santa Barbara
2009 - 2010	**Introductory Chemistry Lab Teaching Assistant**, Emporia State University

Publications

- Janzen E., Shen Y., Liu Z., Blanco C., Chen I.A. **Error minimization and specificity could emerge in a genetic code as by-products of prebiotic evolution.** *Submitted.*
- Shen Y., Pressman A.D., Janzen E., Chen I.A. **Kinetic sequencing (*k*-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters.** *Nucleic Acids Res.*, 2021.
- Janzen E., Blanco C., Peng H., Kenchel J., Chen I.A. **Promiscuous Ribozymes and Their Proposed Role in Prebiotic Evolution.** *Chem. Rev.*, 2020.

- Pressman A.D., Liu Z., Janzen E., Blanco C., Muller U.F., Joyce G.F., Pascal R., Chen I.A. **Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA.** *J. Am. Chem. Soc.* 141(15): 6213–6223, 2019.
- Blanco C., Janzen E., Pressman A., Saha R., Chen I.A. **Molecular Fitness Landscapes from High Coverage Sequence Profiling.** Annu. *Rev. Biophys.* 48, 2019
- Janzen E. **Regulation of telomerase reverse transcriptase expression in Schizosaccharomyces pombe.** *University of Kansas*, 2013.
- Haralalka S., Shelton C., Cartwright H.N., Katzfey E., Janzen E., Abmayr, SM. **Asymmetric Mbc, active Rac1 and F-actin foci in the fusion-competent myoblasts during myoblast fusion in Drosophila.** *Development* 138(8):1551-62, 2011.

Presentations

- Janzen E. **"Understanding Evolutionary Innovation."** *Chemical Science Student Seminar*. December 2019. Talk.
- Janzen E., Pressman A.D., Liu Z., Blanco C., Shen Y., Müller U.F., Joyce G.F., Sutherland J.D., Pascal R., Chen I.A. **"Mapping ribozyme fitness landscapes to optimize activity and specificity."** *Center for NanoScience Workshop on Evolving Nanosciences*. September 2019. Poster and flash talk.
- Janzen E., Pressman A.D., Liu Z., Blanco C., Shen Y., Müller U.F., Joyce G.F., Sutherland J.D., Pascal R., Chen I.A. **"Mapping ribozyme fitness landscapes to optimize activity and specificity."** *Synthetic Biology: Engineering, Evolution, and Design Conference*. June 2019. Poster and flash talk.
- Janzen E., Pressman A.D., Liu Z., Shen Y., Blanco C., Kenchel J., Müller U.F., Joyce G.F., Sutherland J.D., Pascal R., Chen I.A. **"Systematic mapping of fitness landscapes for self-aminoacylating ribozymes."** *Simons Collaboration on the Origins of Life Annual Symposium*. April 2019. Poster.
- Janzen E., Blanco C., Chen I.A., Braun D. **"Deep sequencing and analysis methods."** Simons Collaboration on the Origins of Life Annual Symposium. April 2018. Talk. Pressman A.D., Janzen E., Blanco C., Liu Z., Muller U.F., Joyce G.F., Pascal R., Chen I.A. "Comprehensive ribozyme activity landscape reveals a frustrated evolutionary network." *Simons Collaboration on the Origins of Life Annual Symposium*. April 2018. Poster.
- Janzen E. **"Mapping evolution in an RNA world."** *Friday Noon Seminar*, UCSB. May 2017. Talk.
- Janzen E., Helston R., and Baumann P. "**Regulation of telomerase expression in fission yeast."** *European Molecular Biology Organization Conference on Telomeres and the DNA Damage Response*. October 2012. Poster.

- Janzen E. **"Isolation of secondary metabolites from Myxobacteria."** *Emporia State University Research and Creativity Day.* April 2010. Talk.
- Janzen E., Lovich A., and Bailey M.M. **"Effects of long-chain hydrocarbon industrial solvents on murine fetal development."** *Kansas IDeA Network of Biomedical Research Excellence Annual Symposium.* January 2010. Poster.

Honors and Awards

2020	Graduate Division Dissertation Fellowship
2019	Graduate Student Association Travel Grant
2018	Best Presentation, Friday Noon Seminar
2017	Best Poster, UCSB BMSE/MCDB Retreat
2017	DeWolfe Teaching Fellowship in Organic Chemistry
2010	*Magna cum laude* distinction, Emporia State University
2009-2010	Kansas IDeA (Institutional Development Award) Network of Biomedical Research Excellence (K-INBRE) Scholarship
2009	Shepherd Honors Scholarship, Emporia State University
2006-2010	Presidential Honors Scholarships, Emporia State University
2008	Breukelman Biology Scholarship, Emporia State University

Organizations and Affiliations

2020-2021	American Chemical Society
2019-2020	English for Multilingual Students Oral Skills Workshop, UCSB
2016-2020	Chemical Sciences Student Seminar Series Committee, UCSB
2015-2020	Graduate Biology Mentorship Association, UCSB
2017-2019	Graduate Union of Molecular Biology Investigators, UCSB
2017-2019	Graduate Student Recruitment Committee, Biomolecular Science and Engineering Program, UCSB
2011-2013	Crossroads Student and Postdoc Association, Stowers Institute
2011-2013	University of Kansas Medical Center Physiological Society
2009-2010	American Chemical Society Student Affiliate Program
2008-2009	Educational Opportunity Fund Committee, ESU
2007-2010	Phi Kappa Phi National Honor Society

Abstract

Evolutionary Emergence, Optimization, and Co-Option of Aminoacylation Ribozymes

by

Evan Janzen

Understanding molecular evolution can reveal a great deal about the past, present, and future of biological systems. The evolution of catalytic RNA is of particular interest because of its potential role in an 'RNA World' at the origin of life. Two crucial aspects in the evolution of biomolecules are optimization on the fitness landscape and co-option for new functions. The fitness landscape describes a function of fitness in the space of all possible sequences. Molecules evolve through a random walk on the fitness landscape, with a bias toward climbing peaks. In addition, the ability of enzymes, including ribozymes, to catalyze side reactions is believed to be essential to the evolution of novel biochemical activities. It has been speculated that the earliest ribozymes were low in activity but high in promiscuity, which then gave rise to specialized descendants with higher activity and specificity. One particularly essential activity for the origin of life would be the reaction of ribozymes with activated amino acids to form aminoacyl-RNAs, with co-option of these aminoacyl-RNAs leading to genetic code expansion. In this work, self-aminoacylating ribozymes were identified through *in vitro* selection from full coverage of sequence space and characterized using a massively parallel kinetic assay. Three major sequence motifs were identified on the landscape and analysis of evolutionary pathways revealed that, while local optimization within a ribozyme family would be possible, optimization of activity over the entire landscape would be frustrated by large valleys of low activity. The sequence motifs associated with each peak represent different solutions for catalysis, so the inability to traverse the landscape globally corre-

sponds to an inability to restructure the ribozyme without losing activity. In addition, five families representing the three sequence motifs were further investigated to measure their activity with alternative substrates. Ribozymes in each family displayed high levels of co-optability, with activity on multiple substrates, demonstrating that co-option for a new function can occur more readily than optimization of an existing one. Related ribozymes exhibited preferences for biophysically similar substrates, indicating that co-option of existing ribozymes to adopt additional amino acids into the genetic code 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 how the genetic code may have evolved through the emergence and co-option of aminoacylation ribozymes.

Contents

Chapter 1

Preface

Living organisms represent some of the most complex chemical systems in the known universe. The seemingly limitless solutions designed by nature, which have produced the vast diversity of life on Earth have reshaped the planet through nano-scale machines, chemical factories, and brains with unmatched data processing abilities. These developments are the result of nature's great inventor, evolution. Improved understanding of the fundamental molecular principles of evolution would provide unprecedented advances for human society, from the ability to predict the emergence and progression of disease to the design and application of bio-based tools for solving global crises, the impacts of understanding these processes go well beyond addressing fundamental science questions. In recent years, researchers have attempted to harness the power of evolution through techniques like *in vitro* selection and directed evolution, which have revolutionized the bioengineering community by allowing for the creation of biomolecules with novel functions.[1,2] These methods rely on experimentally designed selective pressures that guide the evolution of desired traits to replicate successful designs and discard unsuccessful ones. Improved understanding of evolutionary principles will continue to aid in the utility of these approaches.

In addition to shaping the future, evolution can provide insight into the past. With advances in DNA sequencing, evolutionary biologists have attempted to reconstruct the evolutionary history of life on Earth, all the way back to the presumptive Last Universal Common Ancestor.[3–5] While researchers in fields from genetics to geology have made great progress in discerning historical information from contemporary observations, the limits of preserved information may forever cloud a complete picture from these methods. For this reason, efforts to recapitulate prebiotic conditions to model possible origins of life may prove to be one of the best avenues for understanding how the complexities of life can arise naturally. Since Miller and Urey,[6] researchers have attempted to simulate prebiotic events to better understand how biomolecules might arise abiotically. This research aims to not only address how life may have formed on an early Earth, but also in what other forms it might exist elsewhere in the universe. Still, how, or even whether, such a complex molecular system could have emerged spontaneously is the subject of much debate. Is life as we know it a predestined consequence of certain physical and chemical environments, simply a natural progression of the prebiotic chemistries that produced it? Or is it a low-probability event that is observable only as a result of producing its own observers? One key challenge to understanding the formation of life is in even defining the question. For every proposed definition of life, one can find a biological exception. However, one common theme among these definitions insists that life requires evolution,[7] thus providing a reasonable starting point for what life is and how it came to be.

The prevailing idea for a precursor to biological life is the RNA World Hypothesis, which posits that ribonucleic acids (RNA) were once the primary informational and enzymatic molecule.[8–10] Evidence for this is perhaps best observed in the biological translation machinery. Here, the informational molecule encoding genetic information is a messenger RNA (mRNA). The genetic code, defined as codons of three nucleobases which encode for corresponding amino acids, is translated through binding of specific transfer RNAs

(tRNA). A tRNA, which has been aminoacylated through the covalent linkage of its cognate amino acid, binds to the mRNA through base-pairing between the codon sequence on the mRNA and the anti-codon region of the tRNA. The ribosome, a large ribonucleoprotein complex, coordinates these interactions and links together amino acids in the growing polypeptide chain. While most cellular reactions are catalyzed by proteinaceous enzymes, the peptide bonds formed during protein synthesis are catalyzed by RNA, as the catalytic core of the ribosome is a ribozyme.[11] The central role that RNA plays in this essential biological process, and its unique ability to both retain genetic information and catalyze reactions, provide compelling evidence for the prior existence of an RNA World.

The manifestation of an RNA world would have required overcoming many challenges, including the evolution of novel molecular functions. Two crucial aspects in the evolution of biomolecules are optimization on the fitness landscape and co-option for new functions. The fitness landscape describes a function of fitness in the space of all possible sequences. Molecules evolve through a random walk on the fitness landscape, with a bias toward climbing peaks. Mapping the topography of fitness landscapes is fundamental to understanding evolution. In addition, the ability of enzymes, including ribozymes, to catalyze side reactions is believed to be essential to the evolution of novel biochemical activities. It has been speculated that the earliest ribozymes, whose emergence marked the origin of life, were low in activity but high in promiscuity, and that these early ribozymes gave rise to specialized descendants with higher activity and specificity. One particularly essential activity leading to the genetic code would be the reaction of ribozymes with activated amino acids to form aminoacyl-RNAs, allowing for the development of translation machinery. Co-option of these aminoacyl-RNAs could then lead to genetic code expansion, with error minimization as a by-product.[12]

The focus of this dissertation is to explore the evolutionary capabilities of catalytic

RNAs, including their emergence from random sequence space, functional optimization through evolution on the fitness landscape, and their co-option for new functionality. This text first introduces the concept of evolutionary fitness landscapes, discusses the challenges for their characterization, and describes recent efforts through the development of high-throughput approaches. Then, the role of promiscuity in evolutionary innovation is explored, particularly as it relates to ribozymes in a prebiotic context. Next, experimental methods for this work are described, followed by the results from the selection and characterization of self-aminoacylating ribozymes and the corresponding fitness landscape for this function. This work then explores the potential for these ribozymes to be co-opted for function with new substrates, along with the implications this process may have had in the origin of the genetic code. Finally, this text concludes with a brief discussion on the impact and outlook of this field of research. While much remains to be discovered on these topics, the hope is that this research will aid in the search for understanding of how new biologies can arise.

Chapter 2

Molecular Fitness Landscapes from High-Coverage Sequence Profiling

2.1 Permissions and Attributions

This chapter was the result of collaboration with Celia Blanco, Abe Pressman, Ranajay Saha, and Irene Chen and has been adapted from a version that previously appeared in Annual Reviews of Biophysics.[13] It is reproduced here with the permission of Annual Reviews, to which further permissions related to the material excerpted should be directed: https://www.annualreviews.org/doi/full/10.1146/annurev-biophys-052118-115333.

2.2 Introduction

Predicting evolution is a key challenge in biological science which not only tests our basic understanding but also has real-world ramifications. For example, prediction of influenza virus evolution[14] is used to select vaccine strains. In principle, evolutionary trajectories could be predicted probabilistically if one knew how any mutation would

affect the fitness of the organism or molecule (as well as knowing other parameters, including population size and mutation rate). The function of fitness in sequence space is known as the fitness landscape.[15,16] Evolution can be seen as a random walk (i.e., exploration by mutation) on a fitness landscape with a bias toward hill-climbing (i.e., selection for higher fitness).[17] Despite the importance of mapping fitness landscapes, the size of sequence space is astronomically large (m^N points for an alphabet size m and sequence of length N), which has previously hampered substantial mapping efforts. While experiments in the laboratory can include a large number of biopolymer sequences (e.g., up to $\sim 10^{17}$ molecules for *in vitro* evolution of RNA), analysis is also limited by sequencing capacity. Therefore, within the last decade, analysis has been transformed by the accessibility of high-throughput sequencing (HTS), as fitness data can now be collected on millions of sequences in parallel. These data form a quantitative framework for addressing classic questions: how does the topography of the fitness landscape constrain evolution? How repeatable are evolutionary outcomes? What does the topography teach us about the emergence of new structures and functions?

This chapter highlights progress that has been made to map fitness landscapes empirically using high-throughput techniques, with a focus on biomolecules. First, to give an initial context for these studies, simple models of fitness landscapes and their properties are introduced. Next, the case study of a classic question is considered: how well selection can optimize fitness on real landscapes and the impact of HTS on this problem. Attention is then devoted to other ways in which HTS has deepened our understanding of molecular fitness landscapes, where fitness approximates functional activity. Finally, organismal fitness landscapes and the importance of the environment are considered, a combination which is daunting in scope but the source of Darwin's "endless forms most beautiful".

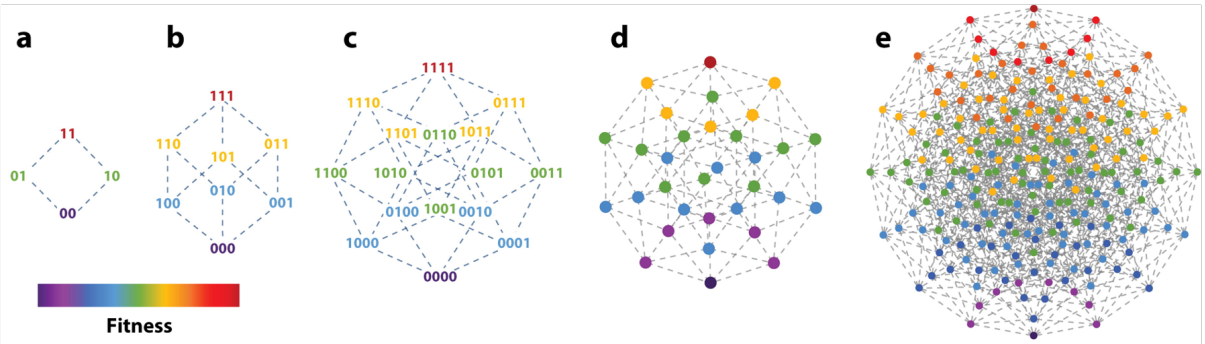

Figure 2.1: **Mock fitness landscapes of small binary sequences, depicted as a projection of the N-dimensional hypercube.** Landscapes are drawn with $m = 2$ and (a) $N = 2$, (b) $N = 3$, (c) $N = 4$, (d) $N = 5$, (e) $N = 8$. The fitness of each point in sequence space is represented by color (see legend) according to a smooth 'Mt. Fuji' landscape (e.g., fitness related to the number of '1's). As N increases, the number of points and neighbors increases exponentially, making a full representation of the fitness landscape difficult to interpret at higher N. Figure based on Wright 1932.[18]

2.3 Sequence Space

Sequence space is discrete, where the number of dimensions N is equal to the number of variable monomer sites in a biopolymer (e.g., with no fixed sites, N is the sequence length), and the number of points in each dimension is the alphabet size m. Fitness is a continuous variable that describes a sequence's evolutionary favorability, and can be defined depending on experimental context. Plotting fitness over sequence space gives the fitness landscape of $N + 1$ dimensions. To gain some intuition, one may draw the space of very small binary sequences, with fitness represented as a heat map (Figure 2.1).

For standard RNA or DNA, with an alphabet size of four nucleotides, the size of sequence space is 4^N ($\sim 10^{0.6N}$). The amount of nucleic acid one might work with *in vitro* would be typically $<10^{17}$ molecules, so sequence space becomes experimentally intractable in the lab for $N >\sim 27$ if one desires full coverage of the space. For standard proteins, composed of 20 amino acids, the space 20^N ($\sim 10^{1.3N}$) becomes intractable *in vitro* for $N >\sim 12$ at full coverage. For experimental evolution *in vivo* (e.g., in microbes),

a 1 L experiment might contain 10^{12} cells, allowing up to ∼20 genome sites to be covered in full. In practice, fitness landscapes can be fully mapped for relatively short sequences, while fitness landscapes for organisms and larger molecules must focus on a small number of variable sites or sparsely sample the sequence space.

Although sequence space is exponentially large, it is still a special subset of the larger space of all possible chemicals. Sequence space for a particular polymer type (biological or artificial) can be thought of as a sort of filigree in chemical space, defined by its particular bonding patterns, which is closely apposed to those for similar polymer types.[19]

2.4 Simple Models of Fitness Landscapes

Experimental investigation of fitness landscapes is difficult due to the complexity of sequence space, so a substantial body of work has involved the development of theoretical models of fitness landscapes. These models can be applied to biological data as a way to represent complex patterns with a small number of parameters. Although theoretical models for fitness landscapes have been reviewed elsewhere,[20,21] two simple and influential models (Mt. Fuji and NK) and related models (Rough Mt. Fuji and House of Cards) are discussed here to provide some intuition for possible topographies and their possible mechanisms of origin.

The simplest theoretical model is the 'Mt. Fuji' landscape,[22] named after Japan's highest mountain because it is a smooth, single-peak landscape. Mt. Fuji landscapes are defined as those in which every point on the sequence space – other than the global optimum – has at least one neighbor sequence (one mutational step away) of higher fitness. The simplest Mt. Fuji model corresponds to a perfectly smooth, monotonic climb along any path toward the center. This topography can be created if the effect of individual mutations are additive (the effect of each site does not depend on the others,

i.e., there is no epistasis). The absence of local optima on Mt. Fuji-type landscapes allows good reconstruction of the topography even when incomplete random sampling is performed. Under conditions of strong selection and weak mutation (SSWM),[23] evolution on Mt. Fuji-type landscapes results in the optimal sequence.

Most empirical landscapes exhibit certain epistatic interactions that the Mt. Fuji model cannot emulate. In particular, Mt. Fuji-type landscapes cannot describe reciprocal sign epistasis, in which the presence of one mutation a changes whether another mutation b is beneficial, and vice versa, creating multiple optima.[24]These non-additive effects disrupt the smoothness of a landscape, creating a need for models with tunable ruggedness. A popular model of this type is the NK landscape,[25,26] in which the system can be solely described by two parameters: the number of sites N, and the epistatic degree K (the number of other sites influencing the effect of a given site). When $K = 0$, the NK model gives a Mt. Fuji landscape. As K increases, the ruggedness of the landscape increases and local optima arise, although a global optimum is still present. In its most rugged incarnation, $K = N - 1$, the fitness contribution of a single position is affected by mutations at every other position in the sequence. In this case, the landscape is dominated by high-order epistasis, leading to a completely uncorrelated landscape with an average number of local optima ($2N/(N-1)$) that scales roughly exponentially with N (Figure 2.2). A landscape in which the fitnesses of related sequences are totally uncorrelated is also known as the random House of Cards model, because pulling a card (i.e., a mutation) from the house results in its collapse (i.e., complete change of the fitness landscape); the house then needs to be entirely rebuilt, reshuffling the genomic deck.[27] Although interesting as a theoretical limit, the completely uncorrelated landscape probably does not occur in reality. Whether incomplete sampling of sequence space can result in a reasonable representation of the topography depends on the ruggedness of the landscape and the properties to be analyzed.

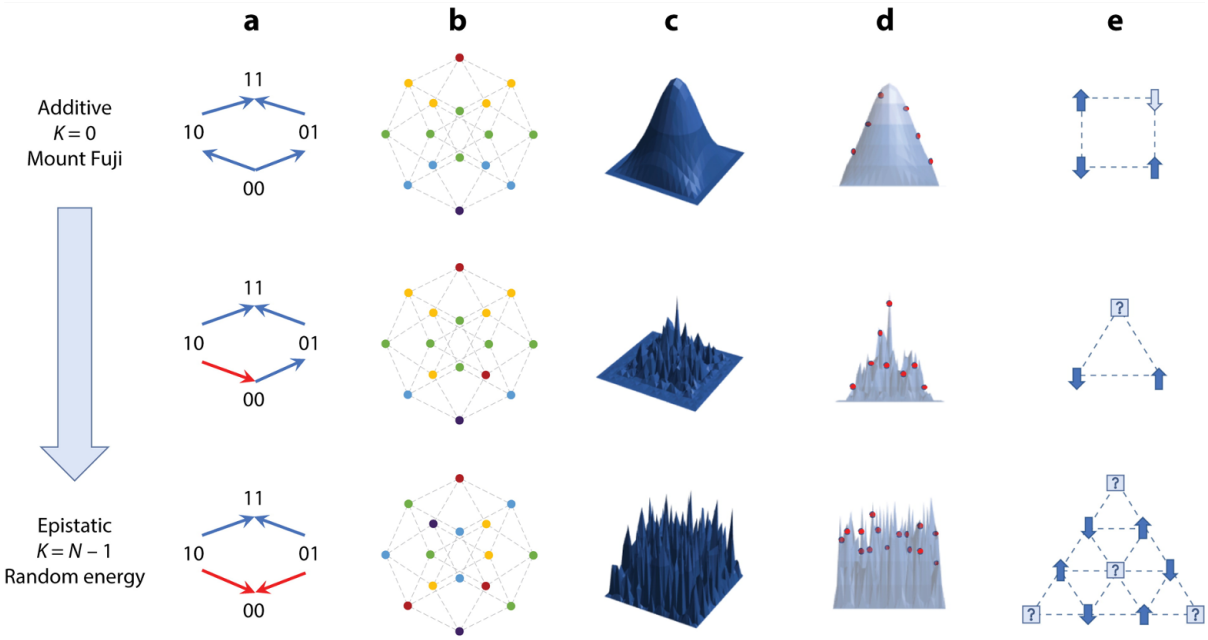

Figure 2.2: **Epistasis and ruggedness on a fitness landscape.** (a) For the simplest possible case ($m = 2$, $N = 2$), a smooth landscape can be climbed upwards from 00 to 11 (peak). Sign epistasis prevents passage over one trajectory, and reciprocal sign epistasis blocks both pathways. Fitness increase or decline is indicated by blue or red arrows, respectively. (b) A similar pattern can be seen for $m = 2$, $N = 4$ (refer to Figure 2.1c). (c) A conceptual 3D depiction of fitness landscapes with varying ruggedness; horizontal axes correspond to sequence space and vertical axis corresponds to fitness values. (d) Random sampling (red dots) can yield a better representation of smooth landscapes than of rugged ones. (e) Representation of frustration (or lack of) in a geometrical lattice of spins. With a smooth landscape, conditions leading to maximum fitness can be satisfied simultaneously. At high K (or p), conditions leading to maximum fitness (or minimum energy) conflict with each other and frustrate optimization.

Two modifications to the NK model can be introduced to increase its realism. First, since biomolecules are often modular (e.g., composed of independent domains), the NK model can be adapted to include different degrees of correlation on the landscape.[28] In the block (or domains) model, mutations in one block only affect the contribution of that block to the overall fitness of the biomolecule, and each independent block can have different values of K. Blocks need not correspond to structural domains from the primary sequence but could represent positions that interact in the tertiary structure.

Second, although the original NK model does not account for the presence of neutral mutations (i.e., mutations that do not change the fitness value), two different adaptations of the model incorporate this feature: the NKP model, where a fraction P of the fitness contributions have a value of zero, and the NKQ model, in which each fitness contribution can only take one of Q possible values. In the limits $P \to 0$ and $Q \to \infty$, the NKP and NKQ models correspond to the original NK model.[29]

Since its initial application to the maturation of the immune response,[26] the NK model has been used to describe experimental protein and DNA fitness landscapes.[30–32] Rugged regions in a landscape are described by high values of K, which can be estimated from the data, for example, by calculating the autocorrelation function for different values of K and comparing to the experimental system.[31,32] It is important to note that, since regions of the fitness landscape that are populated with closely related sequences of low fitness are described by $K \sim 0$, attempts to fit the NK model to landscapes over wide regions might result in artificially low values of K due to averaging over dissimilar regions of the landscape. Different parameters have also been proposed to measure epistasis in fitness landscapes (e.g. number of peaks, ratio of the roughness over additive fitness, or fraction of sign epistasis). Ferretti et al. recently proposed a new measure more directly related to epistasis, namely the single-step correlation of fitness effects for mutations between neighbor genotypes, which can also be used in landscapes with missing data.[33]

Tunable ruggedness can also be introduced into the Mt. Fuji model.[34,35] The 'Rough Mt. Fuji' model is the addition of a Mt. Fuji-type landscape and the uncorrelated House of Cards model. This model can include sign epistasis, in which the effect of a single mutant is positive or negative depending on the presence of another mutation (e.g., Figure 2.2a middle), provided there exists a different single mutant that is more fit than the double mutant. The ruggedness is tuned by varying the proportion of additive and random fitness components. Examples of landscapes with varying ruggedness are given

in Figure 2.2.

2.5 Case Study on Evolutionary Optimization: Neutral vs. Frustrated Networks

An important property of any fitness landscape is the ease with which evolution can optimize fitness. Whether this is feasible depends on the ruggedness of the landscape, and specifically on whether viable evolutionary pathways (i.e., uphill climbs under SSWM) allow access to the global optimum from distant areas of sequence space. Early computational work investigating this problem studied whether viable paths could be found connecting unrelated RNA sequences that were predicted to fold into the same secondary structure. These simulations, which took advantage of the high accuracy of RNA secondary structure prediction,[36] required conservation of the fold to define a viable path. These simulations revealed two related insights. First, they predicted that almost all common folds occur within any small region of sequence space.[37] Second, for common folds, the large set of sequences that share a given fold would form an evolutionary network throughout sequence space.[38–41] The fact that this set is large is important; if the fraction of sequence space that adopts the desired fold is low, then the folded sequences represent isolated regions in the space. However, if the fraction reaches a critical percolation threshold ($\sim 1/N$), the islands become connected and the landscape as a whole exhibits a neutral network.[42] A neutral network could be conceptualized as a fitness landscape topography that is full of 'holes', emphasizing the fact that high-dimensional sequence space has a non-intuitively vast number of potential connections.[43] These computational and theoretical considerations gave rise to the attractive hypothesis that 'neutral networks' might characterize molecular fitness landscapes, allowing evolutionary optimization over

large distances.

In contrast to this view of neutral networks, many empirical examples of epistasis are known in local sequence space, and one might expect that the extension of epistasis through the landscape (i.e., widespread ruggedness) would result in frustrated optimization during selection. This phenomenon can be mimicked in the NK model, which can be interpreted as a superposition of p-spin glass models[44] (Figure 2.2e). In spin glasses, the Hamiltonian of the system exhibits frustration when no spin configuration can simultaneously satisfy all couplings leading to a state of minimum energy. Since there is no single lowest-energy configuration, the energy landscape contains several metastable states separated by a distribution of energy barriers. The parameter p (number of interacting spin glasses) tunes the ruggedness of the energy landscape, much like K in the NK model. In the limit $p \to \infty$, it becomes impossible to satisfy all spin constraints and the system has an extremely rugged, uncorrelated potential surface, equivalent to Derrida's random energy model,[45] which is an analog of the random House of Cards model. Similarly, in the NK model, as K increases, configurations leading to the highest fitness contribution for certain positions become mutually incompatible, leading to blocked evolutionary paths over which optimization by selection is frustrated.

Ideally, experimental detection of a neutral vs. frustrated network would involve mapping the topography of a complete fitness landscape. However, due to the large size of sequence space for even small folded RNAs and the limits of sequencing throughput at the time, early work related to this question focused on construction of a viable evolutionary pathway between two nucleic acid sequences with different functions.[46–48] Several examples of protein evolution to produce new or altered function were also known (e.g., ref.[49]). These efforts were surprisingly successful, suggesting that different functions could be nearby in sequence space, i.e., fitness peaks for different functions can overlap.

Nevertheless, investigating evolutionary optimization on a single fitness landscape

requires identification of a very large number of functional sequences, and thus substantial progress had to await the advent of high-throughput sequencing. The first complete fitness landscape for short RNA sequences ($N = 21$) revealed very few viable evolutionary paths between different functional families.[50] Although this approach cannot be easily extended to much longer lengths, one attempt to evolve an RNA polymerase ribozyme ($N = 168$) at a high mutation rate did not find a new optimum.[51] Although this careful study was able to relate the results of the selection to the topography of the fitness landscape, it is possible that similar results in other systems are under-reported in the literature. These studies hint that frustration may characterize evolutionary optimization of a particular function for RNA for a relatively fixed landscape. Given the contrast between these frustrated cases and the apparent ease of evolving certain new functions, it is tempting to speculate that optimization of a single function might have quite different evolutionary properties than evolution of a new function.

2.6 Measuring Molecular Fitness Landscapes with High-Throughput Techniques

2.6.1 RNA and DNA: From Microarrays to High-Throughput Sequencing

When measuring fitness landscapes, functional nucleic acids present certain advantages compared to more complicated evolvable systems. In particular, an alphabet of only four nucleotides allows far higher coverage of random sequence libraries. Predominantly *in silico* approaches have shown some utility in predicting activity, such as in the generation of an effective anti-HIV aptamer (an RNA-based affinity reagent),[52] but such studies are relatively uncommon. On the experimental side, HTS for studying fitness

landscapes can be seen as the successor high-throughput technique following microarrays, paralleling the trend in genomics applications. Approximately $10^5 - 10^6$ sequences can be studied in reasonable copy number with a single HTS run (or microarray assay), equivalent to full coverage of sequence space with $N = 10$. Nucleic acid microarrays have been used to investigate double and triple-mutational scans of aptamers,[25] used with rational truncation to investigate the importance of structural constraints on aptamer activity,[53] and combined with *in silico* approaches to interrogate large local evolutionary spaces in array-based directed evolution.[54] A 2010 study was able to use array techniques to measure DNA-protein binding over all possible 10-nucleotide sequences, showing that although the fitness landscape contained only a single conserved active motif, the landscape contained sufficient ruggedness to produce many separate local fitness optima.[31]

But microarray approaches have been somewhat limited in their scope and adoption for multiple reasons, including their reliance on reactions or binding events producing a fluorescent signal and limitations stemming from attachment of the nucleic acid to a surface. Instead, HTS-based approaches have increasingly come to dominate RNA and DNA fitness landscape studies.[55] In 2010, Pitt and Ferré-D'Amaré demonstrated the ability of HTS to measure sequence enrichment during *in vitro* selection as an estimate of sequence fitness, generating a local landscape of approximately 10^7 mutant variants of a ligase ribozyme (catalytic RNA, Figure 2.3).[56] The increasing scale and affordability of HTS technology has made such measurements an accessible option. Further development of HTS measurement of fitness landscapes has focused on techniques to improve either landscape coverage or measurement of fitness.

To improve landscape coverage and interrogate larger sequence spaces, the limitation is not pool size (typically $10^{14} - 10^{16}$ molecules) but analytical capability, i.e., sequencing throughput (typically $10^6 - 10^8$ reads). It is possible to overcome this limit with *in vitro* selection – if selection can isolate nearly all of the high-activity sequences, com-

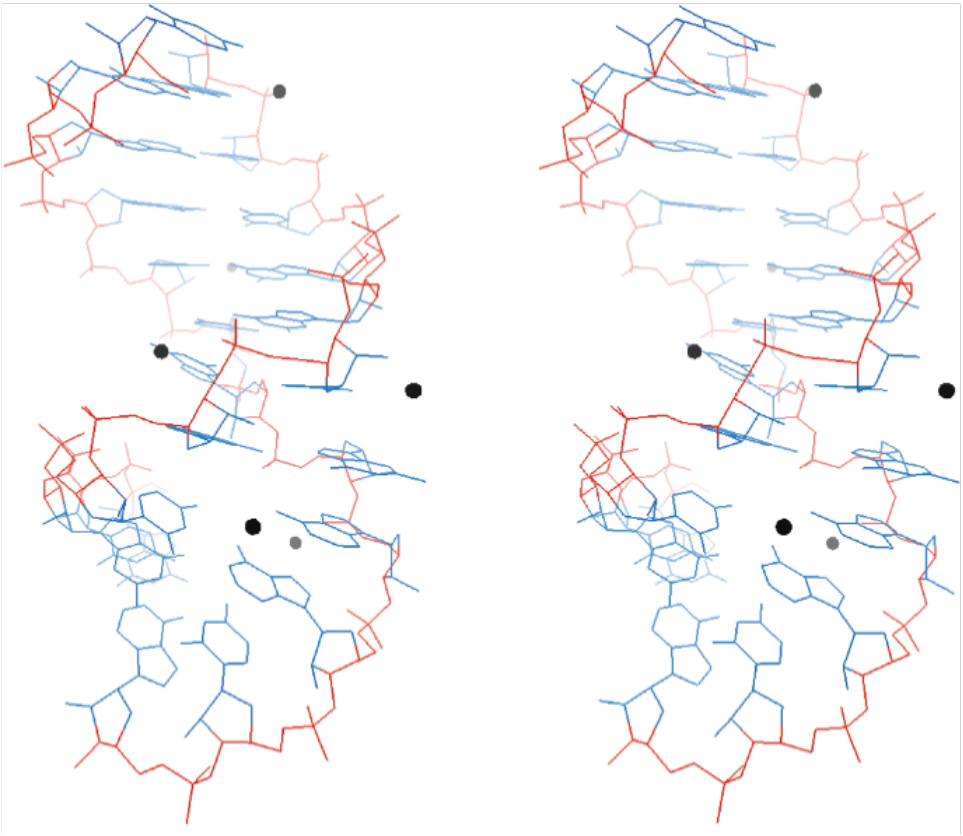

Figure 2.3: **Stereo view of the structure of the class II ligase ribozyme.**[56] PDB ID: 3FTM; image created with Visual Molecular Dynamics.[57]

plete mapping of an RNA fitness landscape becomes possible for short sequences. When studying molecular fitness landscapes *in vitro*, the interpretation of negative information can be powerful.[50] This requires a well-defined initial pool, but potentially expands the analysis, as it is no longer limited by the sequencing throughput but by the complexity of the initial pool, which is larger by several orders of magnitude. Although detailed information cannot be obtained about lost mutants, their disappearance indicates low fitness. It should be noted that epistasis and other studies should be interpreted with respect to the mutants analyzed. For example, if the mutants are not selected at random (e.g., survived a selection), epistasis values for that subpopulation would likely underestimate those for random mutants unless negative information is taken into account. At

the same time, sparse random sampling can also lead to inaccurate estimation of epistasis and ruggedness,[58] and the prevalence of indirect evolutionary pathways that bypass local valleys[59] could lead to underestimates of evolvability if the explored space is too small. However, depending on the hypothesis or question being investigated, *in vitro* selections from a large, random pool that only sparsely covers sequence space can still provide insights into general underlying trends in the larger, un-measurable spaces.[56,60]

For *in vitro* selection experiments, fitness is taken to reflect chemical activity, and can be estimated (or defined) in multiple ways, such as: abundance at the end of selection, enrichment over a single round, or functional activity under selection conditions. Ideally, all of these should be correlated as they are related to the true chemical activity of a given selected species. Abundance, however, can be surprisingly poorly correlated to chemical activity,[50,60] likely due to experimental noise and biases related to sequencing (e.g., PCR). Thus, new approaches use HTS to perform direct activity screens.[61–63] Furthermore, fitness estimates can be notably improved by considering multiple rounds of selection.[60]

High-throughput techniques are also being applied to measurement of RNA and DNA specificity. While these experiments often address different scientific questions than single-function fitness landscapes, they use similar techniques and analyses. HTS techniques were used to characterize the DNA binding landscapes of over a thousand transcription factors (TF).[64] These data enabled mapping of DNA-TF binding energy over large sequence spaces,[65] again illustrating the power of applying HTS to traditional questions.

2.6.2 Beyond DNA and RNA: Exploring New Chemical Space with High-Throughput Sequencing

Recent forays into the chemical space of nucleic acids (NAs) with altered backbones (XNAs) or modified bases raise the prospect that, with modern knowledge and techniques, parallel molecular biology could be developed for these alternative NAs in a relatively short time.[19] Alternative NAs raise many fundamental questions about fitness landscapes, from biologically inspired issues such as the uniqueness (or not) of RNA and DNA, to more abstract problems, such as the shape of the larger fitness landscape in chemical space. While chemical study of alternative NAs dates back to Eschenmoser's pioneering work,[66] investigations into their functional capacity began with altered bases, namely *in vitro* selection on reduced alphabets. Remarkably, ribozymes could be made from alphabets of only three[67,68] or even two letters.[69] In both cases, reduction in alphabet size led to selected ribozymes with lower activity than their larger-alphabet counterparts. On the other hand, artificially expanded genetic information systems (AEGIS) employ additional letters[70,71] and have been used to identify six-letter aptamers with greater affinity than those selected containing four-letters. While AEGIS currently poses some complications requiring probabilistic decoding of HTS data, HTS may still be applied to increase throughput compared to Sanger sequencing. Further advances to functionality are aided by a wider exploration of bases. For example, the incorporation of unnatural hydrophobic nucleobases, (e.g., 7-(2-thienyl)imidazo[4,5-b]pyridine (Ds), SOMAmers, click-SELEX) result in increased binding affinity to their protein targets.[72–74]

For some functions, the activity of functional DNA molecules is comparable to that of RNA molecules.[75] On some occasions, the sequence of a functional RNA can be simply synthesized as DNA and retain functionality,[59,76] sometimes requiring additional evolution.[77] These exceptional cases may arise if the major interactions are electrostatic or

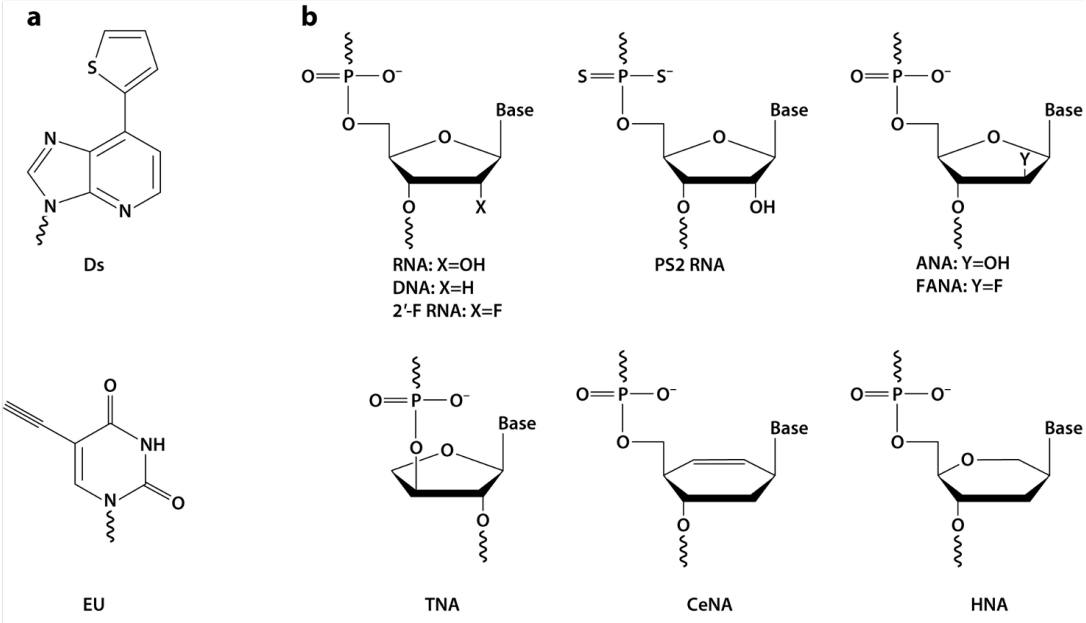

Figure 2.4: **Expanded chemical space of functional nucleic acids.** (a) The modified bases Ds (7-(2-thienyl)imidazo[4,5-b]pyridine) and EU (C5-ethynyl-uracil). (b) Chemical structures for RNA (ribonucleic acid), DNA (deoxyribonucleic acid), 2'-F RNA (2'-fluoro RNA), ANA (arabino nucleic acid), FANA (2'-fluoro ANA), PS2 RNA (phosphorodithioate RNA), TNA (threose nucleic acid), CeNA (cyclohexenyl nucleic acid), HNA (1,5-anhydrohexitol nucleic acid).

nonspecific stacking interactions. XNAs made from non-natural backbone alterations (Figure 2.4) have been selected for binding and catalytic activity, with activities similar to those seen in natural nucleic acids.[78–80] Introduction of phosphorodithioate linkages can improve aptamer binding,[81] with a single modified linkage increasing affinity by ∼1,000-fold in one case.[82] Another aspect of fitness is the chemical and physiological stability of the molecule; for example, many backbone modifications confer resistance to ribonuclease degradation.[83] Other modifications, such as 2'-fluoro and 2'-amino RNA, provide both added stability[84] and sometimes increased functionality.[85] The employment of chemical modifications to improve nucleic acids has been reviewed in more detail in ref.[86–88]

The application of HTS to alternative NAs is not trivial due to the need for engi-

neered polymerases to accept the template and read it out in a decodable way. Still, these challenges are being overcome by ingenious strategies.[70,78,80] Although XNA fitness landscapes are largely unstudied at the moment, it seems inevitable that some may demonstrate different or higher fitness peaks. Whether these changes will lead to new evolutionary properties is currently a fascinating unknown.

2.7 Fitness Landscapes of Organisms: RNA, Proteins and Genomes

Complete coverage of sequence space for an organismal genome – or even a single gene – is intractable due to the size of sequence space involved. However, local sampling around functional proteins (or random sampling of genomic mutants) still provides a rich source of data about the local landscape of the protein or the organism as a whole. Some examples of ways to represent HTS data are shown in Figure 2.5. Fitness landscape studies on sequences *in vivo* access fewer individuals ($\sim 10^{12}$ cells in 1 L) compared to *in vitro* studies. While this limits the diversity of the starting pool, it does not directly affect the number of mutants that can be assayed, since sequencing throughput is still limiting.

The *in vivo* fitness landscapes of small functional (non-coding) RNAs (tRNA and snoRNAs) in yeast have been investigated using HTS to study all single and double mutants. Because these cellular RNAs have smaller sequence spaces than proteins, such experiments can be done at higher mutational coverage, providing a good system for exploring *in vivo* fitness landscapes. In these cases, coverage of the local area around the wild-type sequence indicates that epistatic effects of mutation tend to be negative, with loss of fitness often corresponding to predicted disruption of RNA folding.[90,91] As more

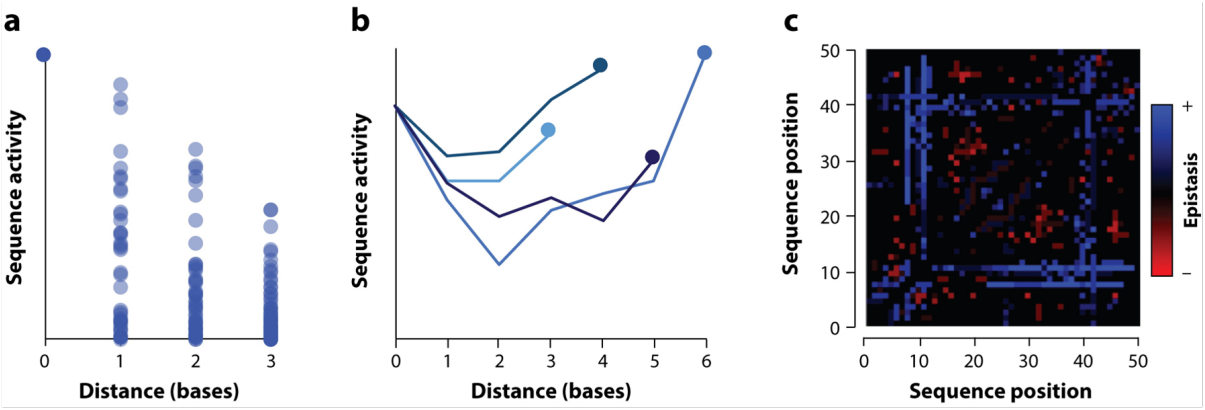

Figure 2.5: **Representing HTS data of fitness landscapes.** (a) A fitness peak with sequence space collapsed onto one dimension representing the edit distance (i.e., number of mutations) from the optimum sequence (after ref.[31,50,56]). (b) Evolutionary pathways between one local optimum and other nearby local optima, with sequence space collapsed as in (a). This representation illustrates fitness valleys and ruggedness (after ref.[31,50]). (c) Heat map representing combinations of mutants, revealing epistatic interactions along the length of a sequence (after ref.[89,90]).

RNA fitness landscapes are examined, it will be interesting to compare landscape characteristics of highly evolved biological RNAs vs. RNAs evolved *in vitro* to understand how >3.5 billion years of natural selection has shaped the landscape itself. Furthermore, the introduction of modified bases into cells[92] suggests the intriguing possibility of measuring fitness landscapes of alternative NAs *in vivo*.

The study of protein fitness landscapes, which began with mutational analysis (e.g., alanine scanning) and combinatorial studies of selected mutants, has been greatly impacted by HTS. Both m and N are substantially greater for proteins than RNA (e.g., the number of single mutant variants to be tested would be ∼6,000 for a typical single domain protein of length ∼300, compared to ∼150 for a typical ribozyme of length 50). The jump from Sanger sequencing to HTS has increased the number of mutants that can be analyzed by at least 4 orders of magnitude.

In an HTS technique known as deep mutational scanning (DMS), the activity of a mutant library is linked to organismal (cell or virus) fitness[93] (e.g., by cell sorting or

simply by reproduction and survival for influenza variants[94]); DMS has been further reviewed.[95,96] The survival of cells (or viruses) harboring the mutant library is measured by HTS, allowing assay of the fitness effect of $10^5 - 10^6$ protein variants. DMS has proven effective for creating high coverage, highly local fitness landscapes centered around a wild-type protein, and can identify sites of conserved function.[97] The local fitness landscape of the green fluorescent protein, measured over thousands of derivative genotypes, was found to be quite narrow, with the majority of single mutants showing reduced fluorescence.[98] On the other hand, DMS of a complete nine-amino acid region of Hsp90 showed that the distribution of fitness was bimodal, with one mode consisting of nearly neutral mutations and the other of deleterious mutations.[99] On a practical side, DMS results within yeast were used to optimize protein engineering, resulting in a new protein (with five point mutations) with a 25-fold increase in binding affinity to the influenza virus hemagglutinin.[100]

DMS is well-poised to measure local epistasis of a protein, since the fitness effect of many combinations of mutations can be measured. Even so, analysis of epistasis on *in vivo* protein landscapes is generally limited to a small number of peptide sites, a limited library of amino acid substitutions, or one specific set of evolutionary paths.[20] Weinreich et al. compiled a comprehensive review of these studies, showing that in these limited-landscape cases, *in vivo* protein epistasis tends to be primarily dominated by low-order epistatic effects of only a few loci,[101] although higher-order epistasis was notable in some cases. A local fitness landscape for four positions in protein GB1 revealed a very interesting feature – although many direct evolutionary pathways were blocked by reciprocal sign epistasis, these evolutionary dead ends could be avoided by following indirect paths in the sequence space.[59] Limited epistasis and evolutionary detours suggest short neutral pathways; whether these could combine over larger sequence space to form a neutral network is still unknown. However, sequencing technology continues to improve,

and may allow study of this question to be taken further in the future.

Although the theoretical models described earlier are highly simplified, one may ask whether empirical fitness landscapes can be fit to them. One 2013 meta-analysis found general trends in ruggedness and epistasis across a number of such studies, with many showing reasonable agreement with patterns expected from a Rough Mt. Fuji model.[102] Efforts to connect empirical data to these models are important for gaining an intuitive grasp of the topography of fitness landscapes. It remains an open question whether these models can also describe effects over organismal fitness landscapes of a larger scale, multiple peaks, or covering evolutionary sites on multiple genes.

2.8 Environment and the Fitness Landscape

It is nearly impossible to overstate the importance of the environment in determining the topography of a fitness landscape (Figure 2.6). At the microscopic level, molecular fitness depends on the temperature, water activity, pH, phase, cosolutes, and nearly any other environmental variable. These effects modulate both basic properties (e.g., RNA stability[103]) as well as sophisticated functions (e.g., ribozyme activity[104–106]). At the macroscopic level, genetic and environmental effects on traits cannot be simply deconvolved, as the heritability of any trait depends on the environment and genetic background in which it is measured. Even without environmental perturbations, the fitness landscape of a metabolizing organism is a continuously dynamic object, as organisms modify their environment, which changes the fitness landscape. Perhaps the most well-known example of this comes from the multi-decade experimental evolution of *E. coli*, in which changes to the genetic background ('potentiating' mutations) enabled evolution of the ability to metabolize citrate.[107] The efforts may also be driven by the potential for biomedical applications, as well: for example, DMS of a kinase involved in antibi-

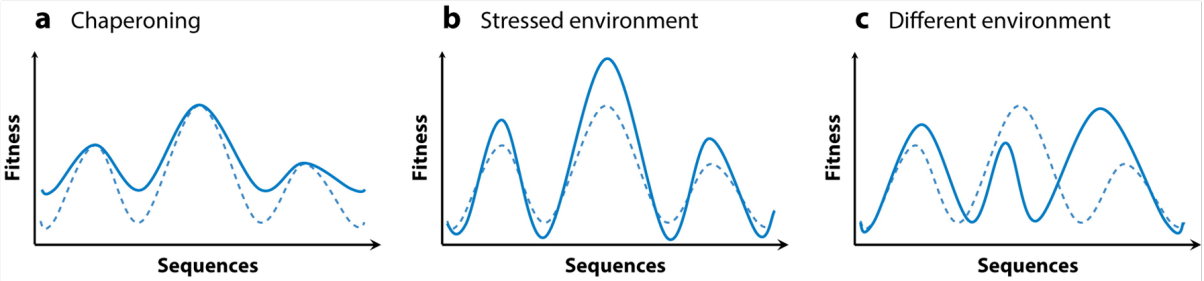

Figure 2.6: **The fitness landscape depends strongly on the environment.** For molecular fitness landscapes, environments might confer (a) stabilization of weakly folded structures (chaperoning), (b) exaggeration of fitness differences under stressed conditions, or (c) completely different structure in a new environment (c). The illustrations indicate the fitness landscape in one environment (dotted line) and in a new environment (solid line).

otic resistance characterized a fitness landscape that varies significantly over changes in both antibiotic concentration and structure.[108] Systematic study of the effect of the environment on the fitness landscape using HTS represents a major goal for this field.

The importance of the environmental context can be seen even in relatively simple molecular fitness landscapes for RNA. While most studies of functional RNA occur *in vitro*, it is clear that *in vivo* conditions may differ, sometimes greatly. For example, aptamer-based biosensors evolved *in vitro* show significantly lower performance in blood than in buffer.[109] Crowded and confined conditions can modify the structure and function of nucleic acids and proteins.[110–114] High levels of molecular crowding have been shown to stabilize mutations in ribozymes,[115] change the binding mechanism of a riboswitch to its ligand,[116] and create a chaperoning effect to assist in aptamer folding.[114] Ribozymes can also modify their environment (e.g., through cooperation[117]), presenting an attractive future target for mapping more complex fitness landscapes.

To study the effect of the environment on organismal landscapes, one common method is to expose the population to a new environment and observe the resulting evolution. In general, organismal fitness drops after environmental changes, but largely recovers through subsequent evolution and delayed adaptation at the genetic level.[118,119] For ex-

ample, changes to the fitness landscape of Hsp90 in *Saccharomyces cerevisiae* were observed in elevated salinity with previously adaptive mutations becoming deleterious in the new environment,[120] and the accessible evolutionary pathways in an esterase were shown to change at different growth temperatures.[121] Interestingly, variation in hosts may alter the topology of a viral fitness landscape, which may drive virus specialization.[122] However, whether the fitness landscape of a gene varies in different environments seems to depend on the details of the system. In contrast to cellular proteins, where a gene's fitness contribution often does vary with environment, studies of tRNA indicate that mutations influence the gene's fitness contribution by a fixed proportion independent of the environment, for four growth environments tested.[123] Further work in the yeast tRNA system also indicates that epistatic effects between loci can vary significantly for the same gene between different organisms.[124] If a mutation has multiple conflicting effects on fitness (antagonistic pleiotropy), adaptation to a new environment might be limited. Landscape analysis of the yeast genome shows that many gene variants display some degree of antagonistic pleiotropy in specific growth conditions.[125] The "environmental landscape" for a single sequence can also be measured, as was done for a riboswitch in nearly 20,000 different environmental conditions.[126] Measurement of such environmental landscapes in conjunction with fitness landscapes is a challenging but essential goal for which high-throughput techniques are essential.

2.9 Discussion

High-throughput sequencing has transformed the study of fitness landscapes, expanding the focus from theoretical models to empirical mapping. Increased sequencing throughput is more than a quantitative extension, as it allows exploration of fundamentally new areas of science, from evolutionary networks to environmental landscapes. To

maximize the knowledge return from this exciting growth of data, perhaps two aspects should be kept in mind. First, attention should be paid to building intuition and understanding, such as by analyzing the fit of data to idealized model landscapes. Second, while raw HTS data can be submitted to databases such as the NCBI Sequence Read Archive, a dedicated resource for submitting and viewing fitness landscape data could facilitate meta-analysis, standardization, and contributions from a greater community of researchers. Regardless, HTS-enabled mapping of fitness landscapes brings the tantalizing prospect of predicting evolution still closer to reality.

Chapter 3

Promiscuous Ribozymes and Their Proposed Role in Prebiotic Evolution

3.1 Permissions and Attributions

This chapter was the result of collaboration with Celia Blanco, Huan Peng, Josh Kenchel, and Irene A. Chen and has been adapted from a version that previously appeared in Chemical Reviews.[127] It is reproduced here with the permission of ACS, to which further permissions related to the material excerpted should be directed: `https://pubs.acs.org/doi/full/10.1021/acs.chemrev.9b00620`.

3.2 Introduction

Catalytic RNA sequences, or ribozymes, are widely accepted to have been central to the origin of life.[9,128] Their dual capacity for information storage and catalytic activity

is the basis for the RNA world theory,[129–131] that an RNA-based metabolism could have preceded the more complex DNA-RNA-protein system that is observed in biology today. Regardless of whether an RNA world existed on the early Earth, ribozymes represent an excellent laboratory model system for molecular evolution. Beginning with a pool of random sequences, strategies can be devised to select for particular activities. Cycles of selection and amplification by PCR allow enrichment and eventually isolation of active sequences. A prerequisite of successful *in vitro* evolution is the presence of one or more molecules with some activity, however slight, in the initial pool or early rounds. Once this kernel of activity exists, the active sequences can be selected and activity possibly improved by mutation during the evolutionary process. In addition to developing new ribozymes, *in vitro* evolution of RNA allows well-controlled experiments to observe and analyze the de novo emergence of biochemical functions.[132–134]

Promiscuous catalytic activities have been invoked as being particularly significant for the origin of enzymes,[135,136] as one might intuit that early, simple ribozymes or enzymes would have little specificity, and therefore might catalyze many reactions, albeit with slow rates. These sequences might possess kernels of activity for many different substrates or reactions (Figure 3.1). One landmark study of such a ribozyme is a sequence which was engineered to adopt two possible folds, one of which acted as a ligase and one of which acted as a self-cleaving ribozyme.[137] This sequence had very low activity for each function, but a relatively small number of mutations would increase function to near wild-type in both directions. Such promiscuity would promote evolutionary innovation by poising sequences at a non-zero fitness for multiple activities, each of which could be potentially optimized by natural selection. This idea also raises the interesting question of whether ribozymes are intrinsically more promiscuous than protein enzymes. From extensive work on the directed evolution of enzymes, it has become clear that much of the success of the field as a whole is due to the presence of low levels of apparently

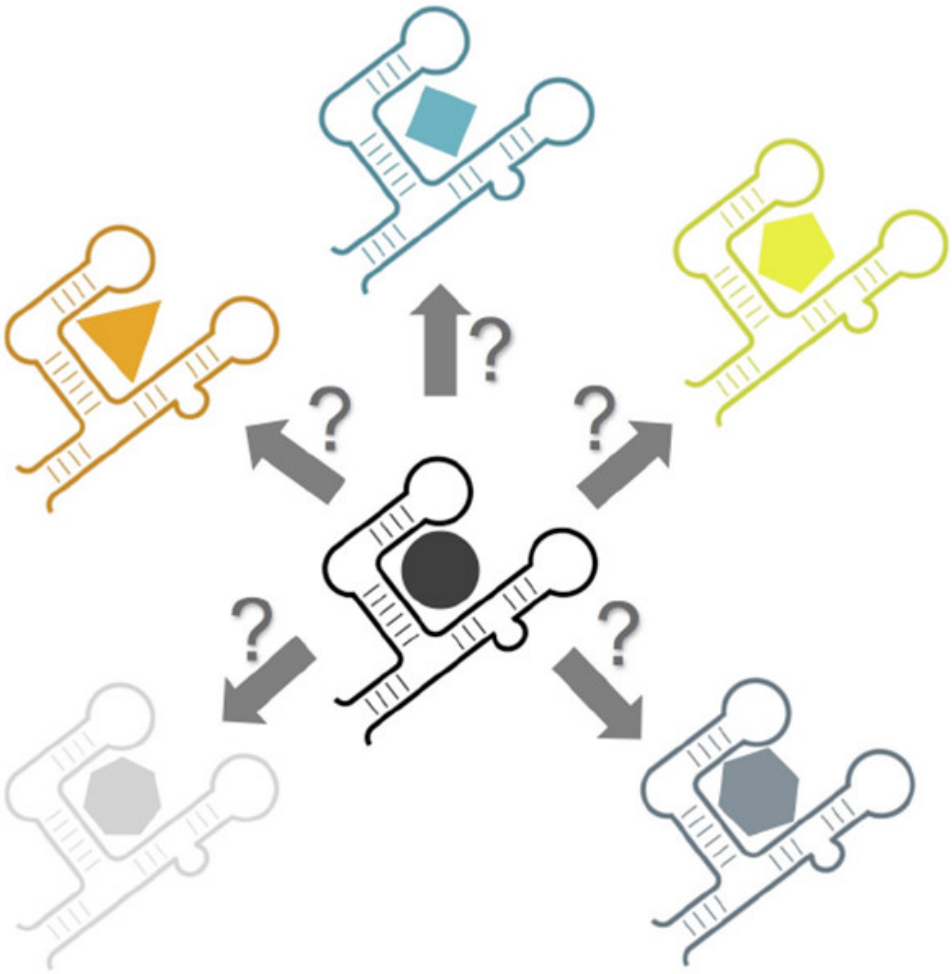

Figure 3.1: **What role might the evolution of promiscuous ribozymes play in the origin of an RNA World?**

promiscuous activity in extant protein enzymes.[138] This surprising degree of promiscuity in highly evolved enzymes suggests that promiscuity is actually the rule rather than an exception for protein enzymes.

This chapter reviews what is known about the specificity and promiscuity of ribozymes. First, major concepts and definitions in specificity and promiscuity are introduced, which were originally developed in the enzyme literature. An interesting concept is the relationship between activity and specificity, which underlies the intuition that early, relatively low-activity ribozymes would be more promiscuous. Then, several

cases of ribozymes in which studies have demonstrated promiscuity in some way are reviewed. When possible, a promiscuity index is calculated from what is known about these ribozymes, a first step toward rigorous comparisons of the promiscuity of ribozymes and protein enzymes. This chapter ends with a discussion of the implications of these comparisons for the hypothesis that early ribozymes were particularly promiscuous.

3.3 Promiscuity and Specificity: Concepts and Definitions

3.3.1 Defining specificity

Specificity is the ability of an enzyme to discriminate between two different substrates, assuming both are present. The question of how to measure enzyme specificity has been a matter of debate in the past (see[139,140] and references therein), but it is generally agreed that specificity in the presence of two different substrates should be compared based on the discrimination factor,[141] defined as the ratio of the catalytic efficiencies (k_{cat}/K_M) for the corresponding reactions. According to transition state theory, the logarithm of the catalytic efficiency (k_{cat}/K_M) is proportional to the free energy difference between the free enzyme and substrate vs. the transition state complex ($\Delta G^{\ddagger}$).[142] When comparing the cognate with an alternate substrate, the discrimination factor is also called the accuracy A. Thus A is exponentially dependent on the difference $\Delta\Delta G^{\ddagger}$ between the cognate and alternate substrates.

In some cases (e.g., polymerases), the use of an error ratio (the rate of incorrect product formation divided by the rate of correct product formation) is more appropriate. To gain an intuition about the possible scale of this ratio, one may note that the theoretical maximum discrimination between alternative substrates undergoing analogous reactions

occurs when the formation of the enzyme-substrate complex is much faster than product conversion and release (as assumed in Michaelis-Menten kinetics). In this case, the theoretical minimum error ratio is equal to the ratio of K_M values.[143,144]

3.3.2 General Mechanisms for Specificity and the Possible Trade-Off with Rate

Discrimination among substrates can arise from different affinities in the initial enzyme-substrate complexes (ground-state discrimination) or in the transition-state complexes (catalytic or transition-state discrimination).[145] The accuracy (A) for a cognate vs. alternative substrate can be increased by three scenarios: (a) higher rate of substrate association (ground-state discrimination with $k_{on}^{cog} > k_{on}^{alt}$), (b) lower rate of substrate dissociation (ground-state discrimination with $k_{off}^{cog} < k_{off}^{alt}$) or (c) higher rate of conversion of the enzyme-substrate complex into the transition state (transition-state discrimination with $k_{cat}^{cog} > k_{cat}^{alt}$).

In ground-state discrimination, lowering the energy of the enzyme-substrate complex has two effects, namely decreasing K_M as well as decreasing k_{cat}. In other words, although selectivity may be improved via increased substrate affinity, the reaction rate suffers. Examples of enzymes exhibiting ground state discrimination include DNA methyltransferases and the ribosome.[145] The tradeoff between accuracy and rate might impose an evolutionary constraint limiting selectivity.[145] Indeed, selection for activity on one substrate does not seem to induce high selectivity by itself,[146] and therefore negative selection against undesired substrates is used when engineering new enzymes.[147,148] Interestingly, a tradeoff between rate and accuracy created by ground-state discrimination would contradict the idea that early, less optimized ribozymes or enzymes were more promiscuous.

On the other hand, in transition-state discrimination, which tends to apply with relatively small substrates (e.g., DNA polymerases[145,149–152]), lowering the activation barrier increases k_{cat} without necessarily affecting K_M. Thus, in principle, transition-state discrimination might achieve higher selectivity at high activity since there is not necessarily a tradeoff between accuracy and rate. In addition, non-equilibrium mechanisms driven by release of chemical energy may improve selectivity with or without a tradeoff between accuracy and rate.[153] Furthermore, such mechanisms can allow accuracy to surpass the theoretical thermodynamic limit based on binding energies. For example, in kinetic proofreading,[143,154] discrimination between two possible substrates is achieved by the presence of one or more irreversible steps in the reaction pathway, whose rate(s) are biased by the identity of the substrate. These steps are made irreversible by consumption of chemical energy, and concatenation of such steps could be used to achieve arbitrarily small error ratios, in principle. Some biological processes can afford high specificity by using this mechanism.[143] For example, although the valine concentration *in vivo* is ∼5-fold higher than that of isoleucine, and isoleucyl-tRNA synthetase favors the reaction with isoleucine over valine by only ∼100-fold, the rapid hydrolysis of mis-incorporated valine-tRNA decreases the error ratio to 1 in 3000. While kinetic proofreading can increase reaction specificity substantially, this comes with a relatively high energetic cost.[145,155]

However, in the absence of proofreading mechanisms, substrate specificity is inherently limited due to physicochemical reasons. Indeed, a recent survey of the BRENDA database (The Comprehensive Enzyme Information System[156]) suggests discrimination is usually much lower than the theoretical maximum.[141] In the case of substrates differing by a single methyl group, discrimination was found to be lower than the theoretical maximum, for 23 out of the 25 enzymes surveyed, by typically 1-2 orders of magnitude. Interestingly, a similar discrepancy is found in non-enzymatic, template-directed polymerization of activated nucleotides,[144] suggesting that this phenomenon is not specific

to enzymes. A discrimination level lower than the expected theoretical maximum might reflect prioritization of increased rate during evolution, if the enzyme is subject to an accuracy-rate tradeoff; in other words, the marginal fitness benefit of increased specificity may come with a larger fitness decrement due to slower rate. Thus, in general, specificity tends to be lower than the theoretical maximum, possibly because of the costs associated with accuracy.[145]

Specificity may appear to be quite suboptimal even for presumably highly evolved enzymes. For example, the carboxylase enzyme Rubisco plays an essential role in fixing atmospheric carbon dioxide into sugars during photosynthesis. However, considering its biomass and critical role, it is surprisingly slow and non-specific, as oxygenation constitutes a major side reaction. Tradeoff models have been proposed to explain the observed correlations between specificity and other kinetic parameters,[157,158] which were recently revisited using an extended dataset.[159] A strong correlation was found between the catalytic efficiencies for carboxylation and oxygenation, indicating that lowering the effective CO_2 addition energy barrier (i.e. faster carboxylation) entails a similar reduction in the effective O_2 addition energy barrier (i.e. faster oxygenation). Therefore, the accuracy of Rubisco appears to be highly constrained.

3.3.3 Promiscuous vs. Multispecific Enzymes

The term 'catalytic promiscuity' was originally used to refer to enzymes known to catalyze more than one type of reaction.[135,160] However, in practice, 'promiscuity' has not been well-defined and thus has been used to refer to fundamentally different phenomena.[141,161] Generally, catalytic promiscuity refers to the capability of enzymes to catalyze reactions mechanistically different from the primary biological reaction[160] and substrate promiscuity refers to the capability of enzymes to transform different substrates.[162] These

terms warrant additional consideration here, as their usage varies and can depend on incomplete knowledge.

The native function of an enzyme refers to the physiologically relevant chemical transformation and substrate for which an enzyme has evolved. Native function is selected for and contributes to organismal fitness. In this context, any physiological functions for which an enzyme has evolved are considered native, even if they are not the enzyme's primary function. For example, while the primary function of aminoacyl-tRNA synthetases is to catalyze the attachment of tRNAs to their respective amino acids, some also catalyze generation of 5',5'-diadenosine tetraphosphate in a reaction that appears to be physiologically relevant,[163] and thus this additional function would be considered native. In practice, whether a particular function contributes to organismal fitness may be difficult to assess.

It is nowadays well accepted that many, if not most, enzymes have multiple side activities.[136,164,165] However, such side activities may or may not be a product of evolution. In the evolutionary biochemistry literature, promiscuity refers to side activities that are non-native (i.e., not evolved), so the alternative transformation or substrate is fortuitous. By definition, there is no evolutionary pressure on non-native activities, as they do not impact organismal fitness (e.g., alternative substrates are not available in the cell).[136,164,166] For biologically evolved enzymes, promiscuity (as defined by evolutionary biochemists) is nearly impossible to ascertain in practice, since we do not know what past environments and selective pressures might have applied to the protein. If promiscuity of a naturally evolved enzyme is suspected, one might use of the term apparent promiscuity, in contrast to true promiscuity, to acknowledge this uncertainty.

Interestingly, there are two special scenarios in which true promiscuity can indeed be characterized unequivocally. First, if the enzyme transforms a man-made compound not present in nature, the enzyme could not have evolved this activity and the activity

must be non-native. Examples are the atrazine chlorohydrolase and melamine deaminase enzymes, which degrade the man-made compounds atrazine and melamine, respectively. Despite very high similarity (98% identity), both enzymes show little activity on the alternative substrate.[139] While it is likely that their host strains evolved the atrazine or melamine degradation function in response to environmental exposure (both strains were isolated from areas contaminated by the substrate[139,167]), it may be presumed that neither strain experienced both contaminants simultaneously. If so, these enzymes can be considered as lacking in true promiscuity.[168,169] The second scenario in which true promiscuity might be determined is in the case of *in vitro* evolved enzymes and ribozymes, in which the different environments and selective pressures applied to the sequences are known.

An important contrast to promiscuous enzymes, whose side reactions are non-native, is multispecific enzymes (or broad-specificity enzymes), which evolved to perform many native transformations, such as on a broad range of available substrates. These enzymes are characterized by small accuracy values, with different substrates having similar k_{cat}/K_M. For example, theta class glutathione transferases from various species can catalyze the conjugation of the tripeptide glutathione to a variety of electrophilic substrates.[170] The enzyme family of cytochromes P450 metabolizes a variety of different substrates, with activities including biosynthesis of steroids, fatty acids, or fat-soluble vitamins as well as the degradation of herbicides and insecticides. In particular, cytochrome P450 3A4 contributes to the metabolism of approximately 50% of marketed drugs.[171–173] Most of the terpene cyclase enzymes (a.k.a. terpene synthases) are also multispecific. For example, the class I sesquiterpene cyclase gamma-humulene synthase generates 52 different products, of which gamma-humulene constitutes less than 30% in abundance.[174,175] Methane monooxygenase oxidizes more than 150 different substrates.[176] The RecBCD nuclease, originally named Exonuclease V, accepts both linear double-stranded DNA

and single-stranded DNA with very low specificity.[177,178] The distinction drawn between promiscuous and multispecific enzymes hinges on whether the additional substrates represented selective pressures on the enzyme. While this is an important conceptual distinction, assessing whether an enzyme is promiscuous vs. multispecific may be difficult in practice due to lack of knowledge of the evolutionary and environmental history of the enzyme.

It should be noted that additional usages of the term 'promiscuous' also exist. Promiscuity is sometimes used to refer to the capacity of an enzyme to transform different physiologically relevant substrates (see[141] and references therein), to be contrasted with 'multifunctional enzymes' whose side activities may be either physiologically useful or detrimental.[179] Unfortunately, this definition of promiscuity can be contradictory to the one given earlier, in which promiscuity refers to the capacity to perform non-native reactions. In addition, the determination of physiological relevance is difficult to make and again raises questions of evolutionary history. A third usage of the term promiscuity refers to enzymes whose catalytic domain executes multiple functions.[179–182] This review favors the definition from the evolutionary biochemistry literature, since ribozymes are often evolved under known conditions *in vitro*, allowing true promiscuity to be characterized, while physiological relevance is unspecified and multiple domains are relatively uncommon.

3.3.4 Promiscuity and Evolutionary Innovation

Fortuitous side reactions of a promiscuous enzyme are believed to be central to evolutionary innovation, as an initial kernel of activity for a side reaction is a starting point for optimization of the new activity by evolution.[135] In addition, an enzyme might exhibit new side activities under new environmental conditions (e.g. temperature, pH),

and such enzymes are called condition-promiscuous enzymes. Conditional promiscuity is a possible path for adaptation in new environments.[183–185] However, in the absence of selective pressure, side activities would be subject to neutral drift and may be lost if they are uncorrelated to native functions of the enzyme.

Interestingly, contrary to native functions, which are usually tolerant of mutations, directed evolution of enzymes has shown that the non-native functions can be greatly optimized by just a small number of mutations.[146,182] The flip side of this so-called plasticity is that newly evolved non-native functions are typically not tolerant to mutations. One might therefore suspect that evolutionary robustness, if it is observed for native functions, likely evolved as a trait or correlate of a selected trait. In addition, the duality of plasticity for non-native functions and robustness for native functions implies that evolutionary optimization of non-native functions might not always lead to a significant decrease of the original native function.[186] However, in the absence of continued selection pressure on the original function, specialization has been shown to occur due to tradeoffs during selection of the secondary function, even without negative selection against the original function.[187]

The idea that enzyme evolution and promiscuity are connected goes back to the mid-1970's, when Yčas and Jensen proposed, independently, the first model for enzyme evolution.[135,188] This general model hypothesizes that primitive life had minimal gene content and the number of available enzymes was limited. It posits that primordial enzymes may have been less specific, being able to catalyze broad classes of reactions on a variety of substrates. Gene duplication and mutation would have then increased genetic diversity, leading to the divergence of new enzymes whose secondary activities might give an evolutionary advantage under newly encountered selective pressures.[189,190] This hypothesis is widely accepted, although direct evidence is scant.[141,182] The molecular processes and evolutionary forces involved in the biological evolution of enzymes

are very difficult to reconstruct, and hence the mechanisms under which duplication and specialization events shape enzyme evolution have been the subject of much debate.[191] Nevertheless, understanding how specificity and promiscuity arise during *in vitro* selection and evolution of ribozymes, recapitulating an origin of life, can address this problem experimentally.

3.3.5 Promiscuity and the Fitness Landscape

The fitness landscape is a well-studied conceptualization of evolution through the space of all possible sequences (sequence space).[13,20] Each point of sequence space is specified by a sequence and its associated fitness (e.g., activity on a given substrate), giving the fitness 'landscape' in sequence space. At each point in sequence space, one might also imagine the large chemical space of possible substrates, and an activity profile for that sequence over substrate space, which reflects the promiscuity of the sequence.

Properties of the fitness landscape are not necessarily expected to correlate with properties of the promiscuity profile. Different fitness landscapes over sequence space can give rise to the same promiscuity profile in substrate space (Figure 3.2). In general, optimization for higher activity need not correspond to increased specificity. However, specific mechanisms, such as a tradeoff between rate and specificity, could produce correlations between the fitness landscape and the associated promiscuity profiles. The idea that early, non-optimized ribozymes were particularly promiscuous would translate into a correlation in which highly active sequences on the fitness landscape have lower promiscuity compared to less active sequences. Whether the specificity of a ribozyme can be improved through mutation or evolution would depend not only on the specificity of that individual ribozyme, but also on the fitness landscapes for the cognate and alternative substrates (Figure 3.2).

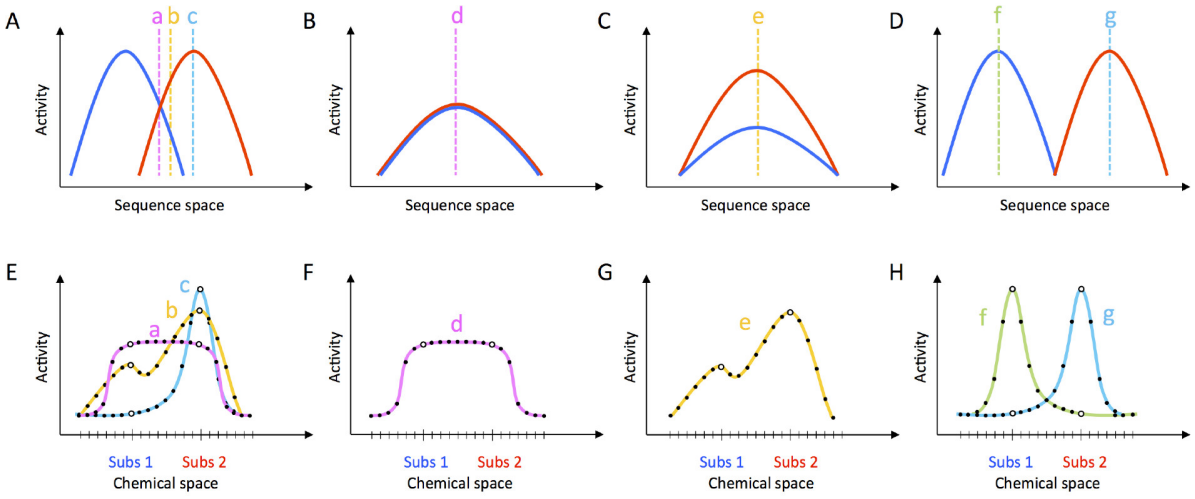

Figure 3.2: **Fitness landscapes and promiscuity profiles.** Different fitness landscapes (A-D) can correspond to the same promiscuity profile (E-H). Vertical dashed lines (a, b, c, d, e, f, and g) correspond to different ribozyme sequences. Ribozyme fitness landscapes (A-D) for two substrates may differ (blue and red) with or without overlap. The promiscuity profile (E-H), depicted here for two substrates (1:blue and 2:red) depends on the sequence tested, as seen in the comparison among sequences a, b, and c in panel E. In addition, similar promiscuity profiles can be derived from qualitatively different fitness landscapes. Compare sequence a from (A,E) with sequence d from (B,F), sequence b from (A,E) with sequence e from (C,G), and sequence c from (A,E) with sequence g from (D,H). While ribozymes a and d have similar promiscuity profiles, their evolutionary potential is strikingly different. Ribozyme a could evolve through mutations to specialized activity, but ribozyme d is already at a local maximum and has no evolutionary potential for increasing activity. Similarly, ribozymes b and e have the same promiscuity profile, but only ribozyme b has the possibility to evolve into a sequence of higher activity and selectivity. Ribozymes c, f, and g are highly specific, but unlike ribozymes f and g, ribozyme c has increased potential to evolve into a promiscuous ribozyme.

3.3.6 Quantifying Promiscuity: The Promiscuity Index

Several possible methods exist to quantify substrate specificity. The promiscuity index (I) proposed by Nath and Atkins is a metric similar to a normalized information entropy:[192]

$$I = -\frac{1}{logN}\sum_{i=1}^{N}\frac{e_i}{\sum_{j=1}^{N}e_j}log\frac{e_i}{\sum_{j=1}^{N}e_j}$$

where N is the number of substrates that can be transformed and e_i corresponds to their individual associated catalytic efficiencies. Due to the normalization, this metric goes from 0 (only uses one substrate) to 1 (equally efficient on all N substrates).

While this promiscuity index is simple and intuitive, it might be strongly influenced by the experimenter's choice of substrates to test. In particular, when comparing promiscuity indices for different ribozymes or enzymes, one sequence might appear more promiscuous only because many chemically similar substrates were assayed. To account for this problem, a weighted promiscuity index (J) factoring in substrate similarity can be calculated:[192]

$$J = -\frac{N}{(\sum_{i=1}^{N}\langle\delta\rangle_i) log N}\sum_{i=1}^{N}\langle\delta\rangle_i \frac{e_i}{\sum_{j=1}^{N} e_j} ln \frac{e_i}{\sum_{j=1}^{N} e_j}$$

Chemical similarity can be calculated using a bitwise dissimilarity metric between a pair of substrates (δ), which is based on the presence or absence of a number of different functional groups.

Any method for quantifying promiscuity from experimental data is likely to be biased in at least two ways. First, there is an experimental bias in the selection of substrates (e.g., synthetically accessible, similar to known substrates). For comparisons among enzymes, differences in these biases might affect the promiscuity index calculated, even when using the weighted value. Second, these metrics do not consider the chemical context in which an enzyme functions. If the environment never provides a certain substrate, it may not be justifiable to include such a substrate in the calculation even if the enzyme has non-zero activity on it *in vitro*. Additionally, the relationship between chemical similarity and promiscuity has not been well established, and often little difference between unweighted (I) and weighted (J) values has been observed.[192,193] Other metrics for promiscuity also exist, such as a measure based on structural information of the catalytic residues.[194,195]

Despite these limitations, the promiscuity index serves as a starting point for charac-

terization and comparison of substrate specificities. In this review, promiscuity indices are calculated for ribozymes for which sufficient data is available in the literature.[196] However, when necessary, catalytic rates were used in place of catalytic efficiency when the catalytic efficiency was inappropriate or unknown. Quantitative metrics like the promiscuity index provide the opportunity to compare the specificity of different molecules and potentially study the relationship between promiscuity and other measurable characteristics, such as activity.

3.4 Ribozymes Illustrating Promiscuity

This section first describes substrate promiscuity using aminoacylation ribozymes, for which different substrates have been studied in some depth. Then, to gain mechanistic insight into a specific case, focus turns to the hammerhead ribozyme, where specificity can be understood in terms of RNA annealing. An important consequence of this mechanism is that promiscuity is dependent on environmental conditions, such as temperature. The expression of promiscuity under new conditions (conditional promiscuity) is a possible mechanism for uncovering latent side activities. Then, a series of *in vitro* evolution experiments seeking an RNA replicase are discussed, in which the presumption of promiscuous activity was essential to the design and success of the experiments. Next, a different kind of promiscuity, catalytic promiscuity, is described, in a case of a nucleotide synthase ribozyme that unexpectedly possesses two distinct catalytic mechanisms. This section ends with a brief discussion of the ribosome, a proteinaceous ribozyme whose promiscuity appears to be unparalleled.

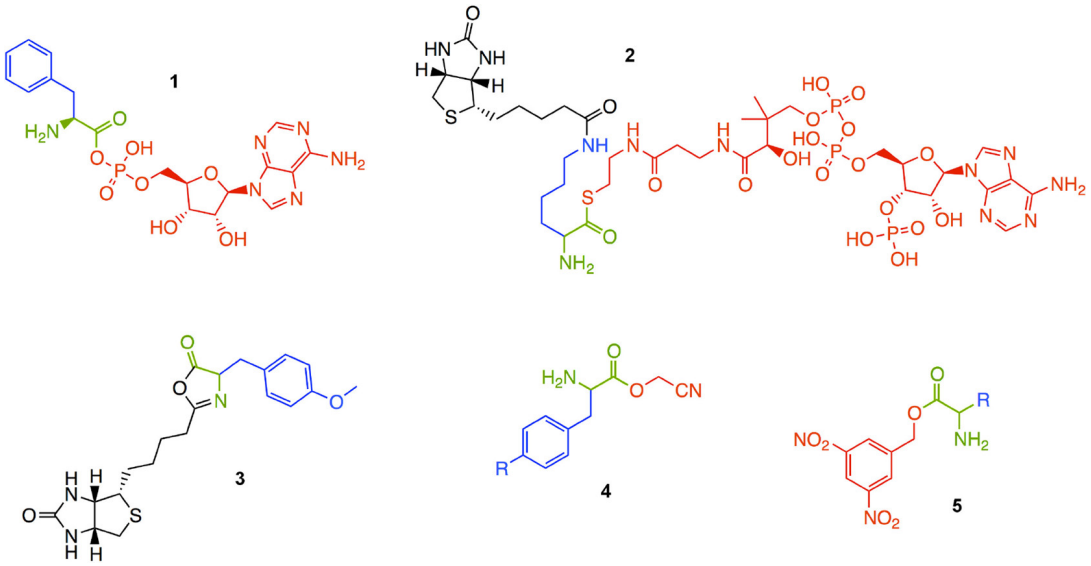

Figure 3.3: **Substrates for aminoacylation ribozymes.** Phenylalanyl-adenosine monophosphate[197,198] (1), BiocytinCoA[199] (2), biotinyl-Tyr(Me)-oxazolone[200] (3), amino acid cyanomethyl ester[201] (4) and amino acid 3,5-dinitrobenzyl ester[201] (5). Substrates 4 and 5 are flexizyme substrates. The amino acid backbone is depicted in green, side chains are depicted in blue, and leaving groups are depicted in red. R indicates possible chemical variation in the side chain.

3.4.1 Substrate Promiscuity: Aminoacylation Ribozymes

Aminoacylation of tRNA is a key step in protein synthesis, and high selectivity for tRNA-amino acid pairs is crucial for the stability of the genetic code.[202] It is presumed that ribozymes carried out aminoacylation reactions in the earliest stages of the evolution of the translation apparatus. Indeed, several aminoacyl-RNA synthase ribozymes have been identified through *in vitro* selection, which use a variety of activated amino acid substrates[197,199,200,203] (Figure 3.3). These aminoacylating ribozymes show a range of specificities for the substrate side chain. For example, selection using a phenylalanine adenylate substrate **1** produced ribozymes that showed little discrimination (i.e., promiscuous ribozymes) as well as ribozymes showing a strong preference for aromatic amino acids.[198] Although they are derived from the same selection, these ribozymes have quite different promiscuity profiles and indices (Figure 3.4, Table 3.1). Selection for aminoa-

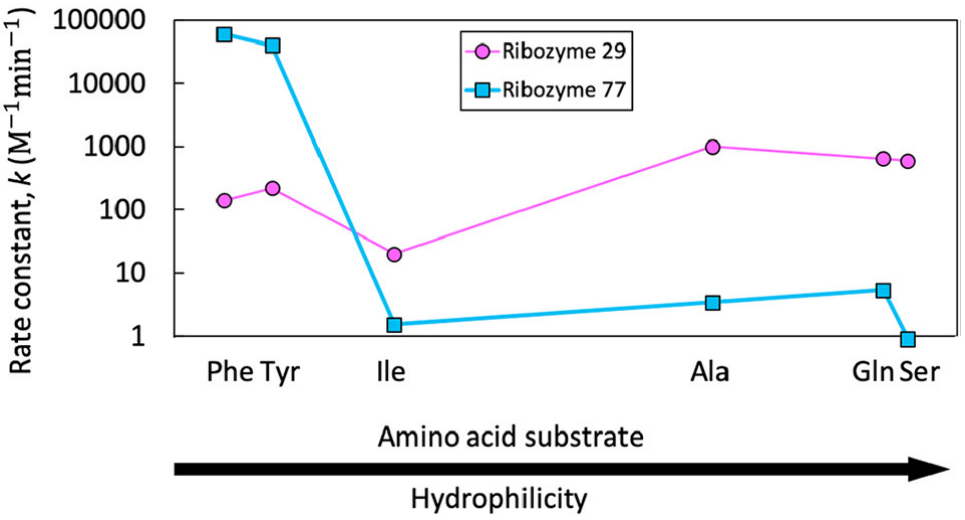

Figure 3.4: **Promiscuity profiles for two aminoacylation ribozymes.** Promiscuity profiles for Ribozyme 77 (blue squares) and Ribozyme 29 (pink circles) show catalytic rates for each tested amino acid substrate,[198] ordered by hydrophilicity as defined by Hopp and Woods[206,207] (Phe = -2.5, Tyr = -2.3, Ile = -1.8, Ala = -0.5, Gln = 0.2, and Ser = 0.3). Also see Table 3.1.

cylation with coenzyme A (CoA) thioester **2** produced ribozymes that could function with other CoA thioesters, but required the presence of a free α-amino group.[199] None of these ribozymes match the specificity of the aminoacyl-tRNA synthetase enzymes found in modern biochemistry. This discrepancy cannot be the result of a tradeoff between activity and specificity, since the ribozymes are generally much less efficient ($\sim$1000-fold) than the corresponding enzymes.[204,205] Instead, the general finding of promiscuity and the variation of specificities found among these ribozymes are consistent with the understanding that newly evolved sequences are not necessarily specific if they have not been selected for specificity.

The aminoacylating ribozymes discussed above have also been observed to catalyze reactions using alternative nucleophilic substrates to generate amide bonds in addition to esterification. In particular, a minimized, 29 nucleotide version of an aminoacylating ribozyme that utilizes Phe-AMP was found to catalyze successive reactions: aminoacylation of the RNA and the subsequent amide bond formation to generate a conjugated

Ribozyme	Substrate side chain	CID	k ($M^{-1}min^{-1}$)	I	J
77	Phenylalanine	6140	60000	0.376	0.439
	Tyrosine	6057	40000		
	Isoleucine	6306	1.5		
	Alanine	5950	3.4		
	Glutamine	5961	5.3		
	Serine	5951	0.9		
29	Phenylalanine	6140	140	0.810	0.807
	Tyrosine	6057	220		
	Isoleucine	6306	20		
	Alanine	5950	1000		
	Glutamine	5961	650		
	Serine	5951	600		

Table 3.1: **Promiscuity indices calculated for two aminoacylation ribozymes (77 and 29).** CIDs (PubChem Compound Identifier) for amino acids were used to determine similarities for calculation of J. The substrates used are aminoacyl adenylates with the side chain indicated. Both the unweighted (I) and weighted (J) promiscuity indices were calculated from the rate constants (k) shown. Rate constants are from Illangasekare, et al.[198]

peptide.[208,209] The rate of peptide formation was approximately 13-fold less than that for aminoacylation, but this difference could be tuned. Extending the 3' tail of the RNA by three nucleotides resulted in a three-fold reduction in the rate of aminoacylation and a 2-fold increase in the rate of peptide formation, presumably by increasing the flexibility around the active site.

The potential promiscuity of aminoacylating ribozymes is highlighted by the 'flexizymes' developed by Suga and colleagues, so named for their flexibility in accommodating a variety of substrates. These ribozymes were generated over a series of selections with the ultimate goal of producing catalysts capable of charging tRNAs with a wide variety of both natural and non-natural substrates. The starting point of the selection was a library containing a 5' random region and a 3' tRNA. Ribozymes were selected for their ability to aminoacylate the 3' terminus of the conjugated tRNA. This first selection produced ribozymes with a high level of specificity to both the tRNA and phenylalanine

substrates.[203,210,211] To broaden the tRNA substrate range, further design and selection was performed with an alternative tRNA sequence, which resulted in ribozymes capable of accepting a variety of tRNAs.[212] These early flexizymes exhibited high affinity for the aromatic side chains. Using the ribozyme's affinity to the aromatic group to broaden the side chain specificity, the initial substrate **4** was redesigned to substrate **5**, which contains a 3,5-dinitrobenzyl ester as the leaving group in the aminoacylation reaction. The idea was that this leaving group could be kept constant, ensuring affinity to the ribozyme, while the side chain itself was varied. This substrate necessitated an altered reaction mechanism, but nevertheless, the strategy was successful, with further selection resulting in ribozymes capable of charging tRNAs without regard to amino acid side chain.[212] More recently, flexizymes have been used to charge tRNAs with various non-natural amino acids, including D-amino acids, β-amino acids, and α-hydroxy acids, and 3'-aminoacyl-NH-tRNA can also be charged.[213–216] Although the flexizyme does exhibit a minor degree of side chain specificity, yields with the non-natural analogs often rival those for the L-amino acids used in the initial selections. The additional substrates represent both promiscuous (non-native) as well as native activities. Overall, the flexizyme demonstrates the surprisingly broad substrate specificity that can be evolved and designed when substrate generality is a desired goal.

3.4.2 Conditional promiscuity: the hammerhead ribozyme

Due to their historical importance in the discovery of ribozymes, much is known about the self-cleaving ribozymes, which function through general acid-base catalysis. The ribozyme fold brings the reactant nucleotides to the vicinity of the cleavage site, with the catalytic strand acting as the general base or acid to activate the nucleophile or stabilize the leaving group, respectively (Figure 3.5). Many of these ribozymes can also

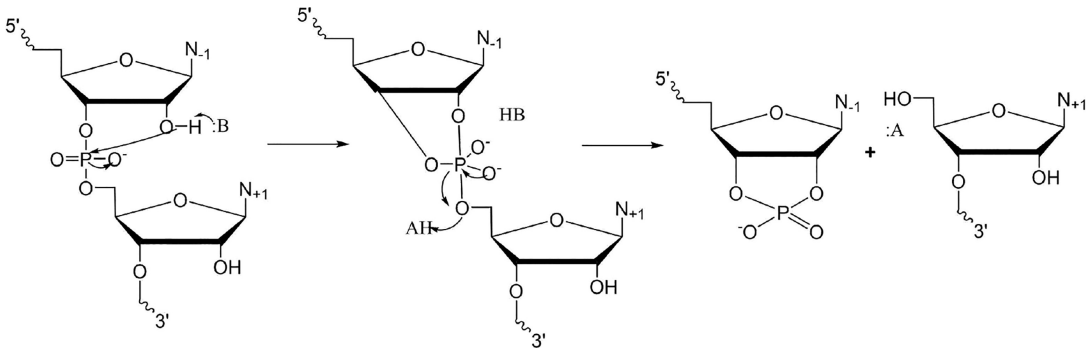

Figure 3.5: **Proposed mechanism of RNA self-scission by general acid–base catalysis.** A general base promotes deprotonation of the 2'-hydroxyl of the nucleophile, initiating formation of the cyclic intermediate. A general acid stabilizes the 5'-hydroxyl leaving group, allowing resolution of the intermediate to generate the cleavage products.

catalyze the same transesterification reaction in reverse, using nucleophilic attack from a 5'-hydroxyl to ligate two substrate strands,[217] which represents a possible case of catalytic promiscuity, a phenomenon discussed in Section 3.4.5 in the context of a different ribozyme. This section focuses on the substrate promiscuity of a self-cleaving ribozyme and how it arises. Although these ribozymes are *cis*-acting *in vivo*, they can be engineered to accept oligonucleotide substrates *in trans* with multiple turnover. While there are numerous self-cleaving ribozymes, discussion here is confined to the case of the hammerhead ribozyme, a naturally occurring ribozyme found in plant viroid transcripts,[218] for which the specificity of *trans*-acting variants has been extensively investigated.

The *trans*-acting hammerhead ribozyme can be engineered from the *cis*-acting ancestor by removing a nucleotide loop of one helical arm, thereby creating a cleavable substrate strand and a catalytic strand.[219–223] In such constructs, the catalytic strand can bind and cleave substrate strands with multiple turnover. In particular, separation of stem I from stem III of the ribozyme (the I/III construct) is most widely studied[224–226] since this construction places most of the conserved nucleotides in the catalytic strand (Figure 3.6). This allows substrate specificity in the substrate strand to be probed.

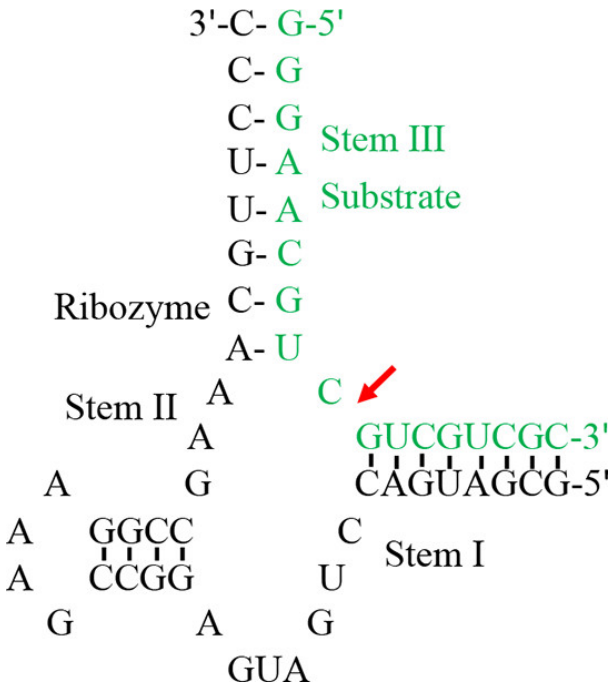

Figure 3.6: **Structure of a *trans*-acting I/III hammerhead ribozyme HH16.** The catalytic strand is shown in black, and substrate is shown in green.[227] The red arrow indicates the cleavage site.

There are two main expectations for sequence specificity of the substrate in a *trans*-acting construct. First, residues critical for the catalytic mechanism are expected to be relatively intolerant to mutations, which would primarily affect k_{cat}.[228] Second, aside from critical residues, promiscuity for the substrate is expected to be determined by binding interactions (K_m) between enzyme and substrate, namely base-pairing, which can lead to large variation in dissociation rates among different substrates. In the HH16 ribozyme (Figure 3.6), substrate affinity, which was dominated by stem III, was very high, implying a low dissociation rate, such that truncation from the 3' end, down to a 2-nucleotide version of stem I, had little effect on the overall rate of cleavage.[227] Specificity in either stem I or III of the substrate was therefore only observed when stem III was destabilized to give a dissociation rate that was on par with or slower than the overall cleavage rate. Conversely, extending the recognition sequence reduced specificity,[229] in

keeping with the idea that, if substrate dissociation is slow relative to cleavage, mutations in the substrate are tolerated since the bound complex is sufficiently populated. In terms of the active site itself, the hammerhead ribozyme has limited substrate promiscuity; substitution of the reactive phosphate with thiophosphate greatly reduces k_{cat}.[230]

The example of the hammerhead ribozyme, particularly the sequence dependence of the substrate, illustrates, at a molecular level, the property of conditional substrate promiscuity, in which the apparent promiscuity depends on the environmental condition. Variants having longer binding regions or higher substrate affinity can tolerate weakening (or strengthening) of binding without much change in population of the bound state, and therefore are relatively insensitive to mutations and have high apparent promiscuity. On the other hand, variants that exist on the threshold of binding can display high specificity as they are sensitive to small changes in dissociation rate. Thus, exhibition of promiscuity depends on conditions such as substrate concentration, pH, ionic strength, and temperature. Conditional promiscuity can be the basis for cryptic genetic variation, in which an altered phenotype is uncovered in new environments. Thus, it is likely to be underappreciated in the literature due to observational bias, since most experimental studies tend to focus on a small set of reaction or environmental conditions. This is an area ripe for future research given the likely importance of conditional promiscuity for evolutionary innovation.

3.4.3 Convergent Mechanism, Convergent Promiscuity: a Tale of Two 'Capping' Ribozymes

The influence of mechanism on promiscuity, illustrated by the hammerhead ribozyme, is exemplified in a comparison of two independently derived ribozymes that share a common mechanism. These ribozymes, isolated under different selection conditions, promote

the formation of a phosphate-phosphate anhydride bond between the terminal phosphate of a nucleotide and the 5'-α-phosphate of RNA. The final product is similar to the 5' cap found on eukaryotic mRNAs.

These two RNA capping ribozymes, the Iso6 and 6.17 ribozymes, were discovered in the Yarus and Unrau groups, respectively.[231,232] Interestingly, both ribozymes were isolated from different selections for which this capping reaction was not the desired function. Iso6 was recovered from a selection originally designed to identify ribozymes that could produce aminoacyl adenylates through reaction between amino acids and triphosphorylated RNA. Instead, pyrophosphate release was observed in the absence of amino acids, and the selection pool even developed labeling with PP_i. Selection for capping activity using UTP instead of PP_i quickly resulted in high activity in the pool and the identification of Iso6.[231] On the other hand, the 6.17 ribozyme derived from a selection initially designed to identify polymerase activity by incorporating labeled UMP into a primer annealed to a poly(A) template. The resultant ribozyme with the fastest kinetics, 6.17, instead was found to act on the 5' end of the RNA, forming a 5'-5' cap.[232]

Iso6 and 6.17 display no apparent sequence similarities and are expected to adopt different secondary structures, consistent with their unique origins. Despite these differences, the molecular mechanisms for these two ribozymes appear to be surprisingly similar. Both ribozymes are predicted to have helices that terminate at the site of capping, with the terminal 5' nucleotide retaining some flexibility; this position is unpaired in Iso6 and requires wobble pairing in 6.17. Both ribozymes also display increased activity at lower pH and require divalent cations for activity, although Iso6 prefers Ca^{2+} while 6.17 tolerates Mg^{2+}, Mn^{2+} and Ca^{2+}. The ribozymes even possess similar substrate binding affinities.[232,233] Finally, both ribozymes appear to have minimal substrate requirements. The identities of the sugar and base have little impact on activity, despite possible hydrogen bonding interactions with these moieties. However, decreasing the

length of the phosphate chain results in a large decrease in substrate binding,[232,234,235] indicating that the phosphate itself is responsible for most substrate interactions.

Thus, these ribozymes suggest a common molecular mechanism for RNA capping that permits a high degree of substrate promiscuity, provided a small number of key features is present. The fact that two independently evolved, structurally dissimilar ribozymes have the same requirements supports the idea that substrate promiscuity is determined by mechanism. In this case, evolutionary convergence on the same mechanism resulted in convergence to similar promiscuity as well.[236]

3.4.4 Relying on Promiscuity: Searching for an RNA Replicase

For those interested in the origin of life, one of the most sought-after *de novo* ribozyme functions is catalysis of template-directed RNA polymerization (an 'RNA replicase'), which is thought to be important, if not essential, to a self-replicating RNA system. One of the major avenues for this search has relied heavily on the promiscuity of newly evolved ribozymes. The first ribozymes developed by use of *in vitro* selection from a large pool of random sequences were RNA ligase ribozymes, including the 'class I ligase' (Figure 3.7). The class I ligase was selected to catalyze the ligation between a 5'-triphosphate on the ribozyme and a 3'-hydroxyl on an RNA oligonucleotide substrate, which caused the ribozyme to tag itself with a sequence on the substrate that was necessary for purification and amplification.[1,237] The 3',5'-phosphodiester bond formed during ligation is identical to that formed during template-directed polymerization. Modification of the original class I ligase to bind a primer-template complex generated a ribozyme that possessed some polymerization activity, being able to extend a primer through the incorporation of mononucleotide triphosphates.[238,239] This reaction occurred with 92% fidelity, though activity decreased with successive nucleotide additions, topping out at six nucleotides

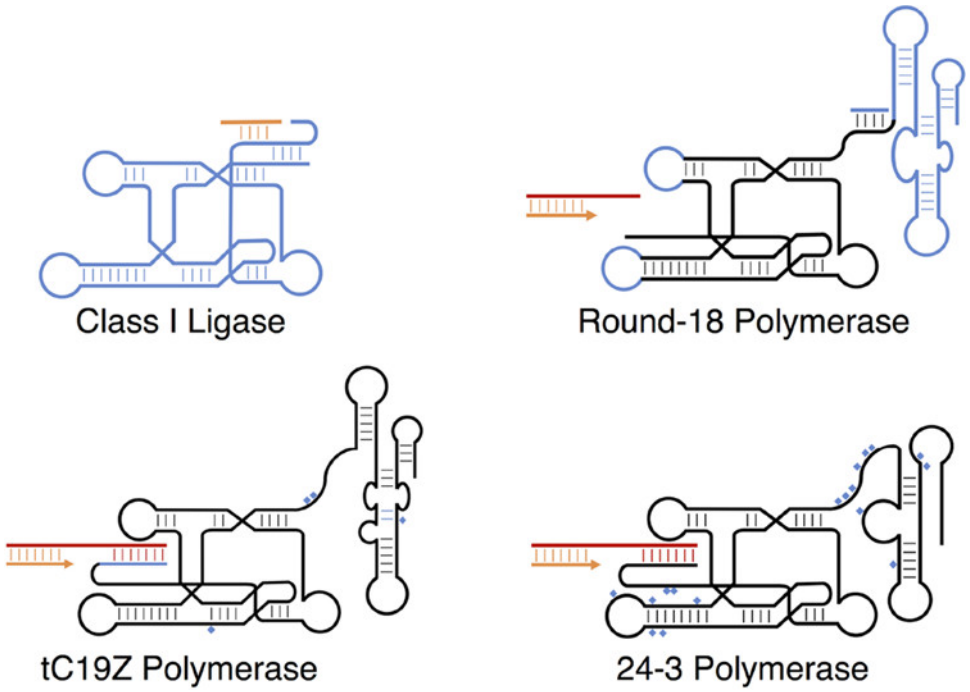

Figure 3.7: **The class I ligase ribozyme and its descendants.** Structures for the class I ligase;[238,240] the round-18 polymerase introducing the new 3' accessory domain (blue), which is known to interact with the loop depicted on the lower right;[241,242] the tC19Z polymerase, introducing a new 5' accessory domain;[243] and the 24-3 polymerase.[244] Blue regions denote new additions to the ribozyme, with point mutations marked by blue diamonds. Primer and template oligonucleotides are shown in orange and red.

added after a six-day incubation. Nevertheless, this initial finding signaled that catalytic promiscuity of the class I ligase could potentially lead to an RNA replicase.

Subsequent work with the class I ligase and its derivatives aimed to increase its processivity, fidelity, and template generality. Important progress was made through attachment of an accessory domain to the 3' end of the ribozyme, which was selected from a 76-nt random sequence with the idea that this domain could facilitate interaction of the ribozyme with the primer-template complex.[241] Polymerization activity was selected for through incorporation of tagged nucleotides opposite an attached primer. To increase the sequence generality of polymerase activity, shorter templates were used to reduce hybridization with the ribozyme, and different primer-template sequences and lengths

were used in different rounds of selection. One ribozyme, isolate 10.2, was found to function without attachment to the primer, and without recognition of a specific sequence. This feature was conferred by the new accessory domain, which increased binding of the primer-template complex, but polymerization activity itself occurred with minimal change to the ligase domain. Mutagenized versions of the 10.2 ribozyme were further selected for function on longer single-stranded templates and in the presence of higher concentrations of untagged nucleotides to improve fidelity. After eight more rounds of selection, the best resultant ribozyme, termed the round-18 polymerase (Figure 3.7), functioned much better with longer templates, and allowed for the extension of up to 14 nucleotides.

While catalytic promiscuity was key to the discovery of an RNA polymerase ribozyme, substrate promiscuity at a given template base is highly undesirable. That is, fidelity is important for an RNA replicase, because error rates represent a serious limit in the transmission of information.[245] The round-18 polymerase copied templates with a per-base fidelity of 96.7%, which corresponds to relatively low promiscuity values (0.01-0.22; Table 3.2). One of the major determinants of this fidelity is misincorporation resulting from G:U wobble pairs, which is reflected by their higher promiscuity values compared to A and C (Table 3.2). While extension across a template A or C resulted in the correct addition (U or G, respectively) in over 99% of cases, G templated with an incorporation fidelity of 95.7%, and U templated with a fidelity of 92.1%, with the vast majority of mismatches resulting in a G:U mispair. This type of mispairing also appears to be the major limitation on the fidelity of non-enzymatic replication, and may be an echo of the thermodynamic limit on specificity.[144,246]

Despite the 3' accessory domain, a major limitation of the round-18 polymerase continued to be low binding affinity for the primer-template complex, which was the primary contributor to the low processivity of the class I polymerase. The affinity also had a high

Substrate	CID	k_A	k_C	k_G	k_U
ATP	5957	0.30	0.057	0.023	5.3
CTP	6176	0.02	0.008	5.4	0.0002
GTP	135398633	0.02	41	0.003	0.23
UTP	6133	87	0.004	0.46	0.001
	Fidelity	0.991	0.9996	0.957	0.921
	$\boldsymbol{I}$	0.020	0.010	0.219	0.126
	$\boldsymbol{J}$	0.020	0.010	0.220	0.125

Table 3.2: **Promiscuity of an RNA polymerase ribozyme.** Fidelity, promiscuity index (I), and weighted promiscuity index (J) for the round-18 polymerase. Rate constants (k_N, for template N = A, C, G, U; $M^{-1}min^{-1}$) are from Johnston et al.[241]

degree of variability with regard to primer-template sequence, suggesting that further reduction of sequence specificity was still needed.[247] Further progress was achieved by selecting directly for activity *in trans* using a water-in-oil emulsion. The first product of this method was the B6.61 polymerase, which was capable of generating sequences 20 nucleotides long.[242,248] B6.61 showed a much faster polymerization rate than its predecessor, with an extension rate over 75-times faster for longer sequences. While there was no significant improvement in binding to the primer-template, this rate increase was accompanied by increased fidelity, including a minimization of G:U wobble insertions. As with the aminoacylation ribozymes (Table 3.1), this trend is a counterexample to the idea of a general tradeoff between activity and specificity.

A substantial improvement to processivity came using a similar compartmentalization technique with the addition of a 5' random region with the aim of improving interactions between the ribozyme and the primer-template complex. This yielded a 5' accessory domain that forms stabilizing interactions with downstream portions of the template, thus increasing binding of the ribozyme to the template. Randomization and selection of the template sequence strengthened these interactions. With the optimized template, the new ribozyme, named tC19, ultimately yielded up to 95 nucleotide extensions with a per-base fidelity of 97.3%. However, the new interactions were largely intermolecular base-

pairing, such that activity was strongly dependent on sequence. Selection on different templates identified four point mutations which, when introduced into tC19 to make tC19Z (Figure 3.7), improved the sequence generality. These new mutations further increased the measured per-base fidelity to 99.1%, largely due to a decrease in G insertion across template U. The tC19Z polymerase was shown to be capable of transcribing a functional 24 nt variant of the hammerhead ribozyme.[243]

Consideration of the promiscuity of the RNA polymerase ribozyme raises an interesting irony: while substrate promiscuity of the incoming monomer across a given template base is undesirable because it leads to copying errors, substrate promiscuity with respect to the template itself is highly desirable to obtain a ribozyme capable of copying many different, and ideally any, sequences. Sequences of particular concern are those with a high degree of structure that would need to be locally melted for ribozyme access, including sequences that comprise the ribozyme itself. Recent selections based on the RNA polymerase ribozyme focus on improving its sequence generality. One such study selected for the polymerase's ability to synthesize complex folded RNA molecules, with selection tied to the creation of two functional aptamers, imposing pressure for sequence generality and high fidelity.[244] The most active selected ribozyme, 24-3 (Figure 3.7), showed a ∼100-fold increased incorporation rate through structured sequences compared to the parent ribozyme. Likely as a result of selection for functional molecules instead of sequence fidelity, the 24-3 ribozyme displayed a higher error rate than its predecessors, in particular an increased tolerance for G-U wobble pairing. Despite this limitation, 24-3 was able to synthesize functional RNAs up to 76 nt long and could perform exponential amplification of an RNA template.

A different approach to overcoming the substrate generality problem takes advantage of plasticity, which occurs when non-native functions can be found through a relatively small number of mutations. In this case, it was hypothesized that copying via ligation of

oligonucleotides could improve copying through structured sequences, since base-pairing to the oligonucleotide would mitigate some of the free energy cost of melting the template. Knowing that the RNA polymerase ribozyme was originally derived from the promiscuous activity of an RNA ligase, Attwater et al. engineered and evolved an ancestor of the tC19Z ribozyme to copy templates using trinucleotide triphosphates instead of NTPs.[249] The triplet oligonucleotides use strand invasion to unfold structured RNA sequences for improved copying. The atavistic ribozyme $t5^{+1}$ displayed reduced fidelity compared to its NTP-using counterpart, but selection for fidelity yielded an improved variant able to synthesize its own catalytic subunit. Interestingly, the $t5^{+1}$ ribozyme consists of a heterodimer of the catalytic subunit and an RNA 'cofactor' that assists interaction with the primer-template complex. Both subunits are descended from the same ancestral pool, illustrating how specialized descendants originated from distinct domains of the ancestor.

An ingenious orthogonal strategy to overcome the problem of sequence generality was developed by Joyce and coworkers, who reasoned that base-pairing between ribozyme and template was the major contributor to the energetics determining template specificity. Base-pairing is essentially absent between D-RNA and L-RNA sequences,[250,251] and thus a D-ribozyme is expected to have little base-pairing interaction with L-substrates. Selection for ligase activity indeed discovered D-ribozymes that could ligate L-RNA oligonucleotides.[252] As expected, the non-natural, mirror-image L-ribozyme could perform the complementary reaction using D-RNA substrates and template. Furthermore, as with the class I ligase, these cross-chiral ligases also possessed polymerization activity. Unlike non-enzymatic templated RNA replication,[253] these ribozymes displayed very little chiral inhibition, showing a high specificity for substrates of the desired chirality. A cross-chiral ligase was efficient enough to produce its mirror image enantiomer, which could then produce the original enantiomer. The cross-chiral ribozymes were not entirely sequence-general, as some substrates, such as those with 3'-terminal C or G residues,

were more efficient than others. Nevertheless, given the precedent of the evolutionary strides demonstrated by the promiscuous class I ligase, the cross-chiral ligases represent an intriguing starting point for further development.

Polymerase (and ligase) ribozymes present a unique challenge in simultaneously requiring broad template accommodation and strict fidelity. Although this work was not undertaken for the purpose of studying promiscuity in ribozymes, the advances made with the class I ligase, spanning more than two decades of work by multiple groups, rely heavily on the promiscuity and plasticity of the ribozyme. Since this lineage of RNA polymerase ribozymes has only been selected on RNA substrates, true promiscuity can be clearly identified if the ribozymes accept different nucleic acids. One ribozyme displays some activity for incorporation of non-natural sugars and nucleobases, although it often stalls if modified nucleotides are present at specific positions.[254] Unlike other ribozymes in its lineage, the 24-3 ribozyme, perhaps as a consequence of its selection for tolerance of different RNA aptamer templates, was observed to polymerize DNA on an RNA template (i.e., reverse transcription), permitting extension by up to 32 deoxyribonucleotides.[255] A later generation of this ribozyme, 38-6, shows remarkable promiscuity, with activity on templates or nucleotides composed of multiple combinations of RNA, DNA, threose nucleic acid (TNA), or arabinose nucleic acids (ANA), though with reduced activity compared to an RNA-only system.[256] 38-6 also performs DNA-templated RNA synthesis and RNA-templated DNA synthesis more effectively than synthesis involving TNA or ANA, likely due to the structural similarities. These results further demonstrate the promiscuity of this ribozyme lineage. While the class I ligase and its descendants constitute a fascinating case study, it is unknown whether other ribozymes could exhibit similar versatility.

3.4.5 Catalytic Promiscuity: The Nucleotide Synthase Ribozyme

While substrate promiscuity appears to be commonly found among ribozymes,[232,257,258] one may ask whether true catalytic promiscuity is also observed. Indeed, an interesting case was found in the pR1 nucleotide synthase.[259] Selected to catalyze a reaction between ribose 5-phosphate (PR) and 6-thioguanosine (^{6S}Gua), this ribozyme was found to also be capable of catalyzing the reaction between ^{6S}Gua and 5-phosphoribosyl 1-pyrophospate (PRPP), an intermediate in the biological synthesis of nucleotides. These two reactions appear to have distinct reaction mechanisms and resultant products, depending on the substrate provided. Reaction with PRPP generates a glycosidic bond, resulting in the corresponding nucleotide, ^{6S}GMP. However, the reaction with PR appears to require acyclization of ribose, allowing ^{6S}Gua to react with the corresponding aldehyde and generate a Schiff base (Figure 3.8). Each reaction has a unique dependence on magnesium ion concentration, supporting the existence of two different mechanisms. Interestingly, ribozymes selected for reactivity with PRPP instead of PR did not exhibit analogous activity on PR. Thus, not all ribozymes with the same function possess catalytic promiscuity. Despite the catalytic promiscuity of pR1, the ribozyme still displays a high degree of substrate specificity toward ^{6S}Gua, as analogous sulfur-containing purines are not recognized, in contrast to purine synthase ribozymes selected independently.[260] The pR1 ribozyme demonstrates that catalytic promiscuity may differ in important ways from substrate promiscuity. While substrate promiscuity might be readily evolved through a relaxed binding mode, catalytic promiscuity requires a new reaction mechanism whose spontaneous emergence might be relatively unusual.

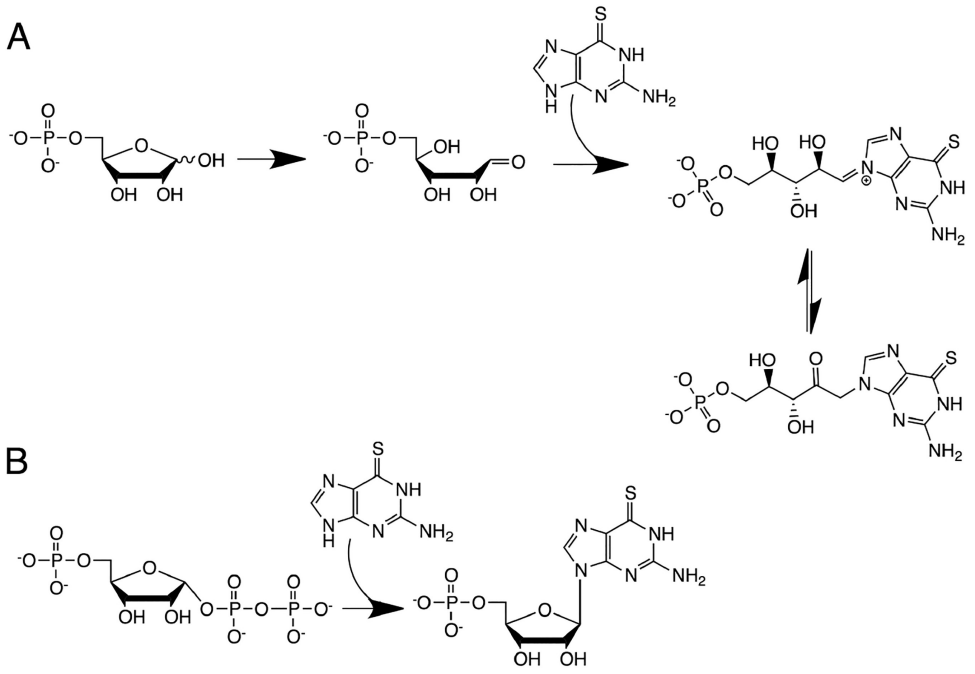

Figure 3.8: **Reactions catalyzed by the pR1 nucleotide synthase ribozyme.** (A) Given a ribose 5-phosphate substrate, the acyclic form of ribose is stabilized and 6-thioguanosine (^{6S}Gua) reacts to form a Schiff base, which can then undergo an Amadori rearrangement. (B) Reaction with 5-phosphoribosyl 1-pyrophospate produces the desired nucleotide, ^{6S}GMP. Adapted from Lau and Unrau.[259]

3.4.6 A Highly Promiscuous Ribozyme: The Ribosome

The ribosome is a ribozyme that translates genetic information of mRNA into protein sequences and is conserved across all the domains of life. The ribosome core consists of catalytic RNA, but farther from the catalytic center both proteins and RNA are found.[261] In eukaryotes, four ribosomal RNAs (rRNAs) associate with about 70 proteins, while in *E. coli*, the ribosome consists of three rRNAs and 52 proteins.[262,263] Because the ribosome is a ribozyme and conserved across all domains, it is presumed to have existed in the last universal common ancestor (LUCA), and is also taken as circumstantial evidence of the RNA World.[264] Protein translation necessitates high fidelity, with error rates of the overall process on the order of 10^{-4} per codon. Fidelity is primarily maintained through factors other than the ribosome, such as aminoacyl-tRNA synthetase editing and EF-

Tu binding.[265,266] While the peptidyl transferase center of the ribosome provides some steric selectivity (e.g., preferring L- rather than D-amino acids[267]), the ribosome itself is a surprisingly promiscuous molecule overall, permitting a wide assortment of substrates so long as there is a correct codon-anticodon match.

The ribosome accepts two aminoacyl-tRNAs at a time and catalyzes the formation of a peptide bond between the amino acids, releasing an uncharged tRNA and retaining a peptidyl-tRNA.[263] The ribosome must accommodate a large variety of substrates: there are 20 canonical amino acids that can be associated with 50 or more different tRNAs, depending on the species.[268] Even if one only considers two canonical amino acids coming together to be joined by a peptide bond, there are 400 possible substrate permutations that the ribosome must accommodate and catalyze. This level of multispecificity is essential for the production of all extant proteins in the organism. In addition to accommodating different amino acids and peptides in the active site, the ribosome must also accommodate different peptides in the exit tunnel. Interestingly, the exit tunnel is lined primarily by RNA and lacks significant patches of hydrophobicity, creating a 'nonstick' character that allows peptides through regardless of sequence.[261]

The ribosome is similarly multispecific with respect to the mRNA templates, on which there are minimal sequence restrictions. Following initiation, which does involve sequence-specific interactions in some organisms, ribosome binding to mRNA is primarily facilitated through interactions with the mRNA backbone.[269] However, the ribosome does display some slight substrate preferences. Early research on the ribosome, for example, discovered roughly two-fold higher reactivity with leucine than phenylalanine.[270] Additionally, ribosomes display codon preferences that can alter the elongation rate,[271] a property which is used for regulation of gene expression. Still, the degree to which ribosomes are capable of utilizing a wide variety of substrates, including many non-canonical amino acids,[272] representing promiscuous activity, is truly striking.

As with the RNA polymerase ribozymes, the substrate promiscuity of the ribosome must co-exist with a requirement for high fidelity of information transfer. The promiscuity of the ribosome is tolerated by the cell in part because translation fidelity is handled during aminoacylation of tRNAs, including proofreading processes.[273] In the ribosome, cognate and non-cognate tRNAs can be distinguished through minor differences in base-pairing to mRNA. Recognition of the cognate tRNA leads to a structural change that is identified by elongation factor proteins which permit translation to proceed.[274] Another restriction imposed on the incoming tRNA is the 3' terminal CCA sequence, which forms specific interactions with the ribosome.[261,269] This CCA sequence is required, and occasionally sufficient, for peptidyl transfer to occur.[270,275] A small number of important interactions between tRNAs and the ribosome provide high fidelity of translation while permitting minimal restrictions on the mRNA or protein sequences.

The innate promiscuity of the ribosome is occasionally exploited by nature. One such example is puromycin, an antibiotic produced by the bacteria *Streptomyces alboniger*. Puromycin is an aminonucleoside, containing nucleoside and amino acid analogs, linked through an amide bond instead of the conventional ester. This structure mimics the 3' terminus of a charged tRNA, which allows it to enter the ribosome and be irreversibly incorporated into the nascent polypeptide, terminating translation.[276,277] The efficacy of this molecule suggests that evolutionary escape from this promiscuous activity has been difficult despite the selective pressure engendered by the antibiotic.

Synthetic biologists have also taken advantage of the substrate promiscuity of the ribosome, fundamentally altering the genetic code itself. tRNAs recognizing the amber stop codon can be charged with non-canonical amino acids. Since the ribosome is essentially agnostic with respect to the side chain of the incoming monomer, the amber codon is translated into the new amino acid.[278,279] Amino acids with a remarkably diverse set of unnatural functional groups have been successfully incorporated by this method, includ-

ing alkanes, polybenzenes, sugars and phosphate-containing species.[280,281] The ribosome can even catalyze the formation of ester bonds, yielding mRNA-encoded polyesters, without mutation in the ribosome itself.[213] Further evolution can push this versatility further, as seen with the ribosome variants ribo-Q1 and ribo-X, which translate quadruplet codons and thus introduce many 'blank' codons to the genetic code.[282–284] Although it was postulated early on that the genetic code might be a 'frozen accident',[130] it now seems clear that the code itself has been the subject of evolution, as evidenced by the different version of the code found in mitochondria[285] as well as statistical analyses suggesting that the code has evolved to minimize the biophysical impact of mutations.[286] The evolvability and malleability of the genetic code attests to the remarkable combination of substrate promiscuity and informational fidelity in the ribosome.

3.5 Primordial Ribozymes: More Promiscuous?

We now return to a question posed at the beginning of this review: would primordial ribozymes be particularly promiscuous? There are two reasons why one might hypothesize this. First, given the importance of promiscuity for evolutionary innovation, one may suppose that primordial ribozymes might have been more promiscuous than highly evolved enzymes due to evolutionary pressure for greater specificity. Second, given the wider chemical diversity of functional groups available to proteins, one may suppose that proteins will have both superior specificity and activity, in general, compared to ribozymes, due to their ability to engage in more types of interactions. Although there is insufficient data in the literature to answer these questions definitively, here we consider two comparisons that bear on these issues. First, we consider whether newly evolved ribozymes are more promiscuous than highly evolved ribozymes. Second, we ask whether ribozymes are more promiscuous than proteins by examining a head-to-head comparison

of a ribozyme and a protein enzyme, both of which were evolved *de novo*.

3.5.1 Newly Evolved vs. Highly Evolved Ribozymes

An interesting comparison can be made between the promiscuity of ribozymes from *in vitro* selection, which have very short evolutionary histories, to highly evolved ribozymes, in particular, the ribosome. The first ribozymes to catalyze amide bond formation were initially selected for a different activity, to catalyze the transfer of an aminoacyl group from the 3'-hydroxyl of a short tRNA mimic to the 5'-hydroxyl of the ribozyme.[287] However, like the ribosome, one of the selected ribozymes was able to use an alternative nucleophilic substrate. When the 5'-hydroxyl was substituted with an amino group, amide bond formation was observed at a similar rate. In both cases, the ribozyme accelerated the respective non-catalyzed reaction by over 1000-fold. Later, ribozymes were selected to perform peptide bond formation by linking a phenylalanine to the 5' end of the RNA and selecting for the ability to attach a biotinylated methionine from a 3' acylated AMP substrate.[288] The best ribozymes from this selection displayed a rate enhancement of $\sim 10^6$. This reaction was inhibited by the presence of AMP, but not other nucleotides or methionine, suggesting that the ribozyme functions primarily through specific interactions with AMP. Consistent with this, activity was also observed with leucine, phenylalanine, and lysine substrates, with methionine and leucine being the best peptidyl donors.

These peptide synthase ribozymes, much like the flexizyme, illustrate that if there are sufficient interactions with other parts of the small molecule substrate, the amino acid side chain may be quite variable. In addition, the fact that one of the ribozymes is capable of both ester and amide bond formation, much like the ribosome, further corroborates its promiscuity of function. (In this case, these non-native activities represent true

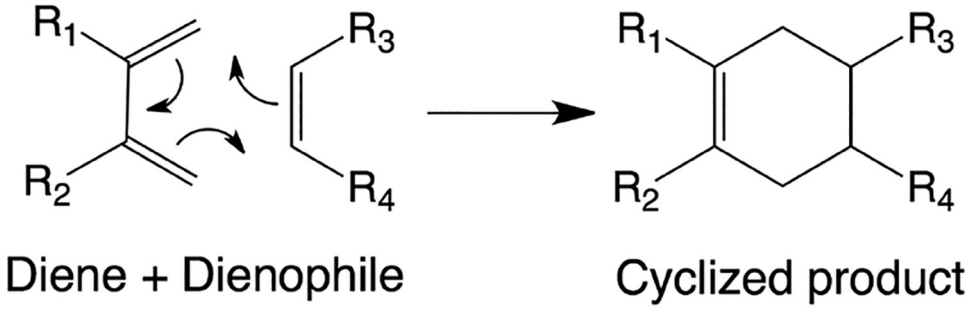

Figure 3.9: **Diels-Alder cycloaddition.** A concerted reaction between a conjugated diene and an alkene (dienophile) results in the formation of a cyclized product.[289] The dienophile substituents shown (R_3 and R_4) are added to the same face of the cyclohexane ring.

promiscuity since the complete environmental history of the ribozyme is known.) However, while it may be possible to evolve increased promiscuity in these ribozymes, it seems a hard task indeed to match or exceed the promiscuity of the ribosome. This comparison at least suggests that newly evolved ribozymes are not necessarily more promiscuous than highly evolved ones. Instead, specificity or promiscuity itself may be a selectable trait, and natural selection may favor either greater or lesser promiscuity.

3.5.2 *De Novo* Ribozyme vs *de Novo* Protein Enzyme: The Diels-Alderases

Are protein enzymes superior to ribozymes, such that ribozymes emerging in the RNA world would be worse than their protein counterparts? While it seems clear that proteins have greater activity in general, nearly all protein enzymes have much longer evolutionary histories compared to ribozymes, most of which have been evolved *in vitro*. To avoid this confounding factor, one may compare a *de novo* ribozyme with a *de novo* protein enzyme. Such a comparison can be made with the Diels-Alderase RNA and protein enzymes, which both catalyze a reaction (Figure 3.9) previously not known to occur in biology.

Interestingly, the first biochemical catalyst discovered for the Diels-Alder reaction was a ribozyme, not a protein. While most ribozyme reactions involve RNA or amino acid modifications and often involve base-pairing interactions, an early discovery that demonstrated the catalytic versatility of ribozymes was carbon-carbon bond formation by Diels-Alderase ribozymes.[290,291] The specificity of one such ribozyme was extensively characterized by the Jäschke laboratory through testing a series of potential substrates.[292,293] The initial experiments selected for cycloaddition of a biotin-maleimide to anthracene, which was conjugated to the RNA via a polyethylene glycol linker. The ribozyme produced from this selection could catalyze this reaction on free substrate with a high degree of enantioselectivity. Additionally, a synthesized mirror image of this ribozyme composed of L-nucleotides produced the opposite enantiomer. This enantioselectivity was the result of a 'tail' group on the anthracene substrate (e.g. the PEG linker), which restricted the molecule's orientation in the binding pocket. Important structural features of both substrates include: the diene must contain three linearly annellated rings, the dienophile must be a five-membered maleimide ring with a hydrophobic tail, and both substrates must be arranged in a stacked, coplanar manner.[293] These results present one of the most rigorous characterizations of ribozyme specificity on a non-nucleotide substrate.

One decade later, a Diels-Alderase protein enzyme was developed by the Baker Lab.[294] This enzyme was created *de novo* using computational design and site-directed mutagenesis to catalyze the reaction between 4-carboxybenzyl trans-1,3-butadiene-1-carbamate and N,N-dimethylacrylamide. Like the ribozyme, this protein enzyme demonstrated a high level of product stereoselectivity (>97%). The best Diels-Alderase enzymes possessed higher catalytic activity than the Diels-Alderase ribozymes, but were still markedly slower than natural enzymes.

Although these catalysts were discovered through different means (*in vitro* selection vs. computational design), both were created in a laboratory setting independent of

	Diene		Dienophile	
DA Enzyme	***I***	***J***	***I***	***J***
RNA	0.620 (0.723)	0.536 (0.765)	0.732 (0.824)	0.665 (0.828)
Protein	-	-	0.723	0.765

Table 3.3: **Promiscuity indices for RNA and protein Diels-Alderase enzymes.** Unweighted (I) and weighted (J) promiscuity indices for RNA (ribozyme) and protein Diels-Alderase (DA) enzymes. Ribozyme promiscuity indices calculated from k_{cat} ($M^{-1}min^{-1}$) and k_{cat}/k_{uncat} (in parentheses) values reported for diene and dienophile substrates.[293] Protein enzyme promiscuity indices calculated from estimated rates (hr^{-1}) with dienophile substrates.[294]

natural evolutionary influences, and therefore they are an interesting test comparison to understand the promiscuity of *de novo* functions. While the reactions catalyzed by these molecules use different substrates, the promiscuity indices can be compared between them (Table 3.3). Note that different values are calculated for the diene and dienophile when possible, and that I and J varies depending on whether they are calculated from k_{cat} or from k_{cat}/k_{uncat} (catalytic enhancement). Despite these differences, in general, the promiscuity indices are not very different; all values for the dienophiles lie in the range of 0.66-0.83, with the values for the protein enzyme lying in the middle of this range. Therefore, comparing these two *de novo* catalysts, it does not appear that the ribozyme is more promiscuous than the protein enzyme.

It is of practical interest to note that the unweighted (I) and weighted (J) promiscuity indices (Tables 3.1 - 3.3) are often not very different from one another, as the difference between these values ranges from 0 to 0.1. This may reflect ruggedness in the promiscuity profile over the chemical space of the substrates. The motivation for creating a weighted index was to account for the expectation that chemically similar substrates would have similar activity. While this must be true to some extent, over the substrates that were tested and reported in these examples, the additional accounting did not alter the overall calculation by much.

3.6 Discussion

Ribozymes identified by *in vitro* selection or evolution represent an ideal model system for studying true promiscuity, because the selective pressures on these ribozymes are controlled by the experimenter and their entire evolutionary history is available for study. In addition, the promiscuity of ribozymes in particular is a fascinating question relating to the origin of living systems. An attractive, but untested, hypothesis is that the earliest ribozymes emerging from the prebiotic milieu of random polymers would be highly promiscuous, presenting a kernel of activity across many functions that could be optimized by evolution individually (e.g., after duplication events). Although a rigorous test of this hypothesis is currently lacking, we may consider how current knowledge informs this hypothesis of promiscuous ribozymes.

What are the likely properties of a ribozyme selected *de novo*, i.e., a primordial ribozyme? It is clear that the activity is likely to be low initially, simply because there are more sequences of low activity compared to high activity (i.e., the frequency of sequences is a decreasing function of activity),[60] leaving room for optimization of activity by natural selection. What about promiscuity? While it might seem intuitive that simple, low activity ribozymes would have high promiscuity, solid evidence for this so far in the literature is lacking. As discussed above, a *de novo* peptide synthase ribozyme is less promiscuous than its highly evolved counterpart (the ribosome). While there are some examples of ribozymes where *in vitro* evolution resulted in both improved activity and specificity (discussed above), it is not clear that there would be a positive correlation between activity and specificity in general. Indeed at least one mechanism (discussed in Section 3.3.2) has the opposite effect, causing a negative correlation, i.e., a tradeoff, between activity and specificity.

The intuition that there should be a positive correlation between activity and speci-

ficity is based on the general idea that increased molecular interactions give both increased activity and increased specificity. This seems to be reasonable, but in one rigorous study of RNA aptamers, activity was found to be uncorrelated to specificity.[295] It is even less clear that increased interactions should increase specificity in ribozymes, particularly *cis*-acting ribozymes, since the entire reaction pathway would be stabilized. How the ground state, transition state, and product would be affected in relative terms is not clear. Therefore it should not be assumed that the earliest emerging ribozymes were particularly promiscuous. Empirical data is required to resolve the relationship between activity and specificity of ribozymes.

The reason why the hypothesis of promiscuous primordial ribozymes is attractive, despite the current lack of evidence to support it, is that it solves an important problem in prebiotic evolution. If the first ribozyme to emerge by chance possesses the ability to catalyze many reactions, albeit at low activity, this ribozyme could serve as the ancestral catalyst to a suite of different reactions, rapidly forming a metabolic network of ribozymes. In evolutionary terms, a network of ribozymes might then arise from exaptation (or pre-adaptation) of a small number of ancestral ribozymes. However, it may be that promiscuity, rather than being an automatic property of a low-activity primordial ribozyme, should be considered as an evolvable or fortuitous trait itself, possibly uncorrelated to activity. In this case, the selective pressures on the RNA world would play an important role in shaping ribozyme evolvability.

Chapter 4

Materials and Methods

4.1 Permissions and Attributions

This chapter was the result of collaboration with Abe Pressman, Ziwei Liu, Celia Blanco, Ulrich F. Müller, Gerald F. Joyce, Robert Pascal, Yuning Shen, and Irene A. Chen. Portions of this chapter have previously appeared in the Journal of the American Chemical Society.[200] It is reproduced here with the permission of ACS, to which further permissions related to the material excerpted should be directed: `https://pubs.acs.org/doi/10.1021/jacs.8b13298`.

4.2 Overview

This chapter describes materials and methods used to obtain the results presented in Chapters 5 and 6.

Table 4.1: **^{1}H NMR spectra for synthesized compounds.**

Compound	**^{1}H NMR Spectrum**
N-tert-Butoxycarbonyl-*O*-methyl-tyrosine methyl ester (Boc-Tyr(Me)-OMe)	^{1}H NMR (400 MHz, DMSO-d_6) δ 7.14 (d, J = 8.5 Hz, 2H), 6.84 (d, J = 8.6 Hz, 2H), 4.11 (ddd, J = 10.0, 8.1, 5.3 Hz, 1H), 3.72 (s, 3H), 3.61 (s, 3H), 2.99 – 2.72 (m, 2H), 1.33 (s, 9H)
Biotinylated *O*-Methyl-Tyrosine methyl ester (Biotin-Tyr(Me)-OMe)	^{1}H NMR (400 MHz, $CDCl_3$) δ 7.14 (d, J = 8.1 Hz, 2H), 6.88 (d, J = 7.9 Hz, 2H), 4.23 (t, J = 6.4 Hz, 1H), 3.80 (s, 3H), 3.78 (s, 3H), 3.24 (qd, J = 14.6, 6.2 Hz, 2H)
Biotinylated *O*-Methyl-Tyrosine (Biotin-Tyr(Me)-OH)	^{1}H NMR (300 MHz, DMSO-d_6) δ 8.05 (d, J = 8.1 Hz, 1H), 7.13 (d, J = 8.5 Hz, 2H), 6.83 (d, J = 8.5 Hz, 2H), 6.37 (d, J = 9.9 Hz, 2H), 4.45 – 4.24 (m, 2H), 4.19 – 4.06 (m, 1H), 3.71 (s, 3H), 3.09 – 2.90 (m, 2H), 2.80 (ddd, J = 20.9, 13.2, 7.3 Hz, 2H), 2.05 (t, J = 7.1 Hz, 2H), 1.69 – 1.08 (m, 6H)

Biotinyl-tryptophan methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ 10.82 (s, 1H), 8.19 (d, J = 7.6 Hz, 1H), 7.47 (d, J = 7.9 Hz, 1H), 7.31 (d, J = 8.1 Hz, 1H), 7.11 (d, J = 2.3 Hz, 1H), 7.04 (dd, J = 11.1, 4.0 Hz, 1H), 6.96 (t, J = 7.5 Hz, 1H), 6.35 (s, 1H), 6.32 (s, 1H), 4.48 (dd, J = 13.8, 8.2 Hz, 1H), 4.30 – 4.25 (m, 1H), 4.09 – 4.04 (m, 1H), 3.56 (s, 3H), 3.11 (dd, J = 14.6, 5.5 Hz, 1H), 3.01 (m, 2H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.55 (d, J = 12.4 Hz, 1H), 2.12 – 2.00 (m, 2H), 1.60 – 1.19 (m, 6H).
Biotinyl-tryptophan	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.53 (s, 1H), 10.79 (s, 1H), 8.03 (d, J = 7.9 Hz, 1H), 7.50 (d, J = 7.8 Hz, 1H), 7.31 (d, J = 8.1 Hz, 1H), 7.10 (d, J = 2.2 Hz, 1H), 7.04 (t, J = 7.5 Hz, 1H), 6.96 (t, J = 7.1 Hz, 1H), 6.35 (s, 1H), 6.32 (s, 1H), 4.45 (td, J = 8.5, 5.1 Hz, 1H), 4.30 – 4.25 (m, 1H), 4.08 – 4.03 (m, 1H), 3.13 (dd, J = 14.6, 5.0 Hz, 1H), 3.03 – 2.93 (m, 2H), 2.79 (dd, J = 12.4, 5.1 Hz, 1H), 2.55 (d, J = 12.4 Hz, 1H), 2.11 – 1.99 (m, 2H), 1.60 – 1.13 (m, 6H).

Biotinyl-tryptophan oxazolone (BWO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 10.86 (s, 1H), 7.50 (d, J = 7.9 Hz, 1H), 7.29 (dd, J = 8.1, 2.6 Hz, 1H), 7.06 (t, J = 2.6 Hz, 1H), 7.03 (dd, J = 11.1, 4.0 Hz, 1H), 6.94 (dd, J = 11.0, 3.9 Hz, 1H), 6.37 (s, 1H), 6.33 (s, 1H), 4.67 (d, J = 4.9 Hz, 1H), 4.33 – 4.24 (m, 1H), 4.09 – 4.04 (m, 1H), 3.23 (dd, J = 14.5, 4.6 Hz, 1H), 3.13 (ddd, J = 14.8, 6.0, 1.9 Hz, 1H), 2.96 (ddd, J = 17.9, 11.6, 6.6 Hz, 1H), 2.81 (dt, J = 12.4, 4.9 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.31 – 2.18 (m, 2H), 1.56 – 1.07 (m, 6H).
Biotinyl-phenylalanine methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ = 8.24 (d, J = 7.8 Hz, 1H), 7.30 – 7.14 (m, 5H), 6.37 (s, 1H), 6.33 (s, 1H), 4.44 (td, J = 9.3, 5.5 Hz, 1H), 4.32 – 4.25 (m, 1H), 4.13 – 4.04 (m, 1H), 3.57 (d, J = 8.5 Hz, 3H), 3.08 – 2.96 (m, 2H), 2.85 (dd, J = 13.7, 9.7 Hz, 1H), 2.81 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.09 – 1.98 (m, 2H), 1.60 – 1.12 (m, 6H).
Biotinyl-phenylalanine	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.57 (s, 1H), 8.03 (d, J = 8.1 Hz, 1H), 7.25 – 7.09 (m, 5H), 6.32 (s, 1H), 6.28 (s, 1H), 4.39 – 4.32 (m, 1H), 4.26 – 4.22 (m, 1H), 4.06 – 4.01 (m, 1H), 3.03 – 2.94 (m, 2H), 2.81 – 2.73 (m, 2H), 2.52 (d, J = 12.4 Hz, 1H), 2.03 – 1.92 (m, 2H), 1.57 – 1.07 (m, 6H).

Biotinyl-phenylalanine oxazolone (BFO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 7.31 – 7.12 (m, 5H), 6.41 (s, 1H), 6.33 (s, 1H), 4.72 – 4.66 (m, 1H), 4.32 – 4.26 (m, 1H), 4.14 – 4.07 (m, 1H), 3.12 (dd, J = 14.0, 5.1 Hz, 1H), 3.07 – 3.02 (m, 1H), 2.96 (dd, J = 14.0, 6.8 Hz, 1H), 2.81 (ddd, J = 12.4, 5.1, 2.2 Hz, 1H), 2.57 (d, J = 12.4 Hz, 1H), 2.39 – 2.25 (m, 2H), 1.62 – 1.16 (m, 6H).
Biotinyl-leucine methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ = 8.13 (d, J = 7.7 Hz, 1H), 6.37 (s, 1H), 6.33 (s, 1H), 4.31 – 4.27 (m, 1H), 4.26 – 4.21 (m, 1H), 4.13 – 4.08 (m, 1H), 3.59 (s, 3H), 3.10 – 3.03 (m, 1H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.10 (t, J = 7.3 Hz, 2H), 1.67 – 1.19 (m, 9H), 0.84 (dd, J = 32.3, 6.6 Hz, 6H).
Biotinyl-leucine	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.37 (s, 1H), 7.94 (d, J = 8.0 Hz, 1H), 6.32 (s, 1H), 6.28 (s, 1H), 4.26 – 4.21 (m, 1H), 4.14 (dd, J = 14.0, 9.0 Hz, 1H), 4.06 (d, J = 6.0 Hz, 1H), 3.02 (t, J = 9.3 Hz, 1H), 2.76 (dd, J = 12.4, 5.1 Hz, 1H), 2.51 (d, J = 12.4 Hz, 1H), 2.05 (t, J = 7.2 Hz, 2H), 1.62 – 1.21 (m, 9H), 0.80 (dd, J = 32.1, 6.5 Hz, 6H).

Biotinyl-leucine oxazolone (BLO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 6.41 (s, 1H), 6.33 (s, 1H), 4.35 (dd, J = 8.1, 6.5 Hz, 1H), 4.31 – 4.26 (m, 1H), 4.14 – 4.09 (m, 1H), 3.12 – 3.06 (m, 1H), 2.81 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.44 (tt, J = 10.8, 5.4 Hz, 2H), 1.86 – 1.33 (m, 9H), 0.89 (dd, J = 14.2, 6.7 Hz, 6H).
Biotinyl-isoleucine methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ = 8.00 (d, J = 8.0 Hz, 1H), 6.32 (s, 1H), 6.27 (s, 1H), 4.26 – 4.20 (m, 1H), 4.13 (t, J = 7.4 Hz, 1H), 4.06 (d, J = 2.6 Hz, 1H), 3.54 (s, 3H), 3.02 (dd, J = 6.4, 3.9 Hz, 1H), 2.75 (dd, J = 12.4, 5.0 Hz, 1H), 2.50 (d, J = 12.4 Hz, 1H), 2.15 – 1.99 (m, 2H), 1.73 – 1.05 (m, 9H), 0.76 (dd, J = 7.0, 5.3 Hz, 6H).
Biotinyl-isoleucine	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.46 (s, 1H), 7.90 (d, J = 8.4 Hz, 1H), 6.37 (s, 1H), 6.32 (s, 1H), 4.31 – 4.26 (m, 1H), 4.15 (dd, J = 8.3, 6.3 Hz, 1H), 4.13 – 4.08 (m, 1H), 3.07 (dt, J = 8.6, 6.1 Hz, 1H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.19 – 2.07 (m, 2H), 1.78 – 1.11 (m, 9H), 0.86 – 0.79 (m, 6H).
Biotinyl-isoleucine oxazolone (BIO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 6.41 (s, 1H), 6.33 (s, 1H), 4.34 (ddt, J = 32.9, 4.1, 1.9 Hz, 1H), 4.30 – 4.26 (m, 1H), 4.14 – 4.10 (m, 1H), 3.13 – 3.05 (m, 1H), 2.81 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.46 (dd, J = 7.3, 1.9 Hz, 2H), 1.92 – 1.14 (m, 9H), 0.94 – 0.69 (m, 6H).

Biotinyl-valine methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ = 7.99 (d, J = 8.0 Hz, 1H), 6.33 (s, 1H), 6.28 (s, 1H), 4.27 – 4.21 (m, 1H), 4.12 – 4.02 (m, 2H), 3.56 (s, 3H), 3.07 – 2.99 (m, 1H), 2.76 (dd, J = 12.3, 5.0 Hz, 1H), 2.51 (d, J = 12.5 Hz, 1H), 2.15 – 2.04 (m, 2H), 1.94 (dt, J = 13.1, 6.5 Hz, 1H), 1.59 – 1.18 (m, 6H), 0.80 (dd, J = 12.1, 6.8 Hz, 6H).
Biotinyl-valine	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.47 (s, 1H), 7.88 (d, J = 8.5 Hz, 1H), 6.37 (s, 1H), 6.32 (s, 1H), 4.31 – 4.25 (m, 1H), 4.11 (m, 2H), 3.07 (dt, J = 8.6, 6.0 Hz, 1H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.55 (d, J = 12.4 Hz, 1H), 2.21 – 2.08 (m, 2H), 2.01 (dq, J = 13.4, 6.7 Hz, 1H), 1.66 – 1.22 (m, 6H), 0.85 (dd, J = 6.8, 1.9 Hz, 6H).
Biotinyl-valine oxazolone (BVO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 6.41 (s, 1H), 6.33 (s, 1H), 4.31 – 4.24 (m, 2H), 4.14 – 4.09 (m, 1H), 3.13 – 3.05 (m, 1H), 2.81 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.46 (dd, J = 8.9, 6.9 Hz, 2H), 2.15 – 2.06 (m, 1H), 1.67 – 1.32 (m, 6H), 1.00 – 0.79 (m, 6H).
Biotinyl-methionine methyl ester	^{1}H NMR (600 MHz, DMSO-d_6) δ = 8.18 (d, J = 7.5 Hz, 1H), 6.37 (s, 1H), 6.33 (s, 1H), 4.37 – 4.31 (m, 1H), 4.31 – 4.26 (m, 1H), 4.15 – 4.08 (m, 1H), 3.60 (s, 3H), 3.10 – 3.04 (m, 1H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.53 – 2.38 (m, 2H), 2.10 (t, J = 7.4 Hz, 2H), 2.02 (s, 3H), 1.95 – 1.78 (m, 2H), 1.66 – 1.20 (m, 6H).

Biotinyl-methionine	^{1}H NMR (600 MHz, DMSO-d_6) δ = 12.53 (s, 1H), 8.05 (d, J = 7.8 Hz, 1H), 6.37 (s, 1H), 6.33 (s, 1H), 4.32 – 4.24 (m, 2H), 4.14 – 4.08 (m, 1H), 3.12 – 3.03 (m, 1H), 2.80 (dd, J = 12.4, 5.1 Hz, 1H), 2.56 (d, J = 12.4 Hz, 1H), 2.51 – 2.38 (m, 2H), 2.10 (t, J = 7.3 Hz, 2H), 2.02 (s, 3H), 1.97 – 1.75 (m, 2H), 1.67 – 1.21 (m, 6H).
Biotinyl-methionine oxazolone (BMO)	^{1}H NMR (600 MHz, DMSO-d_6) δ = 6.41 (s, 1H), 6.33 (s, 1H), 4.45 (td, J = 5.7, 2.0 Hz, 1H), 4.29 (dd, J = 7.6, 5.2 Hz, 1H), 4.15 – 4.09 (m, 1H), 3.09 (dd, J = 12.6, 6.4 Hz, 1H), 2.81 (dd, J = 12.4, 5.1 Hz, 1H), 2.59 – 2.38 (m, 5H), 2.08 – 1.83 (m, 5H), 1.63 – 1.37 (m, 6H).

4.3 Synthesis of Biotinyl-Tyr(Me)-Oxazolone (BYO)

4.3.1 General Synthesis Procedures

Reagents and solvents were obtained from Fluka, Sigma-Aldrich or Bachem, and were used without further purification. NMR spectra in either $CDCl_3$, DMSO-d_6 or D_2O solution were recorded on a Bruker DPX 300 spectrometer (300 MHz) or on a Bruker Avance 400 spectrometer (400 MHz); chemical shifts δH are reported in ppm with reference to the solvent resonance ($CDCl_3$: δH = 7.26 ppm; DMSO: δH = 2.50 ppm; H_2O: δH = 4.79 ppm); coupling constants J are reported in Hz. UHPLC analyses were carried out on a Thermo Scientific Dionex UltiMate 3000 Standard system including an autosampler unit, a thermostated column compartment and a photodiode array detector, using UV absorbance detection at λ = 273 nm. HPLC/ESI-MS analyses were carried out on a

Waters UPLC Acquity H-Class system including a photodiode array detector (acquisition in the 200–400 nm range), coupled to a Waters Synapt G2-S mass spectrometer, with capillary and cone voltage of 30 kV and 30 V respectively, source and desolvation temperature of 140 °C and 450 °C respectively. ESI+ and ESI– refer to electrospray ionisation in positive and negative mode respectively. HRMS spectra were recorded on the same spectrometer, using the same source settings as above.

4.3.2 Preparation of *N*-tert-Butoxycarbonyl-*O*-methyl-tyrosine methyl ester (Boc-Tyr(Me)-OMe)

Synthesis of Boc-Tyr(Me)-OMe was carried out according to a published procedure.[296,297] A solution of Boc-Tyr-OH (7.0 mmol, 2.0 g; Bachem) in dimethylformamide (DMF, 20 mL) was cooled using an ice bath and treated with freshly ground KOH (7.7 mmol, 0.43 g). A cooled solution of CH_3I (7.7 mmol, 0.49 mL) in DMF (5 mL) was added dropwise over 1 min. The mixture was stirred at room temperature for 30 min, then cooled using an ice bath, and additional KOH (7.7 mmol, 0.43 g) and a cooled solution of CH_3I (7.7 mmol, 0.49 mL) in DMF (5 mL) were added dropwise over 1 min. The mixture was stirred for 3 h at room temperature, poured onto ice (40 g), and extracted with ethyl acetate (3 × 20 mL). The organic layers were washed with water (3 × 13 mL), brine (2 × 13 mL), and dried over Na_2SO_4. The solvent was removed under reduced pressure to afford a colorless oily residue. Then the oil was purified by preparative silica gel chromatography (mobile phase: ethyl acetate - hexane, 3:7 v/v) (yield: 1.4 g, 63.6%). NMR spectrum on Table 4.1.

4.3.3 Preparation of Biotinylated *O*-Methyl-Tyrosine (Biotin-Tyr(Me)-OH)

This compound was prepared in three steps. In the first stage, Boc-Tyr(Me)-OMe (1.0 g) was treated by trifluoroacetic acid (TFA) / water solution (9:1 v/v, 2 ml) for 30 min. TFA was removed by evaporation *in vacuo*, the residue was poured into diethyl ether, the TFA salt of H-Tyr(Me)-OMe was collected by filtration as a white precipitate (yield: 0.88 g, 84%). NMR spectrum on Table 4.1.

In the second stage, biotin (345 mg, 1.41 mmol), was activated with 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC, 293 mg, 1.55 mmol) and hydroxybenzotriazole monohydrate (HOBt, 242 mg, 1.55 mmol) in a mixture of CH_2Cl_2 (7 ml) and DMF (7 ml). The mixture was stirred for 5 h. Then the TFA salt of H-Tyr(Me)-OMe (500 mg, 1.55 mmol) was added with *N*-ethyl-*N,N*-diisopropylamine (DIEA, 531 μL, 3 mmol) and the mixture stirred overnight. DMF was removed under reduced pressure, and the residue redissolved in ethyl acetate (100 ml), washed with water (30 ml), 1M $KHSO_4$ (10 ml), $NaHCO_3$ (saturated solution, 10 ml),and brine (10 ml), consecutively. The solution was dried over anhydrous Na_2SO_4 and concentrated under reduced pressure. Residual DMF was removed by dissolving the residue in ethyl acetate (20 ml) and precipitation with hexane (5 ml). The solid was recovered by filtration, washed with hexane and dried *in vacuo* (300 mg, 46%).

The methyl ester biotinyl-Tyr(Me)-OMe (270 mg, 0.62 mmol) was then dissolved in iPrOH:H2O (7:3 v/v) (minimum volume), treated with 1N NaOH (0.93 ml). The mixture was stirred at room temperature overnight. The solvent was removed under reduced pressure and the product was precipitated upon addition of water and acidification with 1M HCl. The free acid biotinyl-Tyr(Me)-OH was recovered by filtration as a white solid, washed with water and then dried under reduced pressure (yield: 226 mg, 86%). NMR

spectrum on Table 4.1. HRMS (ESI+): *m/z* calcd for $C_{20}H_{28}N_3O_5S$ $[M + H]^+$ 422.1750; found, 422.1747.

4.3.4 Preparation of Biotinyl-Tyr(Me)-Oxazolone (BYO)

In a typical experiment, biotinyl-Tyr(Me)-OH (42 mg, 0.1 mmol) was mixed with CH_2Cl_2 (3 ml) and then EDC (21.9 mg, 0.12 mmol) was added. After stirring by magnetic stirrer for 1 h, all the starting material was dissolved. Additional CH_2Cl_2 (3 ml) was added, then the mixture was washed by H_2O (5 ml) twice and saturated brine (5 ml) once. The organic layer was dried by anhydrous Na_2SO_4 and concentrated under reduced pressure. The residue was dried in vacuum in the presence of P_2O_5 for 1 h. The product was stored under -20 °C, or kept in a solution of CH_3CN under -20 °C. HRMS (ESI+): *m/z* calcd for $C_{20}H_{26}N_3O_4S$ $[M + H]^+$ 404.1644; found, 404.1644. See Figures 4.1, 4.2, and 4.3 for NMR data.

4.4 Synthesis of Additional Biotinyl-Aminoacyl Oxazolones (BXO)

4.4.1 General Synthesis Procedures

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 (δH = 2.50 ppm). Spectra were analyzed using MNova software.

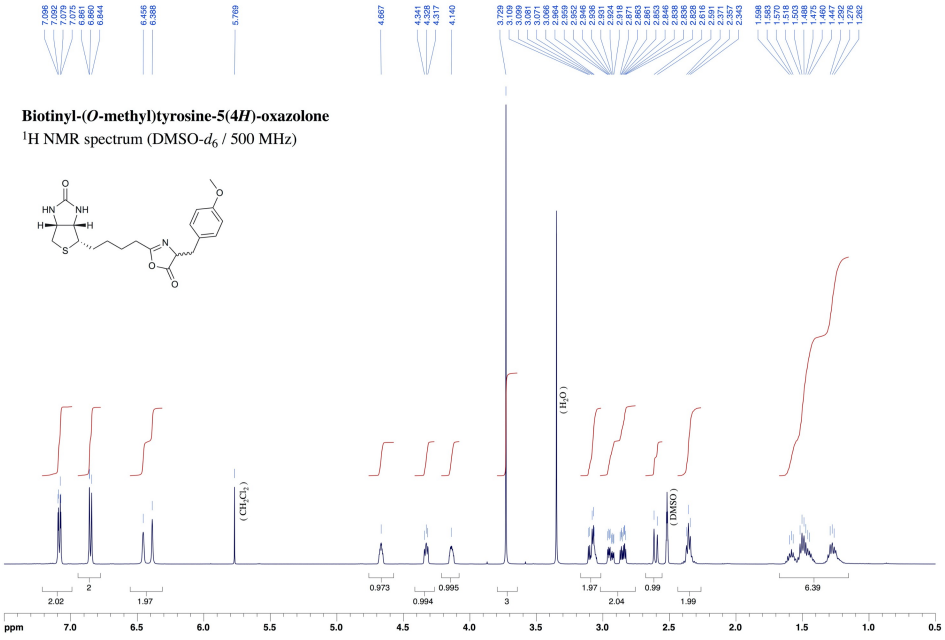

Figure 4.1: **^{1}H NMR spectra for biotinyl-tyrosine oxazolone (BYO).**

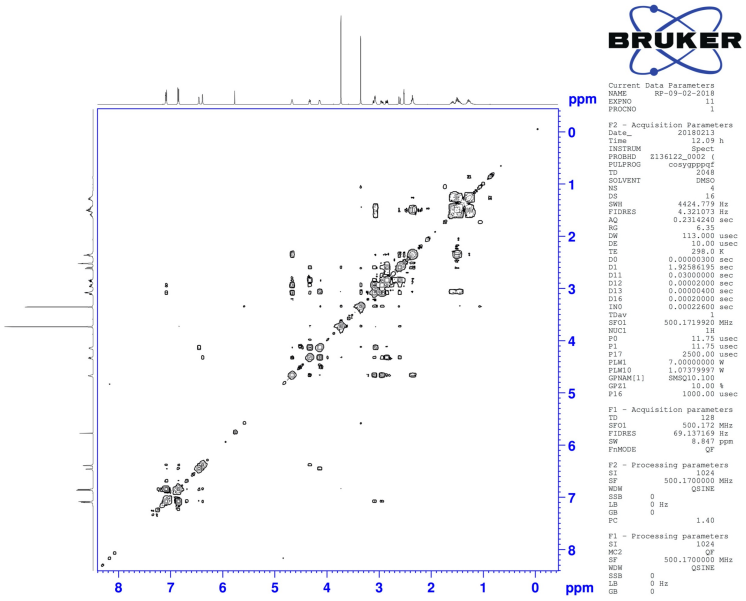

Figure 4.2: **2D-COSY NMR spectra for biotinyl-tyrosine oxazolone (BYO).**

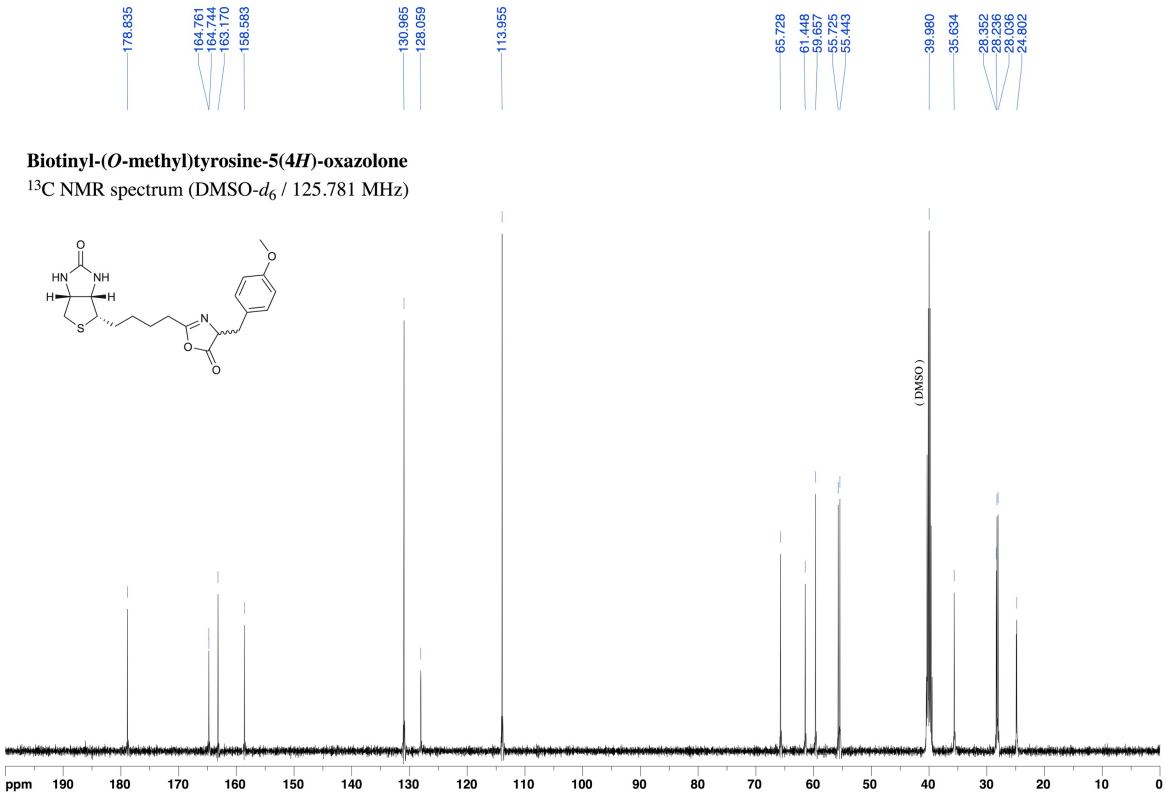

Figure 4.3: **^{13}C NMR spectra for biotinyl-tyrosine oxazolone (BYO).**

4.4.2 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 (^{1}H NMR chemical shifts reported in Table 4.1).

The recovered compound was dissolved by sonication in iPrOH:H2O (2:1 v/v) (15 mL), to which 1 mL of 3M 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 1M HCl to produce a white solid. This compound was recovered by filtration, washed with water, and dried *in vacuo* (Table 4.1).

4.4.3 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 (Figures 4.4 - 4.9 and Table 4.1). NMR characterization was

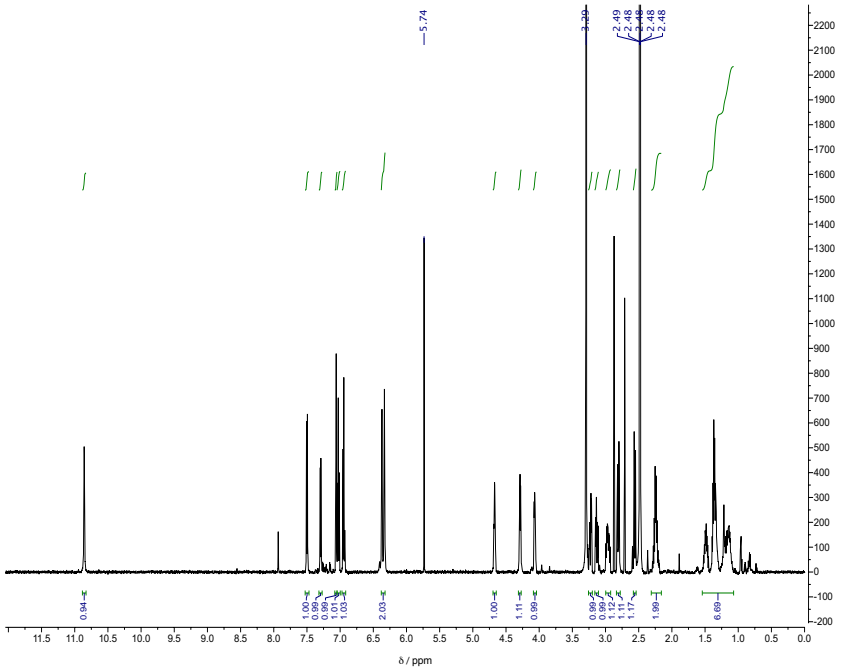

Figure 4.4: **^{1}H NMR spectra for biotinyl-tryptophan oxazolone (BWO).**

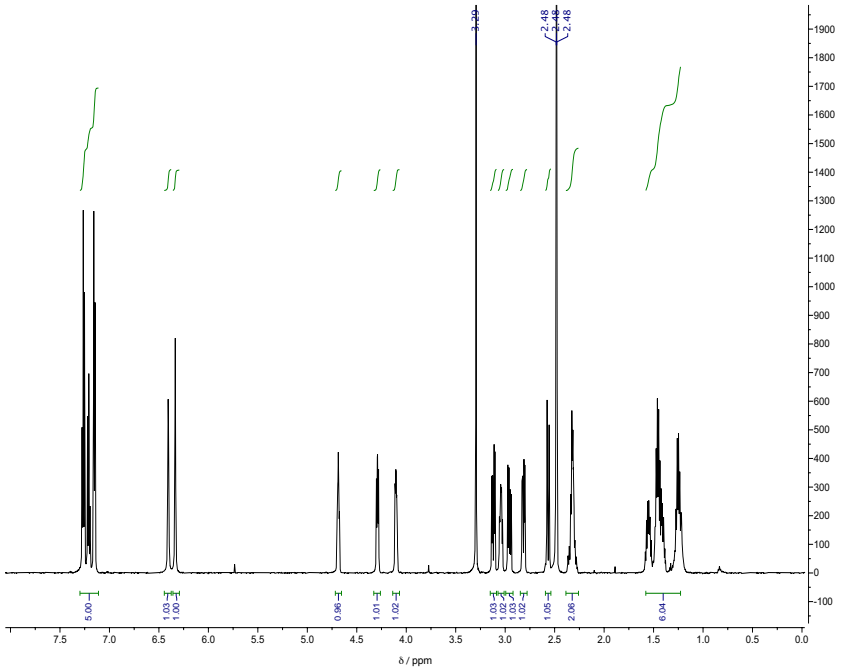

Figure 4.5: **^{1}H NMR spectra for biotinyl-phenylalanine oxazolone (BFO).**

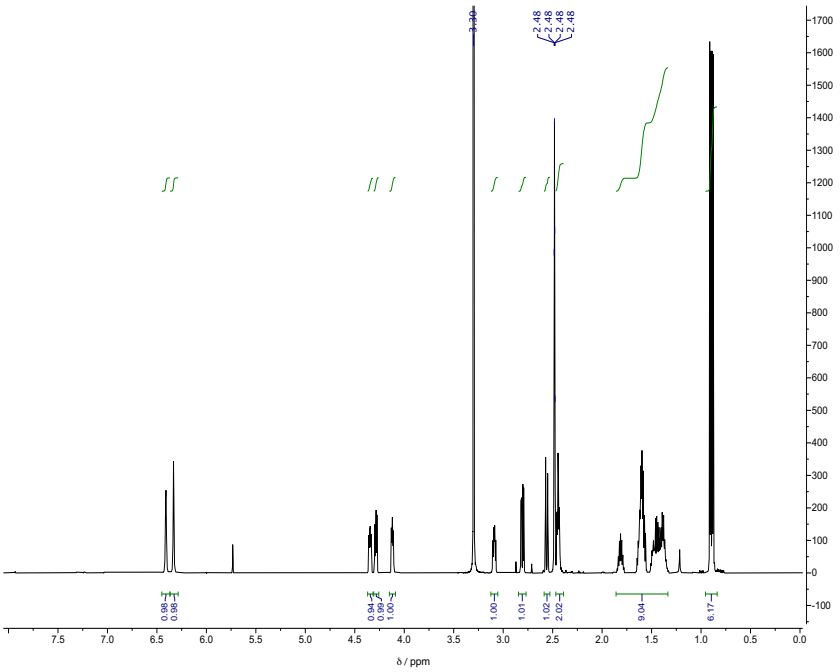

Figure 4.6: **^{1}H NMR spectra for biotinyl-leucine oxazolone (BLO).**

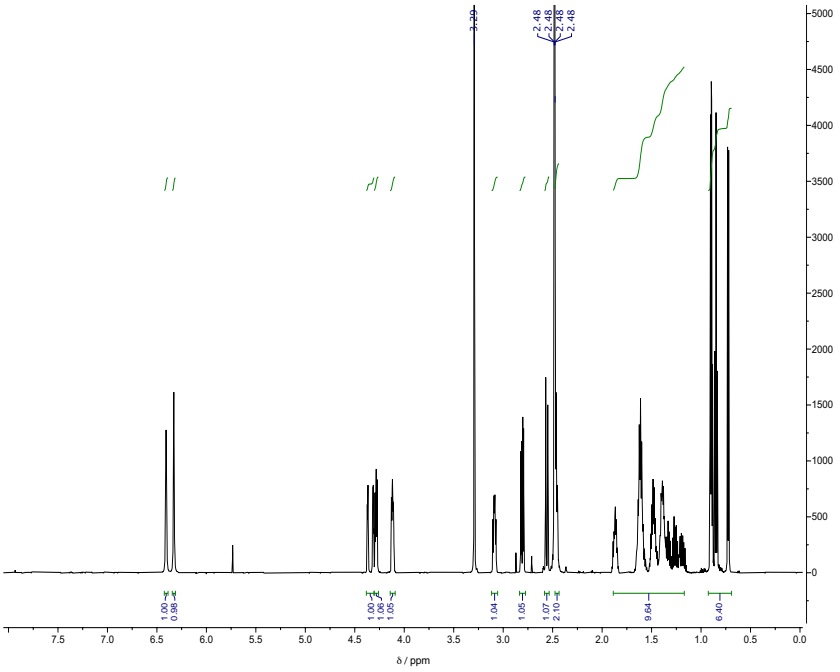

Figure 4.7: **^{1}H NMR spectra for biotinyl-isoleucine oxazolone (BIO).**

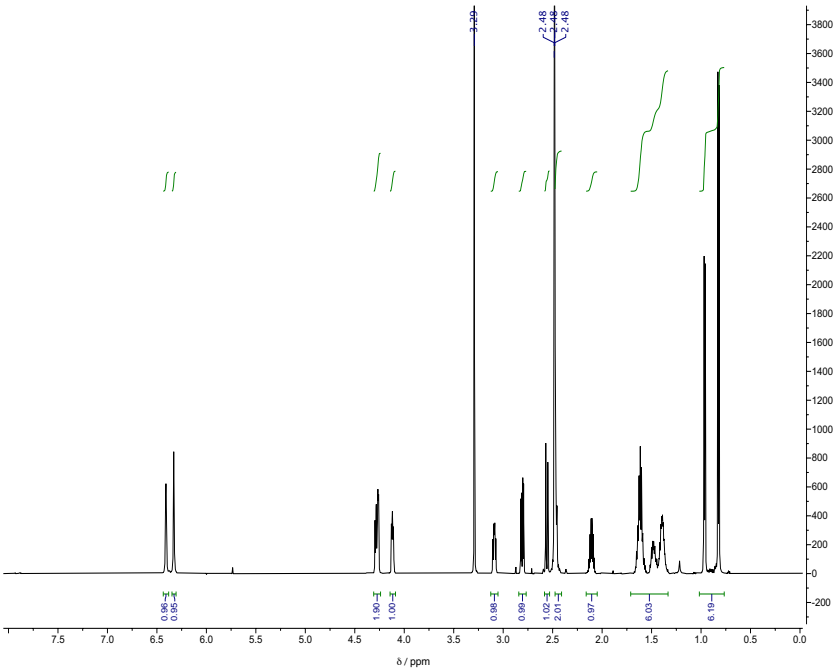

Figure 4.8: **^{1}H NMR spectra for biotinyl-valine oxazolone (BVO).**

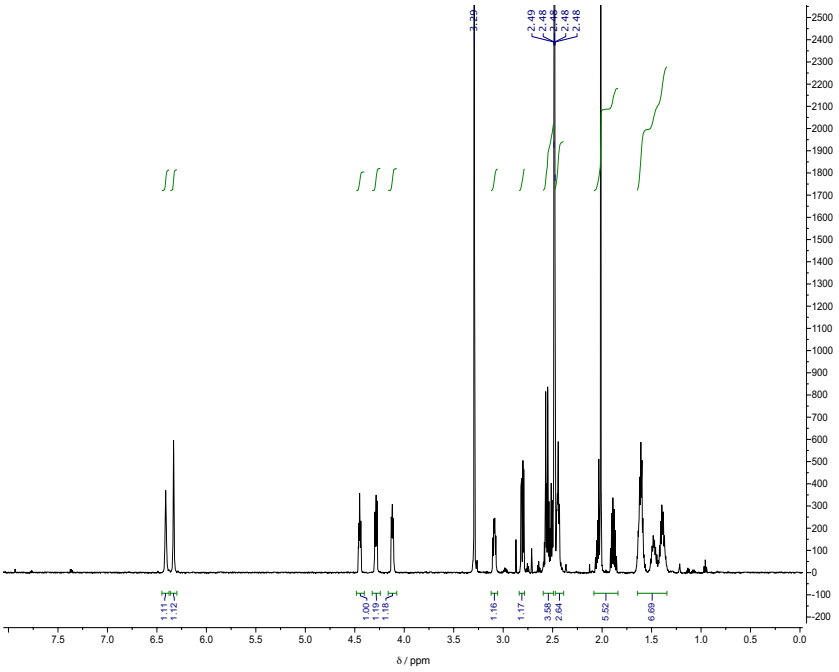

Figure 4.9: **^{1}H NMR spectra for biotinyl-methionine oxazolone (BMO).**

Compound	[*biotin*] (mM)	σ
BWO	9.57	0.98
BFO	10.37	1.88
BLO	14.35	0.42
BIO	13.49	1.34
BVO	14.00	2.13
BMO	12.56	1.51

Table 4.2: **Biotin quantification of BXO solutions.** Average measured biotin concentration ([*biotin*]) and standard deviation (σ) of prepared BXO solutions (where X = W (Trp), F (Phe), L (Leu), I (Ile), V (Val), or M (Met)). Expected concentrations were 25 mM.

performed as described above.

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 Table 4.2 (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.

4.5 Aminoacylation Ribozyme Selections

Chemical synthesis (IDT, PAGE purification) was used to obtain a library of DNA molecules having the sequence 5'-GATAATACGACTCACTATA-GGGAATGGATCCACATCTACGAATTC-N21-TTCACTGCAGACTTGACGAAGCTG-

3', where N21 denotes 21 consecutive random positions and nucleotides upstream of the transcription start site are underlined. Two replicates of the selection were performed (RS1 and RS2), beginning with 9.1 (coverage ≈ 1.3-fold) and 145 pmol (coverage ≈ 20-fold) of DNA for RS1 and RS2, respectively. RNA was transcribed using HiScribe T7 polymerase (New England Biolabs) and purified by denaturing polyacrylamide gel electrophoresis (PAGE). In the first round of selection, $3.4x10^{14}$ or $1.9x10^{15}$ RNA sequences (RS1 and RS2, respectively) were incubated with 50 μM BYO in the aminoacylation selection buffer (100 mM HEPES (pH 6.95), 100 mM NaCl, 100 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$) for 90 min, at an RNA concentration of 1.4 - 3.2 μM. The reaction was stopped by removing unreacted substrate using Bio-Spin P-30 Tris desalting columns (Bio-Rad). Streptavidin MagneSphere paramagnetic beads (Promega) were used to isolate reacted sequences at a volume ratio of 1:4, which were then eluted with a 5 min incubation at 65 °C in a solution containing 95% formamide and 10 mM EDTA. Sequences were prepared for the next round of selection by reverse transcription and PCR (RT-PCR), with primers complementary to the fixed sequence shown above. Five additional rounds of selection were performed using the same procedure, with ∼400 pmol ($\sim 2x10^{14}$ molecules; 2 μM) of RNA in each round. DNA samples from each round were barcoded and pooled for sequencing by Illumina NextSeq 500 (Biological Nanostructures Laboratory, California NanoSystems Institute at UCSB).

Selections for self-aminoacylating ribozymes with BFO and BLO were conducted as described for BYO aminoacylation.[200] Libraries were obtained from IDT with the sequence 5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-N21-TTCACTGCAGACTTGACGAAGCTG-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, reac-

tions 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)

4.6 Clustering Analyses of Sequences from Selections

Clustering of BYO selection data was performed on the Galaxy platform[298] for sequences in Rounds 4-6. Multiple families containing the same motif were designated as 1A.1, 1A.2, etc., or 2.1, 2.2, etc. Center sequences were used to assign each family to a motif. SeqLogo plots[299] representing motifs were generated from all sequences identified among every family grouped into that motif. Sequences are named according to the convention: S-Motif.family rank-sequence rank, where rank is determined by relative abundance in Round 6. For example, S-1B.1-a is the top-ranked sequence from the top-ranked family of Motif 1B.

Sequences from BFO and BLO selections were clustered into families based on sequence similarity, using a custom Python script. The script ClusterBOSS.py uses the enumerated read output files generated from the EasyDIVER package.[300] 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 ($d_{cutoff} = 3$ 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}), 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, the same procedure was followed until all sequences had been explored.

4.7 Kinetic Sequencing (k-Seq) Experiments and Analyses

4.7.1 Kinetic Sequencing (k-Seq) of BYO Selection Pool

2 μmoles of RNA from Round 5 (RS1) were incubated with BYO substrate at various concentrations (2, 10, 50, and 250 μM), under buffer conditions and reaction time otherwise identical to those during selection. Streptavidin beads were added at a volume ratio of 1:1, and bound RNA was eluted as described above. To enable absolute quantitation of the products, 4, 12, 17, and 42 fmol, respectively, of a control RNA sequence were

spiked into the RNA eluted from each concentration point. The spike-in control sequence was transcribed by T7 RNA polymerase from a DNA oligonucleotide (IDT) having the sequence 5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-AAAAACAAAAACAAAAACAAATTCACTGCAGACTTGACGAAGCTG-3' (promoter underlined). The k-Seq reactions were performed in triplicate, barcoded, and sequenced as described above.

Every unique sequence detected in Round 5 was tracked across all 12 k-Seq samples. The absolute concentration of each sequence was calculated as $(n_s/n_{spike})[spike]$, where n_s and n_{spike} are the number of reads found for sequence s or the spike-in sequence, respectively, and [spike] is the known concentration of the spike-in sequence in the sample. Concentrations were averaged across triplicates and fit to the first-order rate equation $F_s([BYO]) = A_s(1 - e^{(-k_s\alpha[BYO]t)})$, where F_s is the measured fraction of sequence s reacted, A_s is the maximum reacted fraction, t is the incubation time of 90 min, and k_s is the effective rate constant of the reaction catalyzed by sequence s. α is the coefficient accounting for the hydrolysis of substrate BXO during the reaction time (t=90 min), and a fixed value (0.479, see below).[200] To obtain an estimate of error, each set of 12 observations was randomly grouped into three series of four concentrations, k_s and A_s were fit individually for each set, and the standard deviation among the three series was calculated.

For sequences of low activity, the parameter A_s could not be accurately estimated over the concentrations tested, leading to a fitting artifact with $A_s = 1$ and underestimation of k_s. However, while A_s and k_s are poorly estimated individually, the combined chemical activity parameter k_sA_s is estimated more accurately. Thus k_sA_s was used to compare catalytic activity across the broad range of observed activity. The ratio of k_sA_s to k_0A_0 (the uncatalyzed activity, see below) is defined as the catalytic enhancement of sequence s (r_s).

4.7.2 Kinetic Sequencing (k-Seq) of Variable Pools

DNA libraries for kinetic sequencing experiments were designed as described.[301] Libraries were obtained from Integrated DNA Technologies (IDT) or Keck Biotechnology Laboratory with the sequence 5'-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-[central variable region, length 21]-TTCACTGCAGACTTGACGAAGCTG-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 (Table 4.3) 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).

Family	Wild-type sequence for 21-nt selected region
1A.1	CTACTTCAAACAATCGGTCTG
1B.1	CCACACTTCAAGCAATCGGTC
2.1	ATTACCCTGGTCATCGAGTGA
2.2	ATTCACCTAGGTCATCGGGTG
3.1	AAGTTTGCTAATAGTCGCAAG

Table 4.3: **Wild-type sequences for ribozyme families used in this study.**

Kinetic sequencing experiments were performed as previously described.[200,301] Reactions were performed in 50 μL aqueous solutions containing selection buffer (100 mM HEPES, 100 mM NaCl, 100 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$) 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. 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'-CAGCTTCGTCAAGTCTGCAGTGAA-3'. Serial dilutions of random library ssDNA were prepared in triplicate from $5x10^{-5}$ to $5x10^{2}$ pg/μL alongside each experiment for generating standard curves.[302] 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).

4.7.3 Computational Analyses of Variable Pool k-Seq Data

Sequencing reads were processed using trimmomatic SE CROP:90 to facilitate joining,[303] and then paired-end reads were joined and unique sequences were enumerated using EasyDIVER.[300] Joining was performed using the following PANDAseq[304] flags: -a -l 1 -A pear -C completely_miss_the_point:0. These flags strip primers after assembly rather than before (-a), requires sequences to have a minimum length of 1 after removing primers (-l 1), sets the assembly algorithm to PEAR[305] (-A pear), and excludes sequences with mismatches in overlapping paired-end regions (completely_miss_the_point:0). Primer sequences were extracted using CTACGAATTC as the forward primer and CTGCAGT-

GAA as the reverse primer.

k-Seq analyses were performed using developed 'k-seq' package.[301] 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 $F_s^{BXO} = A_s(1 - e^{(-k_s \alpha [BXO] t)})$, 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 [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 BYO[200]) 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 product $k_s A_s$ was used to represent the activity of each sequence.

4.8 Determination of Aminoacylation Rates by Electrophoretic Mobility Shift Assay

Ten sequences were chosen from among the top 20 peaks for experimental testing. The corresponding DNA oligonucleotides were obtained from IDT (HPLC-purified) and RNA was transcribed using T7 RNA polymerase. In addition, a control sample of random pool sequences was used to determine baseline uncatalyzed activity (k_0A_0, measured as a combined parameter). RNA was labeled using a 5' EndTag Labeling Kit (Vector Laboratories) with Alexa Fluor 488 (Fisher), and purified by phenol-chloroform extraction. Labeled RNA sequences were then incubated (RNA concentration of 100 nM) with BYO for 90 min under conditions described as above for k-Seq. Following desalting, samples were incubated with 2 μM streptavidin for 15 min in 10 mM Tris (pH 7.0), then analyzed by native PAGE. Gels were scanned and fluorescence was quantified with ImageQuant software on an Amersham Typhoon 5 Biomolecular Imager. Bands corresponding to the streptavidin complex and the free RNA band were quantified to calculate the fraction of each sequence that had undergone aminoacylation. Values determined by k-Seq were compared to gel shift percentages to determine the average fraction loss l during streptavidin bead pull-down. This value of l was used as a correction factor when calculating catalytic enhancements using k-Seq data, as k_0A_0 was measured by gel-shift assay.

For determining the uncatalyzed reaction rate with BFO, aminoacylation reactions were performed in 50 μL selection buffer with 5% acetonitrile and contained 0.43 μM random library RNA and BFO at 1250, 250, 50, 10, or 2 μM. 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 and run on an 8% polyacrylamide gel. Gel

shift assays for qualitative observation of reactivity were performed as above with 500 μM BXO per sample unless otherwise noted.

4.9 Degradation Rate of BYO

RNA sequence S-1A.1-a was added to 250 μM BYO that was pre-incubated with reaction buffer for 5 - 180 min. The initial rate of the reaction was compared to the reaction kinetics for this sequence determined without pre-incubation of BYO (see above). The effective concentration of BYO at the start of reaction was calculated assuming a first-order reaction (i.e., effective [BYO] = (250 μM x initial rate)/($k_s A_s$), where $k_s A_s$ is the activity of the ribozyme without pre-incubation of BYO), giving a half-life for BYO of 36.5 min. Reaction rates for ribozymes were adjusted accordingly to account for lower effective substrate concentrations.

4.10 Identification of Reactive Nucleotides

Ribozyme aminoacylation reactions were performed in selection buffer containing 1 μM RNA and 500 μM BYO and incubated with gentle agitation for 90 min. RNA was concentrated using Amicon Ultracel-3 filters (EMD Millipore) and an adapter oligo having the sequence 5'-AACCTGCTGTCATCGTCGTCCCTATAGTGAGC-3' was adenylated using a 5' adenylation kit (NEB) and ligated to the 3' end using T4 RNA Ligase 2, truncated KQ (New England BioLabs) (see exception noted below). The ligated products were gel purified and reverse transcribed using a 5' Rhodamine Green-X-tagged reverse primer complementary to a region of the adapter sequence (5'-CTCACTATAGGGACGACGATGACAGCAGG-3') and SuperScript III Reverse Transcriptase (Thermo Fisher), with a 10 min extension at 55 °C. Reverse transcripts were

run on a 12% denaturing sequencing gel, scanned on an Amersham Typhoon 5 Biomolecular Imager. The likely site of truncation was identified by gel position from the primer (bands at single nucleotide resolution could be visualized at high contrast; see Figure 5.3). To verify specific 2'-OH positions, RNA sequences containing 2'-O-methyl modifications were obtained from IDT and tested for aminoacylation activity by streptavidin gel shift (described above).

Sequences from families 1A.1 and 1B.1 were ligated to an alternative adapter oligo (5'-AAAACGGGCTTCGGTCCGGTTC-3'), as ligation to the original adapter oligo (listed above) was noted to interfere with folding of these sequences. The corresponding RT primer was 5'-GAACCGGACCGAAGCCCG-3'.

4.11 Background Reaction Rate Estimation of Results from Variable Pool k-Seq

Histograms (100 bins) of $\log_{10}$-transformed $k_s A_s$ values for sequences from all families were fit to a bimodal Gaussian distribution (Figure 4.10 and Table 4.4). The mean of the low-activity peak (μ_1) was used as the estimated uncatalyzed rate ($k_0 A_0$) and the standard deviation of the fit (σ_1) was used to inform the choice of catalytic enhancement threshold.

4.12 Promiscuity Index Calculations

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 Atkins.[192]

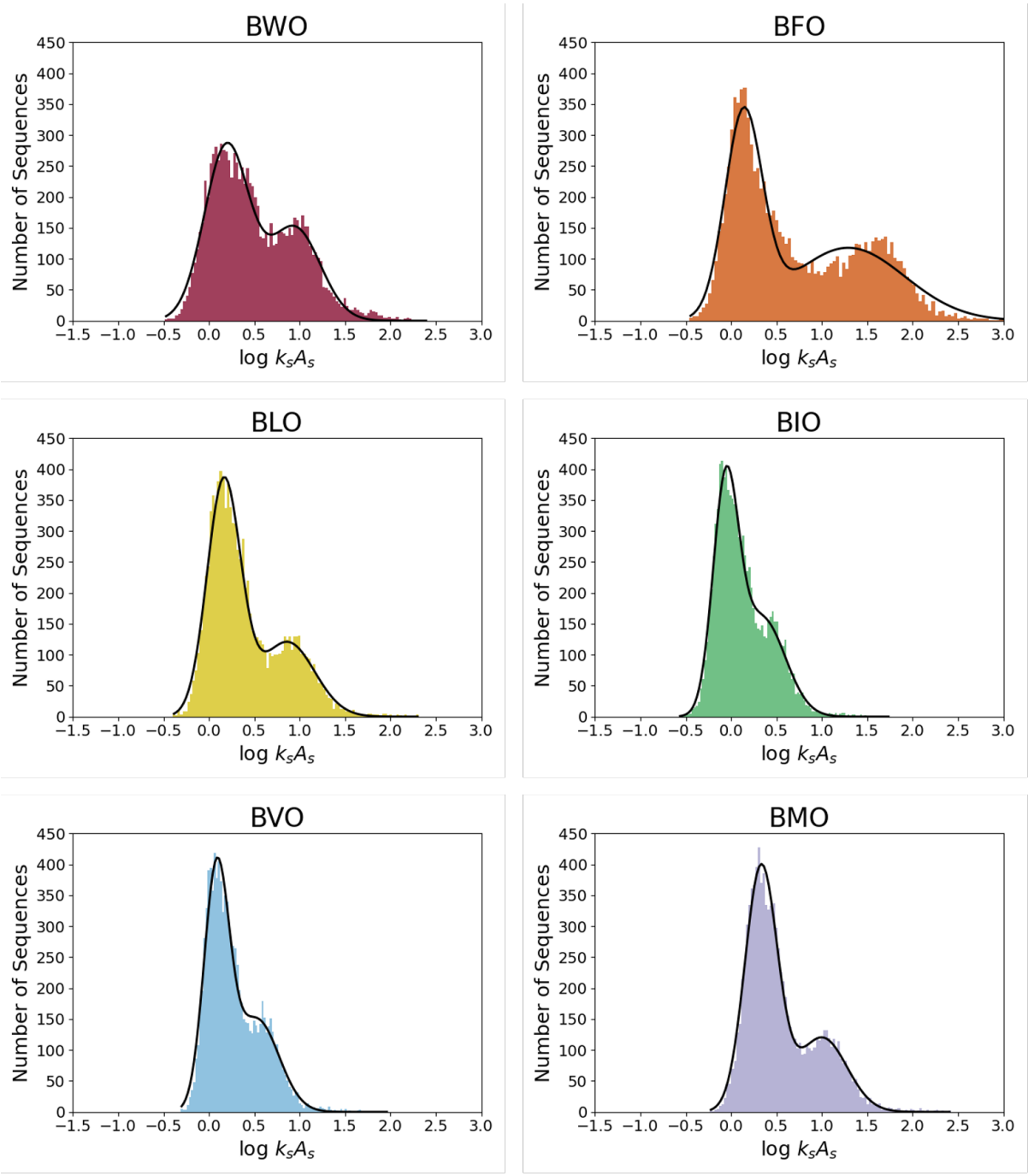

Figure 4.10: **Estimation of background rate k_0A_0.** Frequency distribution (histogram) of log_{10}-transformed k_sA_s for the ribozyme variants reacted with each substrate. The frequency distribution of ribozymes has been previously found to be log-normal.[60,200] Bimodal Gaussian fits (black lines) were used to characterize the low-activity peak using the equation $y = A_1e^{-(\frac{x-\mu_1}{2\sigma_1})^2} + A_2e^{-(\frac{x-\mu_2}{2\sigma_2})^2}$. The centers of the low-activity peaks (μ_1) and their standard deviations (σ_1) are shown in Table 4.4.

Substrate	μ_1	σ_1	2.5%	97.5%	10^{μ_1} ($M^{-1}min^{-1}$)	95% range ($M^{-1}min^{-1}$)
BWO	0.196	0.247	-0.289	0.680	1.57	0.51 - 4.79
BFO	0.138	0.205	-0.265	0.541	1.37	0.54 - 3.47
BLO	0.164	0.185	-0.200	0.527	1.46	0.63 - 3.36
BIO	-0.060	0.141	-0.337	0.217	0.87	0.46 - 1.65
BVO	0.083	0.140	-0.191	0.357	1.21	0.64 - 2.28
BMO	0.331	0.180	-0.021	0.684	2.14	0.95 - 4.83

Table 4.4: **Characterization of the background peaks.** The center of the low-activity peak, 10^{μ_1} (Figure 4.10), was used as the presumed background rate (k_0A_0) for each substrate.

4.13 Data and Code Availability

Galaxy computer code used is available as previously reported.[298] Additional scripts and files for review are available on GitHub (`https://github.com/ichen-lab-ucsb/SCAPE-BYO`). The HTS datasets generated and analyzed during the current study will be available in the UCSB Dash Data Repository.

Data from high-throughput sequencing and *k*-Seq analysis of the variable pools will be available at the Dryad Digital Repository (`https://doi.org/10.25349/D92C9C`).

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).

Chapter 5

Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA

5.1 Permissions and Attributions

Portions of this chapter were the result of collaboration with Abe Pressman, Ziwei Liu, Celia Blanco, Ulrich F. Müller, Gerald F. Joyce, Robert Pascal, and Irene A. Chen and have previously appeared in the Journal of the American Chemical Society.[200] It is reproduced here with the permission of ACS, to which further permissions related to the material excerpted should be directed: `https://pubs.acs.org/doi/10.1021/jacs.8b13298`.

5.2 Introduction

Molecular evolution is largely governed by the function of fitness in the space of all possible sequences, known as the "fitness landscape".[15,16] Evolution corresponds to a biased random walk on this landscape, in which mutation enables exploration of neighboring points in sequence space, and natural (or artificial) selection favors hill-climbing toward higher fitness. Therefore, knowledge of the fitness landscape is necessary for a systematic, quantitative understanding of molecular evolution.[17,20,306] For example, a deep question is whether the landscape allows selection to optimize biochemical activity. If the topography of the fitness landscape is relatively smooth, optimization by selection can occur readily through hill-climbing. However, if the landscape is riddled with low-fitness valleys between local fitness optima, then many potential evolutionary pathways through sequence space will be inaccessible, inhibiting global optimization of activity. A comprehensive map of the fitness landscape would enable understanding of such fundamental issues.

Fitness landscapes of ribozymes are of special interest because RNA may have been the first evolving molecule during an "RNA World" at the origin of life.[8,9,36,55,130,131,307] In addition, ribozymes have been proposed as the genetic and catalytic basis for a minimal synthetic cell.[308] On the practical side, ribozymes can be relatively short in length (L),[209] so it is possible to interrogate the entirety of sequence space in a laboratory setting (e.g., for $L = 21$, $4^{21} \approx 4x10^{12}$ possible sequences). Recent studies have emphasized the importance of comprehensive coverage vs. sparse sampling of sequence space for understanding evolutionary pathways. For example, sparse sampling (e.g., based on known genotypes) can miss viable evolutionary pathways and create a biased view of the fitness landscape.[59,309] Exhaustive data could also aid computational efforts to explore larger sequence spaces.[310,311] Therefore, mapping the comprehensive fitness landscape for

ribozymes is an important goal.

The Chen Lab previously developed a method for mapping the comprehensive fitness landscape of an RNA aptamer by *in vitro* selection,[50] with abundance used as a proxy for fitness. However, binding is qualitatively different from catalysis,[312] which involves a reaction pathway, often including covalent modification of the ribozyme, in addition to binding of the substrate and stabilization of the transition state. Furthermore, previous work has established methods for measuring affinity constants, ribozyme reaction rates, and RNA processing and thermodynamic stability by high-throughput sequencing, raising the prospect of mapping the landscape in terms of affinity or activity.[31,56,60–63,313–315] Although prior studies measuring chemical activity were applied to small populations or sparse samples of sequence space, these studies, combined with the ability to map a comprehensive fitness landscape, point toward the possibility of mapping the comprehensive chemical activity landscape for ribozymes.

This current work uses this combined approach, termed SCAPE (sequencing to measure catalytic activity paired with *in vitro* evolution), to map a comprehensive ribozyme activity landscape. We focus on an activity that would be foundational to protein translation, perhaps the most impressive invention of the RNA World. Despite its importance, the emergence of protein translation is poorly understood. A key activity is the covalent attachment of specific amino acids to specific tRNAs, which establishes the biophysical information content of the "second genetic code".[316] In modern biology, this attachment is catalyzed by aminoacyl-tRNA synthetases, but self-aminoacylating ribozymes could have been the original basis of the tRNA / synthetase system. Ribozymes that react with aminoacyl adenylates or other activated substrates have been discovered,[197,198,212,317,318] illustrating the ability of ribozymes to catalyze formation of aminoacyl-RNAs, although the substrates studied previously are prebiotically implausible or highly unstable. In contrast, *N*-carboxyanhydrides (NCAs) and the related 5(4*H*)-oxazolones can be produced

from amino acids (or peptides) by multiple prebiotically plausible reaction pathways (e.g., with carbonyl sulfide,[319] cyanate,[320] or cyanamide[321] as activating agents). These compounds react with amino acids to form peptides,[322] and therefore have been proposed as a prebiotic form of chemically activated amino acids. At high concentration, NCAs and 5(4*H*)-oxazolones react with phosphate esters, including nucleotides, to form aminoacyl-RNA mixed anhydrides in low yield,[323–327] suggesting this reaction as a candidate for ribozyme catalysis. Use of 5(4*H*)-oxazolones avoids uncontrolled polymerization in comparison to NCAs, making oxazolones a practical and prebiotically relevant substrate for *in vitro* selection. Thus, SCAPE is employed to map the catalytic activity landscape for ribozymes that self-aminoacylate using a prebiotically plausible form of chemical activation and the evolutionary and mechanistic implications of the empirically determined ribozyme landscape are analyzed.

5.3 Research Strategy

The SCAPE strategy begins with a population of molecules containing a randomized central region of 21 nt flanked by two constant regions used for PCR amplification (total length = 71 nt). In a first step, this library is subjected to *in vitro* selection for aminoacylation activity to isolate the ribozymes. In a second step to assay the ribozymes' activities, a pool of the selected molecules that includes many ($\sim 10^4$ to 10^5) different active sequences is allowed to react with various concentrations of substrate, and the products are isolated and sequenced on the Illumina platform. The sequencing output is used to quantify reaction products[313] and thereby measure the catalytic rates of potentially hundreds of thousands of sequences in parallel. This second step is referred to here as kinetic sequencing (k-Seq).

5.4 Selection of Aminoacylation Ribozymes

Beginning with a pool of random-sequence RNAs (central random region length $L = 21$) with high coverage of sequence space (∼70-99.99% coverage), six rounds of *in vitro* selection for aminoacylation activity were conducted (Figure 5.1A). In each round, the RNA pool was reacted with a biotinylated tyrosine analog, biotinyl-Tyr(Me)-oxazolone (BYO). RNAs that react with BYO become covalently attached to the biotin tag, allowing their isolation by binding to streptavidin beads. These RNAs are reverse-transcribed and amplified by PCR, providing templates for the next round of selection and amplification. The progress of the selection was followed by high-throughput sequencing, which yielded $2x10^6 - 1x10^7$ sequence reads per round of selection. Two replicates of the selection were performed (RS1 and RS2). Analysis was conducted using RS1, with data from RS2 used to confirm reproducibility of the selection.

For each round, sequences were first clustered into families using a maximum edit distance of 3 mutations (substitutions, insertions, or deletions) from the center sequence, which was defined as the sequence of highest abundance in the family. Sequence families could be identified starting in Round 4 (Figure 5.1B). The 20 ribozyme families of highest center abundance identified in the RS1, Round 5 pool were compared manually to identify conserved sequence motifs. The top 20 families comprised 80% of sequence reads by Round 6 and were consistent in RS1 and RS2. These 20 families could be characterized by one of three distinct motifs, numbered as Motif 1, 2, and 3. Motif 1 contained the shortest conserved region (Figure 5.1C) and the greatest number of unique sequences contained Motif 1. This motif could be further categorized into three submotifs (1A, 1B, 1C) based on differences in the conserved region, with 14 of the top 20 families containing Motifs 1A or 1B. Motif 2 characterized fewer unique sequences than Motif 1, but more than Motifs 1A, 1B, or 1C. Motif 2 also characterized Family 2.1, the most abundant

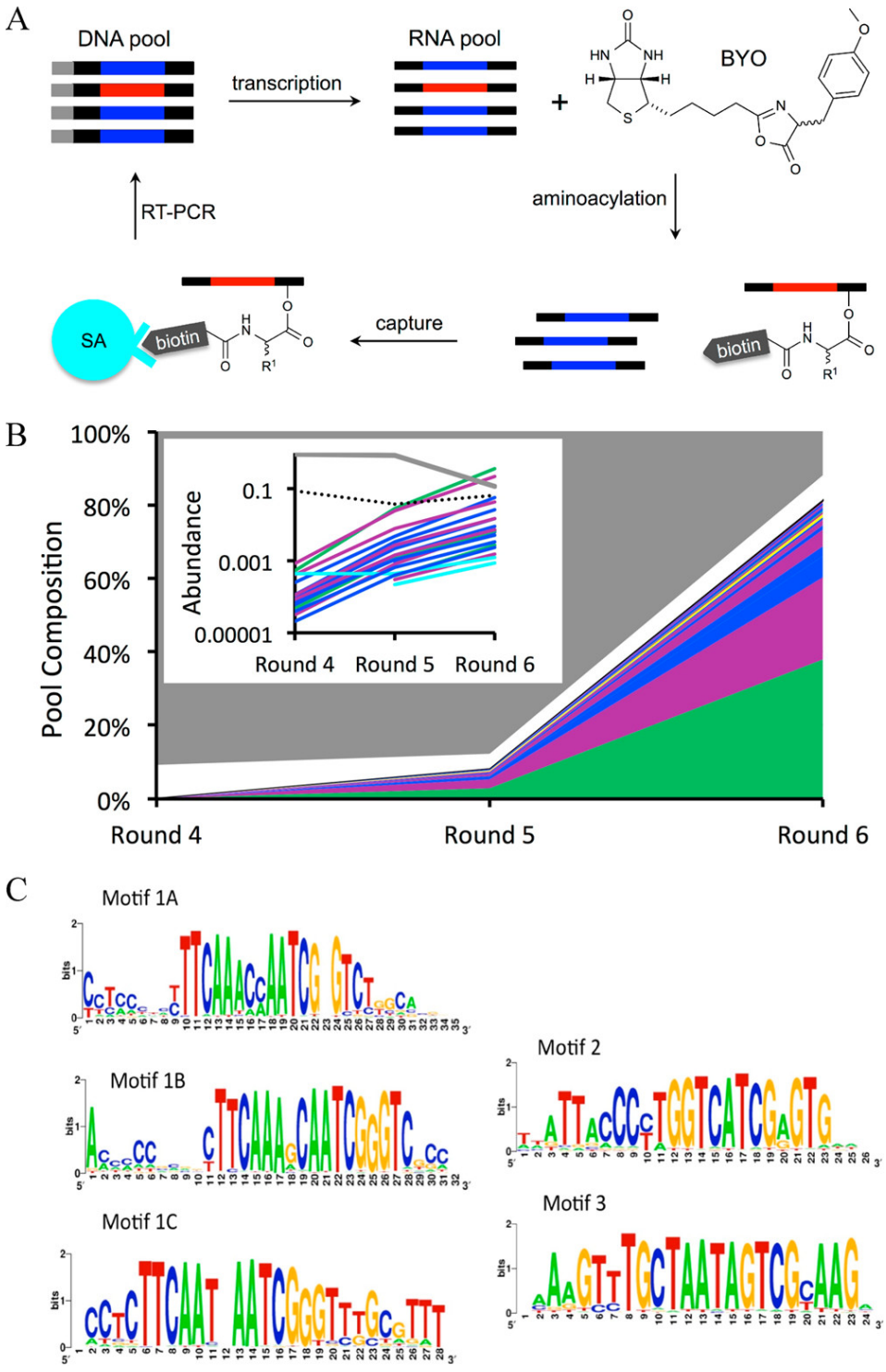

Figure 5.1: ***In vitro*** **selection for aminoacylation ribozymes.**

Figure 5.1: ***In vitro* selection for aminoacylation ribozymes.** (A) Selection began with DNA templates containing a transcription promoter (gray) and a central region of 21 random-sequence residues (red or blue) flanked by constant regions (black). These templates were transcribed into RNA and incubated with BYO. Aminoacylated RNAs (red) were isolated using streptavidin beads and amplified by RT-PCR for the next round of selection. (B) Pool composition over Rounds 4-6 after clustering. The top 20 families are indicated in non-neutral colors; gray corresponds to unclustered sequences; white corresponds to families with rank by abundance >20. Multiple families from submotif 1A (purple), 1B (dark blue), 1C (cyan), Motif 2 (green) and Motif 3 (yellow) are shown. Inset: Abundance of the top 20 families in Rounds 4-6 (same color scheme, except that the dotted black line corresponds to families of rank >20). (C) SeqLogo representations of the motifs.

family in the pool. Of these motifs, Motif 3 was found in the smallest fraction of the pool and characterized the fewest unique sequences.

5.5 Kinetic Sequencing (k-Seq)

The rate constants of the selected ribozymes were determined by a massively parallel assay (kinetic sequencing, or k-Seq; Figure 5.2A). In a gel-based assay to measure the rate constant of aminoacylation, a single RNA sequence was mixed with BYO and product formation was monitored by gel shift of the RNA in the presence of streptavidin. In the k-Seq assay, a heterogeneous pool obtained from *in vitro* selection, which contained many different RNA sequences, was reacted with BYO and the aminoacylated RNAs were isolated using streptavidin beads. These RNAs were analyzed by high-throughput sequencing (HTS), yielding the relative abundance of each sequence in the products, which were converted to absolute concentrations by comparison to a standard of known concentration in the product pool. Rate constants (k_s for sequence s) and maximum amplitude of reaction (A_s) in both assays were obtained from the dependence of product formation on the concentration of BYO. k-Seq estimates for activity could be obtained for $8.9x10^6$ sequences, but the majority of sequences were present at low abundance

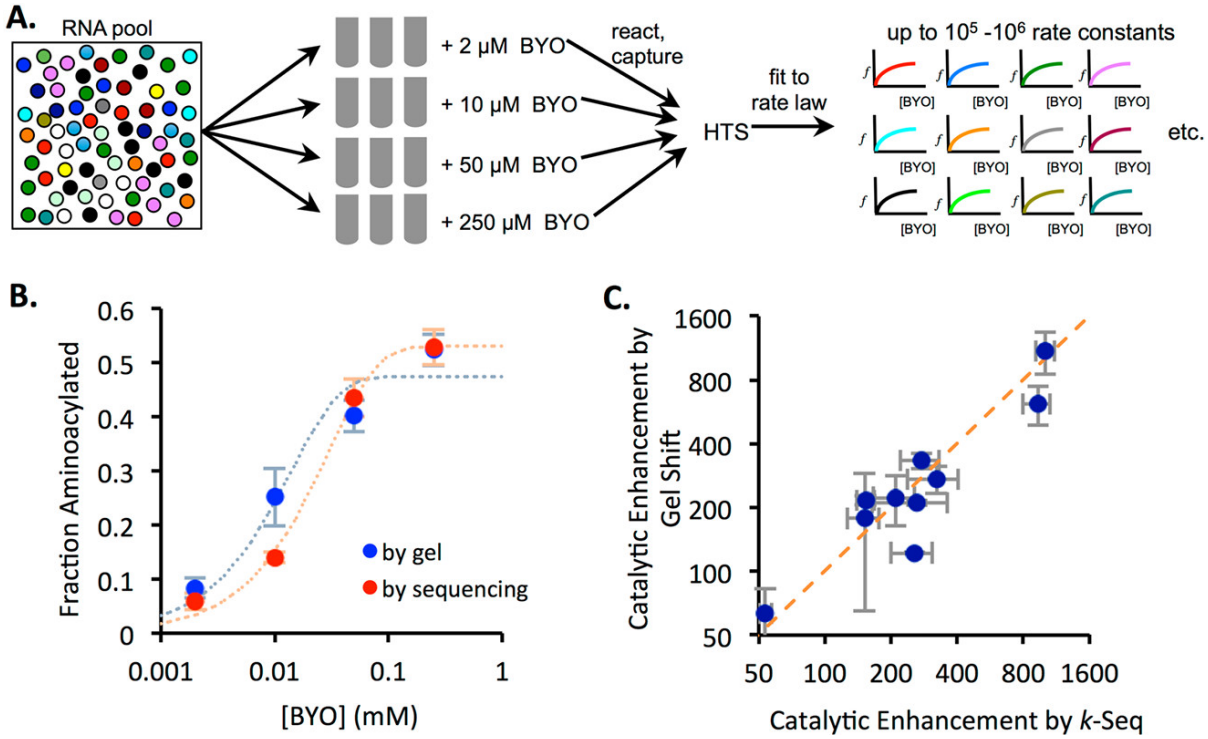

Figure 5.2: **Kinetic characteristics of ribozymes.** In k-Seq, an RNA pool enriched for active ribozymes is reacted at multiple BYO concentrations, in triplicate. Captured RNA is then reverse-transcribed and sequenced. Activity curves are constructed for sequences detected in the enriched pool. (B) Aminoacylation at various [BYO] for ribozyme S-2.1-a observed by both gel shift and k-Seq. Error bars correspond to standard deviation among triplicates. (C) Correlation between catalytic enhancement of ten ribozymes, measured by gel shift assay and k-Seq. Error bars correspond to standard deviation among triplicates (k-Seq) or 2-3 replicates (gel assay) ($R^2 = 0.87$). Dotted orange line indicates line of unity.

and correspond to low activity. $\sim 10^5$ unique sequences, out of $\sim 4^{21}$ possibilities, were found to have activity >10-fold above the non-catalytic background rate (i.e., catalytic enhancement $r_s > 10$, where $r_s = \frac{k_s A_s}{k_0 A_0}$, and k_0 and A_0 are the rate constant and amplitude of reaction of the non-catalyzed reaction, measured in the randomized RNA pool).

To determine how well k-Seq results corresponded to results of the standard assay, ten sequences were chosen that are close to the consensus sequences of the high- or medium-activity families (with all five motifs and submotifs represented) and aminoacylation activity was measured by the gel-shift assay.[328] Rate constants determined from

k-Seq matched well with gel-shift measurements (Figure 5.2B-C). All k-Seq and gel-shift measurements were performed in triplicate and the standard error was similar between k-Seq and gel-shift measurements. Measurement error during k-Seq decreased as sequence read abundance increased, as expected for stochastic noise. For most sequences with count >10, and nearly all sequences with count >100, the noise of k-Seq measurements appeared to be within a factor of 2.

High-activity sequences (e.g., the center of Family 2.1, with $r_{S-2.1-a} = 1010$ and $k_s = 779 \pm 21$ $M^{-1}min^{-1}$) exhibit saturating kinetics from k-Seq, providing both the rate constant (k_s) and the maximum amplitude of reaction (A_s). However, the reaction for lower activity sequences (approximately $k_s < 20$ $M^{-1}min^{-1}$) appears linear under the conditions tested, so that k_s and A_s are difficult to estimate separately using these data; instead the combined parameter k_sA_s can be estimated.

5.6 Aminoacylation Site and True Catalytic Enhancement

The most highly abundant sequences from each major motif were chosen (S-1A.1-a, S-1B.1-a, S-2.1-a, S-3.1-a; see Section 4.6 for sequence nomenclature) for characterization of the reactive site. Identification of the reactive site was performed in two steps. First, reverse transcription is known to be sensitive to 2' adducts, such that stalled products can be used to identify the sites of 2' acylation.[330,331] The putative ribozymes were ligated to a 3' adapter to test for stalling of reverse transcription along the entire length of the ribozyme. Stalling resulted in a truncated product whose length, determined by gel electrophoresis, suggested a likely site of aminoacylation (Figure 5.3). Second, the nucleophilic importance of the 2'OH at the candidate site was verified by testing the

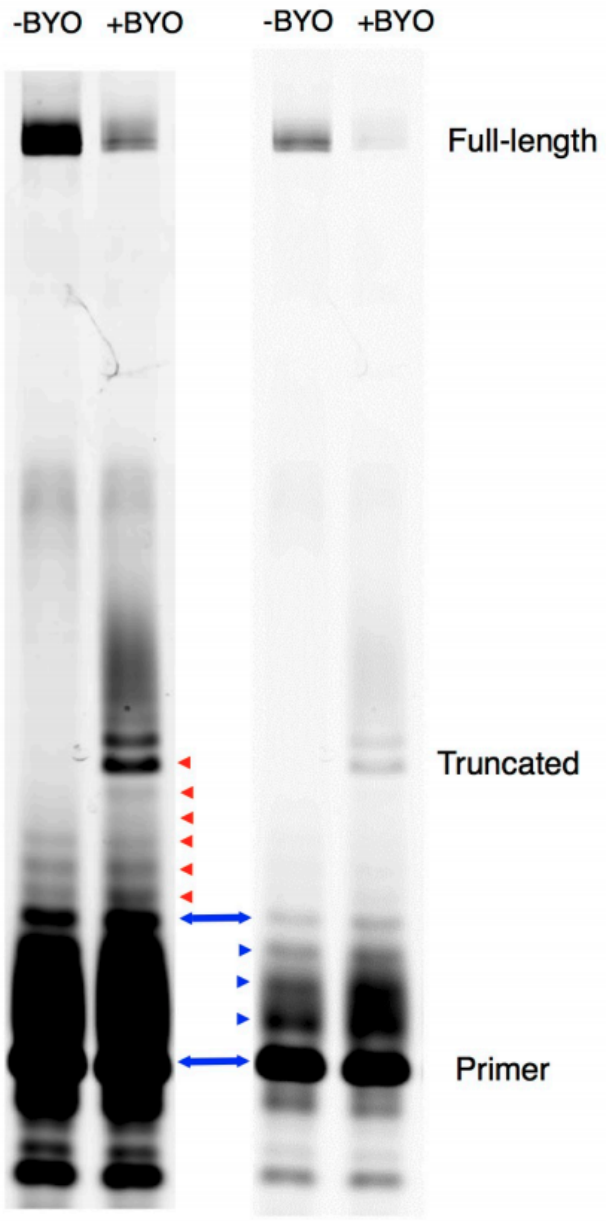

Figure 5.3: **Identification of aminoacylation site by reverse transcriptase stalling.** The likely site of BYO modification was identified by stalling of reverse transcription, resulting in a truncated product. Here, high (left) and low (right) contrast versions of a gel show single-nucleotide resolution of banding used to identify the suspected site of aminoacylation. Blue marks indicate bands terminating within the ligated adapter; red marks indicate bands terminating within the ribozyme sequence. In this case, for ribozyme S-1A.1-a, the main reverse transcription stall occurs immediately before the 7th position from the end of ribozyme sequence, implicating G65 as the site of aminoacylation.

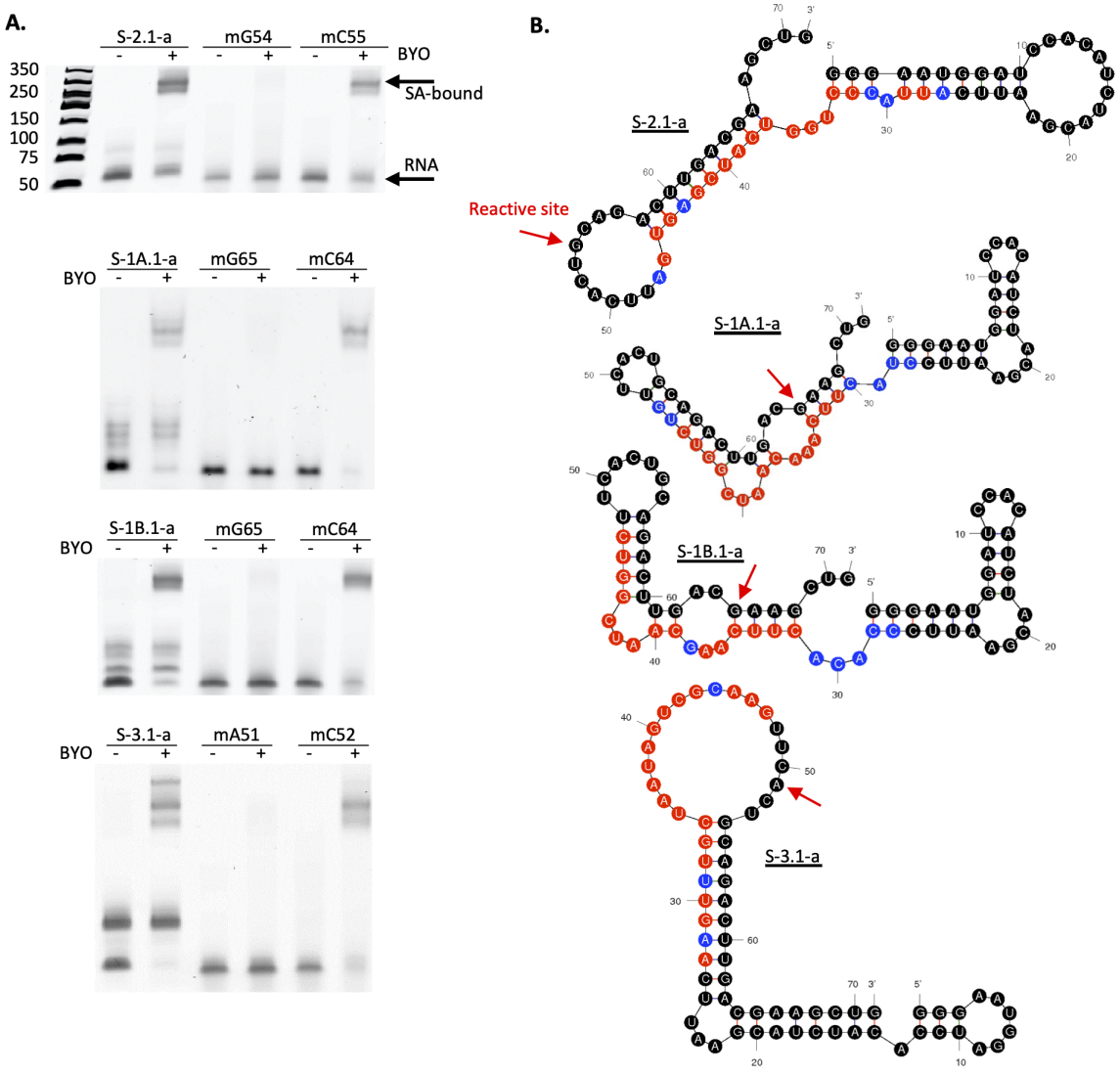

Figure 5.4: **Verification and position of aminoacylation sites.** (A) Streptavidin (SA) gel shift assays for ribozymes S-2.1-a, S-1A.1-a, S-1B.1-a, and S-3.1-a. The site of aminoacylation predicted by reverse transcriptase stalling (Figure 5.3) was verified by loss of activity upon 2'-O-methylation. 2'-O-Methylation of an adjacent site did not show loss of activity. (B) Minimum free energy secondary structures for the sequences indicated, as predicted by mfold.[329] Note that these structures have not been experimentally verified in this work. Black denotes constant regions. Sites in the selected region conserved with information content <1 bit are shown in blue; sites with information content >1 bit are shown in red (also see Figure 5.1C). Red arrows indicate the observed aminoacylation site.

activity of a synthetic RNA modified at this position by 2'-O-methylation. In each case, a control synthetic RNA that was instead modified at an adjacent position was also tested. Blocking of the candidate site (but not the control site) by O-methylation is expected to abolish the reaction. For all sequences tested, the results were consistent with aminoacylation at a specific internal 2'-OH position within the 3' constant region of the sequence (Figure 5.4). While the reactive site was conserved for sequences from the same major motif (e.g. S-1A.1-a and S-1B.1-a, both from Motif 1), the site differed among sequences from the three major motifs, indicating that ribozymes with different motifs utilize different detailed reaction mechanisms.

Note that the catalytic enhancement r_s calculated here underestimates the true catalytic enhancement at the modified site. The potential nucleophilic sites include 70 internal 2'-OH groups, the vicinal diol at the 3' end, and the 5'-triphosphate. Thus the uncatalyzed reaction rate at a particular site is at least 73-fold lower than $k_0 A_0$, which was measured for the entire RNA. In addition, previous work on oxazolone modification of small RNA oligonucleotide models indicates that the vicinal diol and terminal phosphates (2', 3', or 5') are strongly preferred as nucleophiles, with no detectable reactivity at internal 2'-OH sites.[326,327] In contrast, all ribozymes tested, representing each motif (1A, 1B, 2, 3), were modified at an internal 2'-OH. Therefore, the true catalytic enhancement provided by these ribozymes at a specific internal 2'-OH is likely to be at least 700-fold greater than the r_s as reported here.

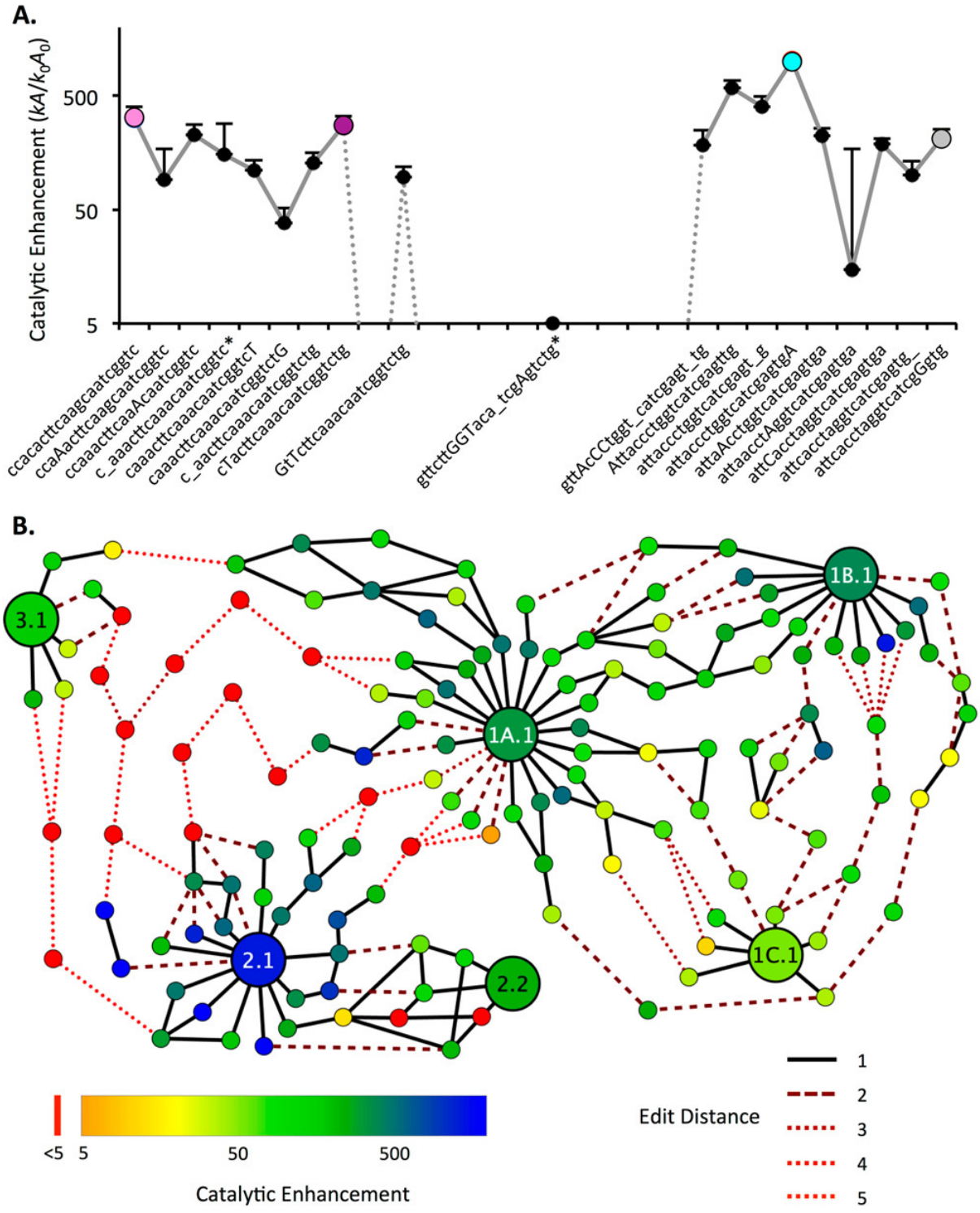

Figure 5.5: **Evolutionary pathways for aminoacylation ribozymes.**

Figure 5.5: **Evolutionary pathways for aminoacylation ribozymes.** (A) Catalytic enhancement along a best pathway discovered from the center of Family 1B.1 (pink, S-1B.1-a), to 1A.1 (purple, S-1A.1-a), to 2.1 (cyan, S-2.1-a), to 2.2 (gray, S-2.2-a). Capital letters denote sequence positions changing at each step; underscore indicates a deletion. A large drop in activity is required for several mutations between Motif 1 and Motif 2. Error bars are standard deviation from triplicate measurements (only top bar is shown). Asterisk (*) indicates a sequence that was found in only one replicate (RS1). (B) Evolutionary network displaying the 10 best pathways discovered between the centers of six key families (1A.1, 1B.1, 1C.1, 2.1, 2.2, and 3.1) representing each motif and submotif and the two most active centers from Motif 2. Each node is an individual sequence with activity measured by *k*-Seq indicated by color (see legend; red indicates activity at or below the baseline rate). The lines indicate mutational distance between sequences (solid black line = 1 mutation). Dotted lines indicate sequences at baseline activity (see legend). The majority (67%) of the edits along these pathways are substitutions; the remainder are indels.

5.7 Evolutionary Pathways between Ribozyme Motifs

A series of single mutations defines an evolutionary pathway between two sequences. Although there are very many conceivable pathways, many of these include intermediate sequences of low fitness. Under selection, such fitness valleys represent dead ends that effectively block evolution. An open question is whether viable evolutionary pathways exist between different sequences that catalyze the same reaction. Using the chemical activity data from *k*-Seq, we searched for viable evolutionary pathways between center sequences of the major ribozyme families (Figure 5.5).

A broad network of pathways was found among Families 1A.1, 1B.1, and 1C.1, with a <10-fold catalytic rate decrement at the lowest point of the best pathways. Thus the families of Motif 1 form a 'plateau' in the chemical activity landscape, corresponding to the small size of Motif 1. Similarly, viable pathways exist between the top two families of Motif 2. Although Motif 2 encompasses a smaller region of sequence space compared to Motif 1 due to a larger conserved region, Motif 2 contains the global optimum of the

landscape. Viable pathways were not found between families of Motif 3, likely due to the small number of unique sequences in this motif. Within Motifs 1 and 2, the number of viable pathways was relatively small, suggesting that evolution within a motif would be fairly reproducible.

However, evolutionary pathways between motifs appeared strikingly different. The only pathways that could be constructed between different motifs contain fitness losses down to baseline activity, with multiple mutational steps occurring at near baseline activity. The closest apposition of motifs was a pathway between Family 3.1 and Family 1A.1, which involves 5 consecutive intermediates expected to have baseline activity (i.e., $r \sim 10^3$-fold less than $r_{S-2.1-a}$). The global optimum (Family 2.1) is especially isolated, with >10 mutations at baseline activity required along any pathway toward a different motif. These pathways would not be viable under selection, indicating that optimization of activity over the global fitness landscape would be frustrated.

5.8 Discussion

In the SCAPE method, a ribozyme fitness landscape can be mapped in two steps. First, the vast majority of inactive sequences are removed from the pool through *in vitro* selection. Second, the catalytic activities of the remaining sequences are directly assayed by kinetic sequencing (k-Seq). In this case, k-Seq yielded estimates for $\sim 10^5$ unique sequences (a number that in general depends on pool diversity, activity distribution, and sequencing depth). Using SCAPE, the first comprehensive fitness landscape was mapped for catalytic activity, subject to the following caveats. First, in order to survive the selection, sequences must be both catalytically active and replicable (by transcription and RT-PCR). Because RT stalls at the aminoacylated site, ribozymes that aminoacylate within the randomized region are presumably disfavored. Consistent with this, all of

the ribozymes tested here react within the 3' constant region, as modification does not preclude primer binding. Sequences may also have been lost during selection for other reasons (e.g., transcription or RT-PCR bias, and genetic drift in early rounds). While such occasional losses might affect the details of evolutionary pathways, they would likely not affect the overall findings given the extensive fitness valleys found. Alternatively, if the starting library were relatively small ($\sim 10^6$ sequences), k-Seq alone (without selection) could be used to build a comprehensive map of the library; advances in sequencing technologies may push this bound further. Second, fitness is measured in the specific environment applied, in this case for aminoacylation activity under the chemical conditions of the selection. How the environment would affect the fitness landscape, and how aminoacylation activity might relate to the replicative fitness of an RNA World organism (a variety of relationships are possible[332–336]), are difficult to address at present.

Ribozymes were discovered that self-aminoacylate using a $5(4H)$-oxazolone, a key step toward the genetic code. The best ribozyme found here has a rate constant comparable to that of ribozymes obtained using a biologically derived aminoacyl adenylate,[197,318] indicating that these reactions could proceed efficiently even with only prebiotic substrates. Interestingly, all ribozyme families discovered here react at an internal 2'-OH of the RNA. These sites stand in contrast to the modification of modern tRNAs at the vicinal diol (3' terminus), which is also found to be more reactive in model oligonucleotides.[326,327] It is possible that an internal reaction site facilitates establishment of multiple contacts with BYO, and the rate acceleration caused by these structural features outweighs the intrinsic reactivity of the vicinal diol. Similarly, it is unknown whether the identity of the 3' terminal sequence (CUG in this study, compared to CCA in tRNAs) may contribute to this finding. This difference raises the interesting question of whether ribozymes such as those discovered in this model system could be on the pathway toward the modern implementation of the genetic code; whether they have the evolutionary capacity to adopt a

mechanism more similar to the aminoacyl-tRNA synthetase system is currently unknown.

In addition to discovering novel ribozymes, a primary motivation for SCAPE analysis is to learn about molecular evolution by exhaustively determining the viable evolutionary pathways and networks through sequence space. While some viable pathways exist locally around an optimum, most conceivable pathways toward the global fitness optimum (Family 2.1) are blocked by extensive fitness valleys. The likely reason is that the three major motifs differ substantially in structure, as indicated by their different aminoacylation sites. It appears that the ribozyme structure cannot be changed without essentially destroying the structure of one ribozyme and building another, requiring extensive mutations at negligible activity. Such evolutionary walks would be essentially impossible while under selection for catalytic activity, frustrating optimization over the network.

This landscape can be compared to other landscapes and evolutionary pathways that have been described for functional RNAs. Extensive work on *in silico* folding of RNA sequences has predicted the existence of large neutral networks for secondary structure, in which evolutionary walks over long distances could maintain a given structure.[38,42,337] Such neutral networks would permit facile exploration of sequence space through evolution. In addition, multiple examples of ribozymes evolving to perform different functions are known.[137,249,338,339] In contrast, the previously described landscape for RNAs selected to bind GTP (based on sequence abundance rather than activity measurement) showed that the landscape consisted of several evolutionarily isolated peaks.[50] Thus, it appears that, although preservation of secondary structure could occur over a neutral network, the additional tertiary structural requirements of a functional RNA leads to a qualitative change in the nature of the evolutionary network. Such a change is analogous to the phase transition-like behavior of percolation through a network;[340] as the frequency of active nodes decreases, the network suddenly switches from highly connected, as in the case of neutral networks of RNA secondary structure, to essentially impermeable, as

observed for evolutionary networks of functional RNAs. An important caveat is that the landscape reported here was mapped under constant selection for a single catalytic activity and cannot be directly compared to evolutionary pathways leading to new functions; changing environments[341] or selection pressures may significantly alter this picture.

The phenomenon of frustration arises when competing interactions prevent overall optimization of a system, and the following discussion uses the term 'frustration' in this general sense. A classic illustration of frustration is the anti-ferromagnetic spin glass, in which energy would be minimized by antiparallel placement of neighboring electronic spins. In certain configurations (e.g., a triangle), no placement of spins can satisfy all desired constraints, leading to rugged energy landscapes.[342] An example of frustration in biology is the folding energy landscape of proteins,[343–345] where individual local molecular arrangements that minimize energy may be mutually incompatible, resulting in rugged energy landscapes and misfolded states. The analogy to frustrated spin glasses is also being explored theoretically to understand fitness landscapes,[102,346,347] gene expression networks,[348] morphological innovation,[349] and even the evolution of biological complexity.[350] These results show that the experimentally determined ribozyme activity landscape exhibits frustration, as individually beneficial mutations are often mutually incompatible, leading to ruggedness on the fitness landscape.[13,351] Walks on such energy or evolutionary landscapes are characterized by sensitivity to initial conditions, frustrated optimization, and multiple possible outcomes. It should be noted that mechanisms that favor greater genetic diversity, such as recombination, gene duplication, or epistasis among genes, could enable crossing of fitness valleys.[352,353] Recent work suggests that recombination, in particular, can occur spontaneously in pools of RNA.[354,355] The quantitative effect of such mechanisms on traversal of the fitness landscape is unknown at present. Nevertheless, in the absence of such mechanisms, the emergence of a globally optimal sequence is likely to result from chance events rather than natural selection.

Chapter 6

Error Minimization and Specificity Could Emerge in a Genetic Code as By-Products of Prebiotic Evolution

6.1 Attributions

This chapter was the result of collaboration with Yuning Shen, Ziwei Liu, Celia Blanco, and Irene A. Chen.

6.2 Introduction

The origin of life is believed to have progressed through an RNA World in which ribozymes catalyzed critical biochemical reactions.[9,10] In principle, ribozymes performing new functions could arise either by chance or by adaptation of pre-existing ribozymes having promiscuous activities. Co-option of a pre-existing sequence (i.e., exaptation) is a well-established mechanism for evolutionary innovation.[135,141,182,188,191,356] Gene duplica-

tion coupled with co-option could lead to a more complex system as the ribozymes adopt additional substrates.[357] However, the degree to which the evolution of complex systems in the RNA World would rely on chance vs. co-option is unclear.[127]

The genetic code of protein translation is one of the most complex products of the RNA World, and its emergence is considered a 'major evolutionary transition'.[358] 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.[316,359,360] However, during the emergence of protein translation itself, these functions were presumably performed by ribozymes. Indeed, evolutionary analysis of the aaRS proteins indicates that these enzymes evolved after the establishment of a primitive genetic code[358,361–363] and have heterogeneous genetic origins.[364] Several ribozymes catalyzing aminoacylation reactions have been discovered by *in vitro* selection, including self-aminoacylating RNAs.[197–200,203,212] Such ribozymes could serve as precursors to the aaRS/tRNA encoding system.

A well-documented feature of the standard genetic code is robustness to errors, i.e., that non-synonymous point mutations tend to result in amino acid substitutions that conserve biophysical properties.[129,130,286,365,366] This 'error minimization' confers a clear selective advantage as it reduces the deleterious impact of mutations on the resultant protein.[367,368] However, the standard genetic code does not appear to be particularly optimal with respect to error minimization.[369–372] This raises a fundamental open question about the origin of error minimization, namely, whether error minimization of the standard genetic code is the product of natural selection, or a serendipitous by-product of the evolution of protein translation.[130] In other words, in contrast to direct natural selection for error minimization, it is possible that expansion of an early version of the code, initially comprising a small number of amino acids, to the full set of 20 amino acids, involved an evolutionary mechanism that happened to conserve the biophysical character

of the amino acids.[12,373]

This work evaluates the evolutionary potential of self-aminoacylating ribozymes to adopt new amino acid substrates. Previously, *in vitro* selection and high-throughput sequencing were used to exhaustively search sequence space (21 nt) for self-aminoacylating ribozymes.[200] 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.[319–323,325,326,374] Three distinct, evolutionarily unrelated catalytic motifs had been discovered from the exhaustive search. Here, the co-option potential of these ribozymes is determined 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). This assay and related techniques leverage high-throughput sequencing to measure the activity of thousands of candidate sequences in a mixed pool.[62,314,375,376] The six substrates (analogs of tryptophan, phenylalanine, leucine, isoleucine, valine, and methionine) represent a range of sizes and biophysical classes (aromatic, aliphatic, sulfur-containing), as well as supposed early (Leu, Ile, Val) and late (Trp, Phe, Met) incorporations into the genetic code.[377–381] The results indicate extensive opportunities for co-option to incorporate new substrates into the system. In addition, two major by-products of evolution of these ribozymes are described. 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 biophysically similar amino acids, suggesting that expansion of the code by co-option would incorporate a biophysically 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.

6.3 Aminoacylation Substrates and Design of the Ribozyme 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: tryptophanyl (BWO), phenylalanyl (BFO), leucyl (BLO), isoleucyl (BIO), valyl (BVO), and methionyl (BMO) (Figure 6.1A). Compounds were synthesized using previously described methods[200] and verified by NMR spectroscopy. An initial test by a gel shift assay at high substrate concentration (500 μM) indicated that each oxazolone served as substrate for at least one ribozyme tested, although the tested ribozymes - S-1A.1-a and S-2.1-a - differed in selectivity (Figure 6.1B). To study the cross-reactivity of these ribozymes and their mutants systematically, pools of sequence variants were designed to explore the sequence space around the major ribozyme families obtained from the selection on BYO (Table 4.3). 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 containing these motifs had comprised $\sim$80% of the selected pool.[200] Sequencing of the variant pool showed that it included 13.5% of the unique sequences (having abundance $\geq 10^{-6}$) from the originally selected pool. Thus, the variant pool, based on these five ribozyme families, was designed to be representative of ribozymes having aminoacylation activity.

Because the ribozymes had been identified through selection with substrate BYO, it was possible that entirely new ribozyme families might react with different BXO sub-

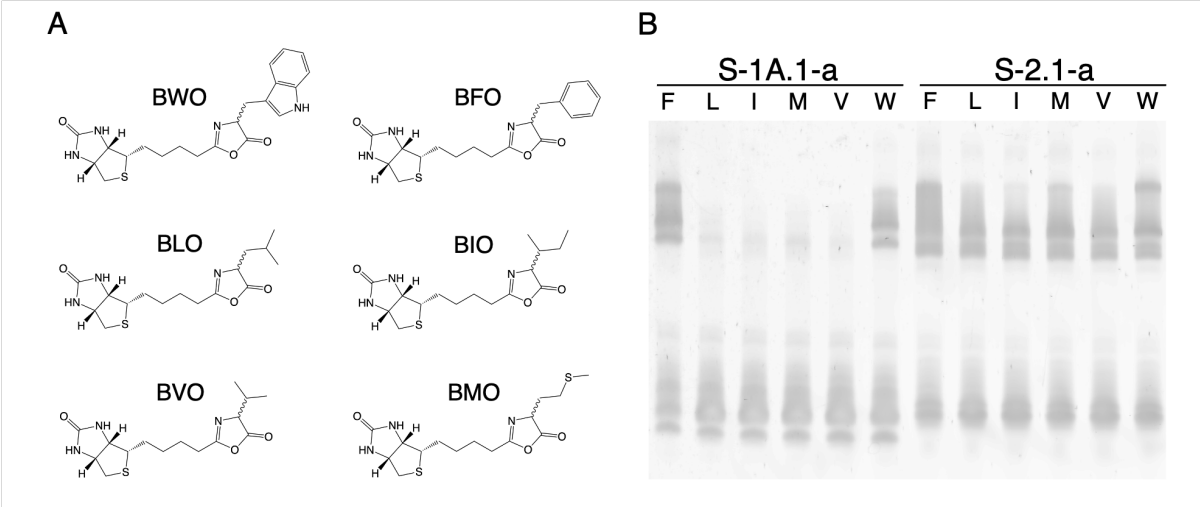

Figure 6.1: **Aminoacylation activity of two ribozymes with BXO substrates.** (A) Biotinyl aminoacyl oxazolones (BXO) used in this study: tryptophanyl (BWO), phenylalanyl (BFO), leucyl (BLO), isoleucyl (BIO), valyl (BVO), and methionyl (BMO). (B) Aminoacylation activity of two ribozymes (S-1A.1-a, the center of Family 1A.1, and S-2.1-a, the center of Family 2.1) with BXO substrates analyzed by streptavidin gel shift (X = F, L, I, M, V, or W, as indicated). Reactions were conducted for 90 min at 500 μM BXO. The reacted RNA is detected by its slower migration through the gel due to complexation with streptavidin. Multiple bands may be caused by the presence of multiple conformers or streptavidin oligomers.

strates. To assess this possibility, *in vitro* selections for self-aminoacylating ribozymes were performed for two of the new 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 been previously identified in the earlier BYO selection (Figure 6.2). 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 (Figure 6.1B). These results indicate that the designed pool of variants would probe the major motifs of the active sequence space for these substrates.

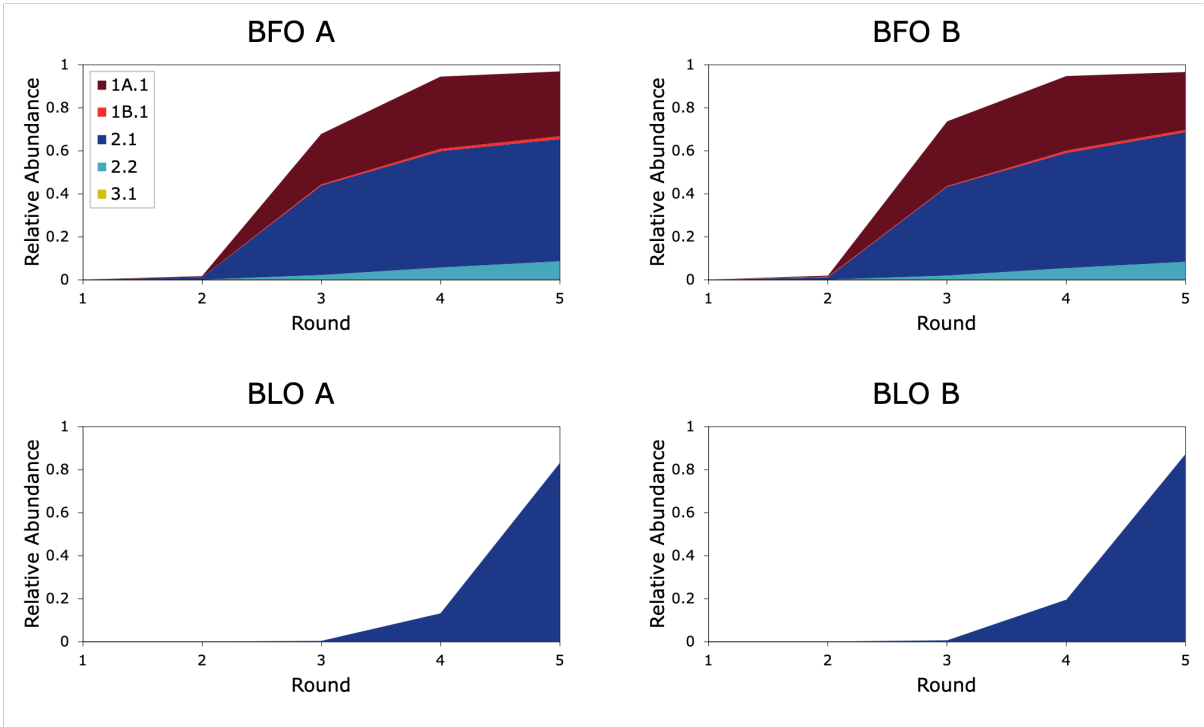

Figure 6.2: **Convergence of selections for aminoacylating ribozymes with BFO and BLO.** Duplicate selections (A and B) for aminoacylating ribozymes with BFO and BLO result in convergence on the same primary families identified previously under selection with BYO.[200] Note that only Motif 2 emerges in substantial fraction following selection with BLO, while the BFO selection yields families from multiple Motifs.

6.4 Cross-Reaction of Self-Aminoacylating Ribozymes with Alternative Aminoacyl 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).[200,301,302] During k-Seq, a pool containing thousands of candidate ribozymes is reacted with a substrate at multiple concentrations. The reacted molecules, having been biotinylated through reaction, are isolated by streptavidin binding, and then sequenced on the Illumina platform. Quantitation of the reacted fraction allows fitting to a kinetic model to determine ribozyme activity. Data obtained from this method correlate well with traditional bio-

chemical assays, provided a sufficient number of sequencing counts, and confidence intervals of the measurements are obtained by experimental replicates and bootstrapping.[301] In each k-Seq experiment here, 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 μM in triplicate. Reaction data were fit to a pseudo-first-order kinetic model ($F_s^{BXO} = A_s(1 - e^{(-k_s\alpha[BXO]t)})$) 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. Although data over a fixed concentration range are inadequate for separately estimating k_s and A_s for low activity ribozymes, the product k_sA_s can be accurately estimated across a wide range of activities, due to the inverse correlation of k_s and A_s during curve fitting[200,301] (Figure 6.3). The product k_sA_s reflects ribozyme activity at non-saturating conditions and was used in the following analyses. The data yielded k_sA_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 (Figure 6.4).

k-Seq cannot distinguish catalyzed and uncatalyzed (background) reactions, and thus both reactions are measured. To determine catalytic enhancement, i.e., the ratio of catalyzed to background reaction rates, the rate of the background reaction for BFO was measured by gel shift assay with the randomized RNA library. The background rate was 0.55 ± 0.18 $M^{-1}min^{-1}$ ($\mu \pm \sigma$), which is similar to that measured previously for BYO (0.65 ± 0.28 $M^{-1}min^{-1}$).[200] Comparing to the frequency distribution of k_sA_s measured by k-Seq (Figures 6.4 and 4.10 and Table 4.4), 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

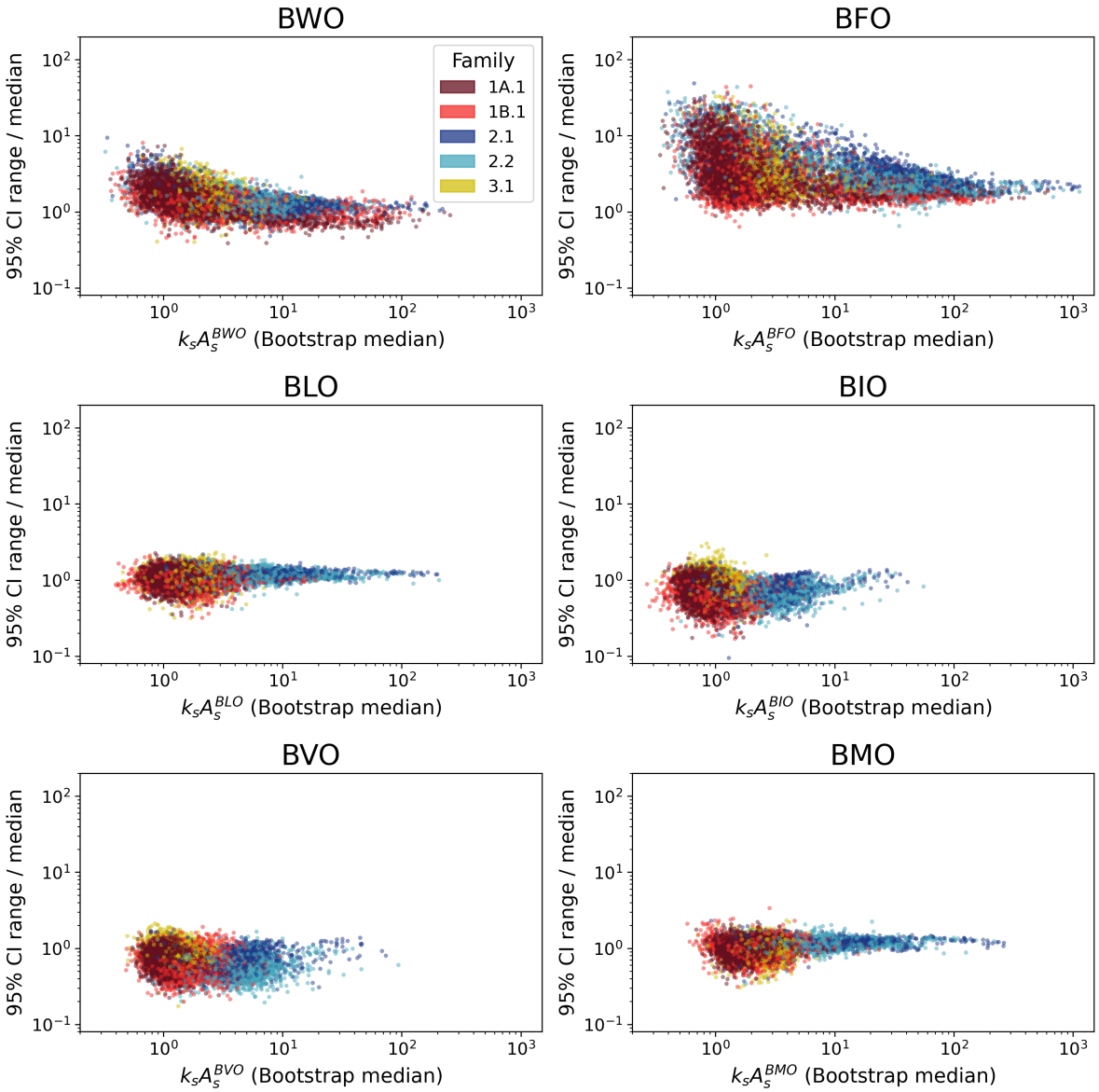

Figure 6.3: **Precision of *k*-Seq estimates of $k_s A_s$.** Bootstrapping ($N = 1000$) was used to estimate 95% confidence intervals (95% CI range, i.e. 97.5%-2.5%) and medians as previously described.[301] Confidence intervals were normalized to the medians estimated from bootstrapping. It can be seen that normalized confidence intervals are generally within one order of magnitude.

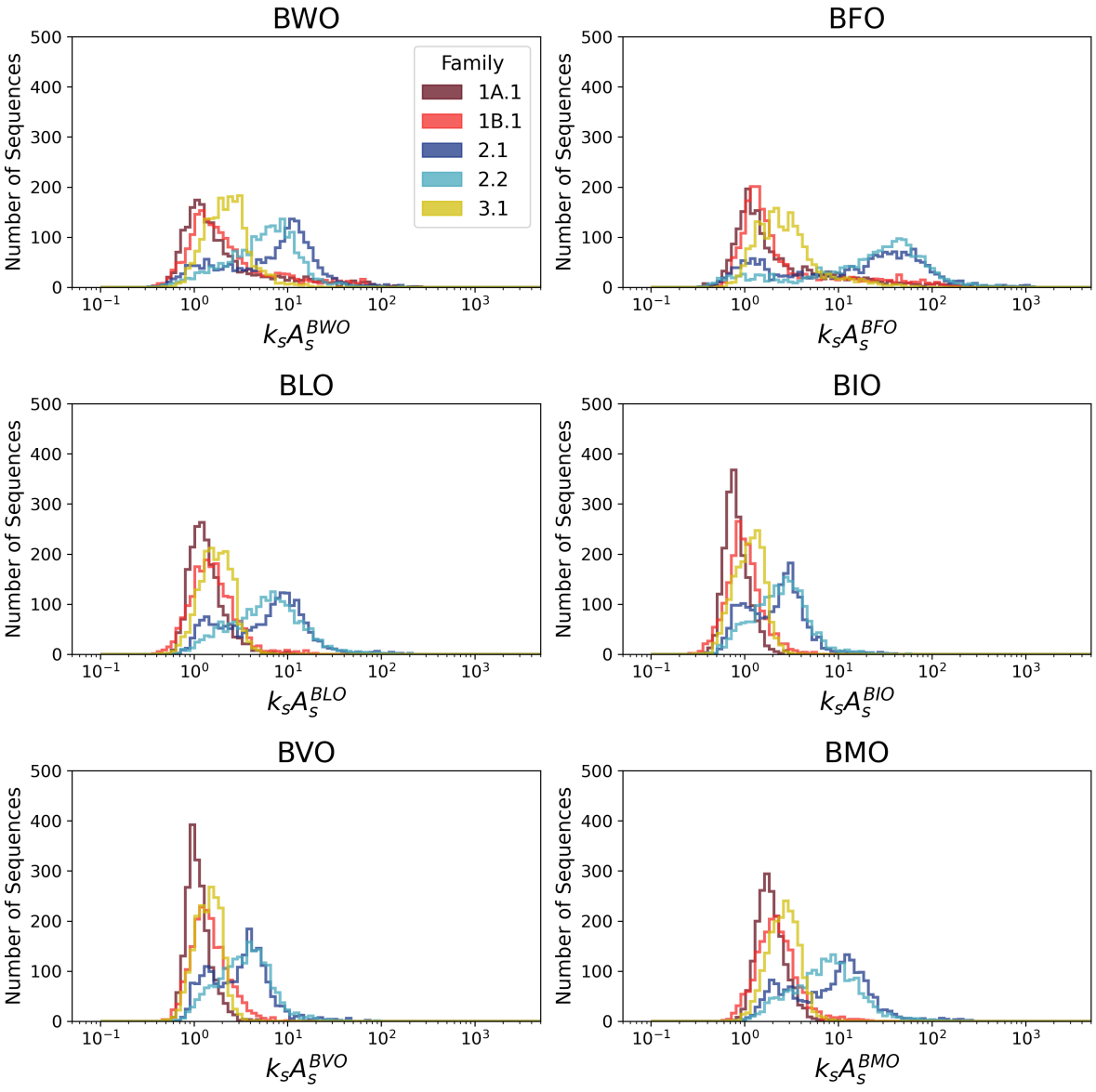

Figure 6.4: **Histograms of ribozyme $k_s A_s$ values with each substrate, for each family.**

with observations that individual Motif 1 ribozymes display little activity with some substrates at high concentration when analyzed by a gel-shift assay (Figure 6.1B). 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 (Figure 4.10 and Table 4.4). 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 (Figures 6.5 and 6.6). 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. 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.

6.5 Ribozyme Families Distinguish Different Biophysical Features of Substrate Side Chains

To assess the activity and specificity of individual ribozymes for each substrate, catalytic enhancement values for different substrates were compared in a pairwise fashion (Figure 6.7 and Figure 6.13). All families displayed a high degree of correlation among

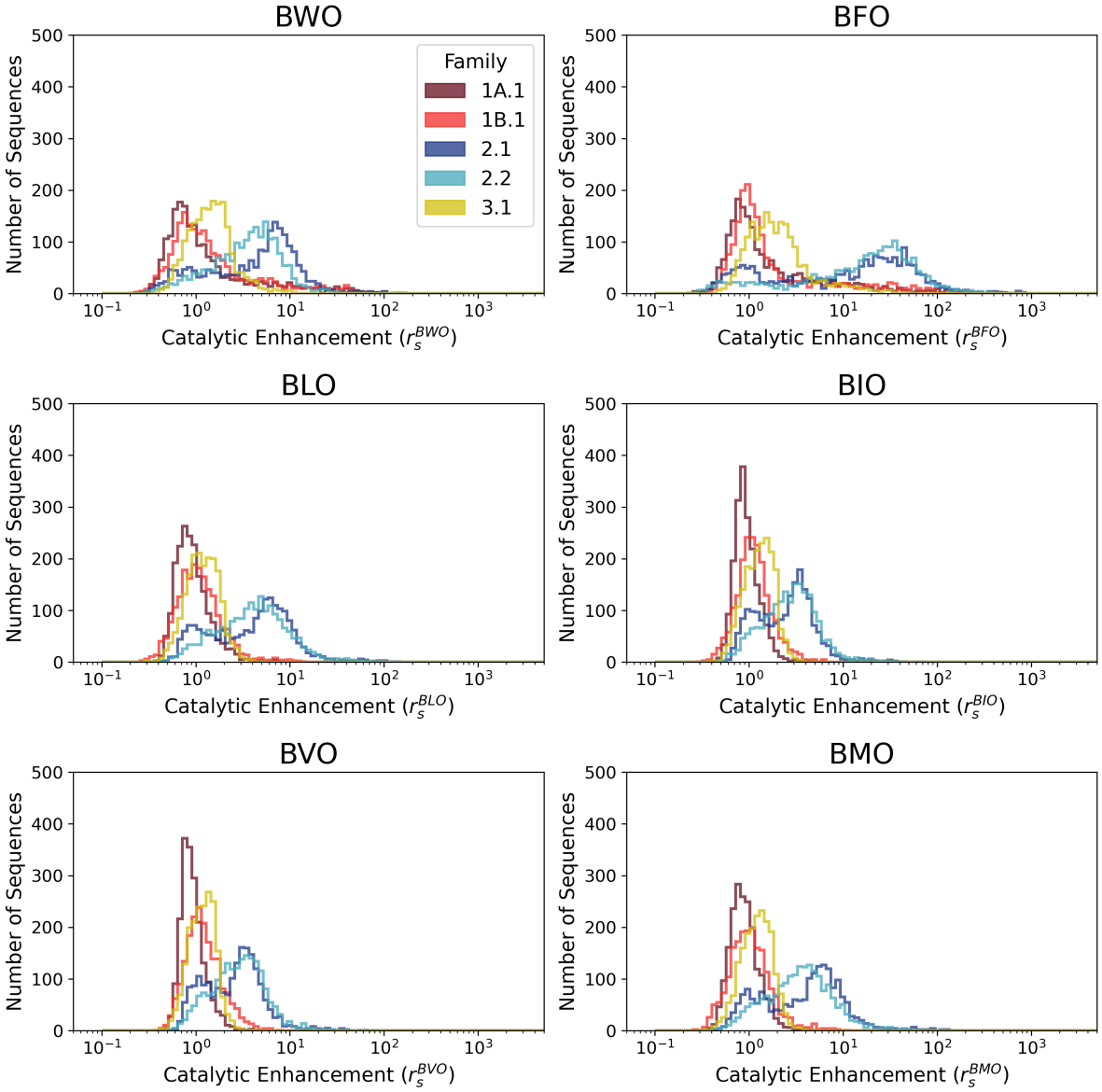

Figure 6.5: **Catalytic enhancement of ribozyme families for different substrates.** Histograms of catalytic enhancement values ($r_s = \frac{k_s A_s}{k_0 A_0}$) with each BXO substrate, measured by k-Seq, for ribozymes in Family 1A.1, 1B.1, 2.1, 2.2, and 3.1. While many ribozyme mutants in Motif 2 families have activity on each substrate tested, many ribozyme sequences containing Motif 1 or 3 are inactive.

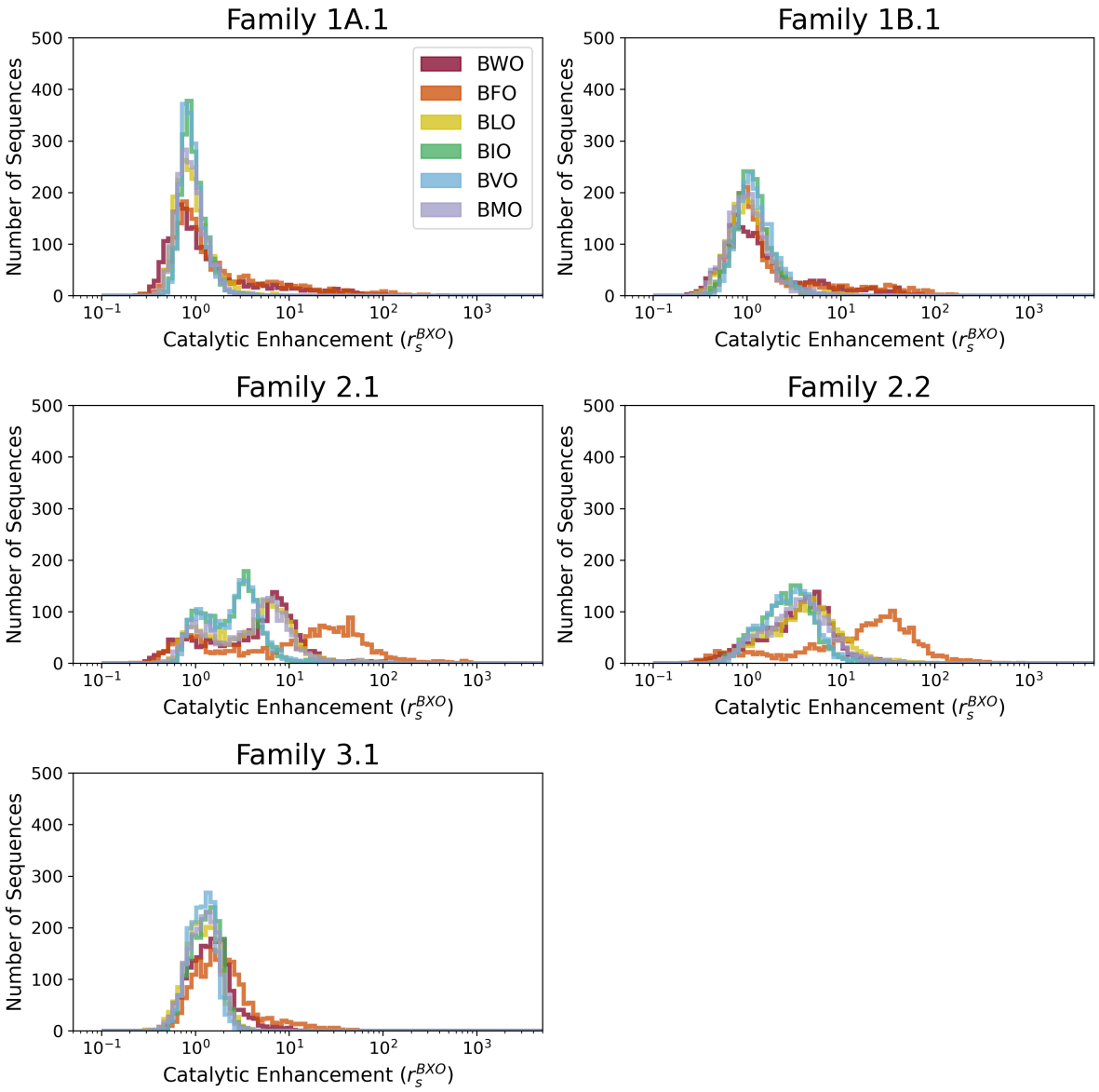

Figure 6.6: **Catalytic enhancement of ribozyme families for different substrates.** Histograms of catalytic enhancement values (r_s) for each family with each substrate. The same data are represented in Figure 6.5 (but presented by family instead of by substrate).

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)) (Figure 6.8A). The high correlations indicated that few sequences exhibit large activity differences between amino acids within the same biophysical class.

However, when comparing amino acids of different classes (i.e., aromatic vs. non-aromatic), strong correlations were only observed for Families 2.1 and 2.2, indicating that the effects of mutations in Motif 2 ribozymes 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 (Figure 6.7), resulting in lower correlations between activity on aromatic and non-aromatic side chains (Figure 6.8A). These preferences were also captured by the slopes on the correlation plots (Figure 6.8B), which confirm that Motif 1 ribozymes strongly favor aromatic side chains, while Motif 2 ribozymes demonstrate less pronounced preferences, and Motif 3 ribozymes display an intermediate strength of preference. While less pronounced than for Motif 1, some preferences were still observed for Motif 2 ribozymes, in which BFO was most preferred, BWO, BLO, and BMO were weakly preferred, and BIO and BVO were disfavored. Interestingly, BIO and BVO, in contrast to the other side chains, are both branched at the α-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 biophysical features of the side chains.

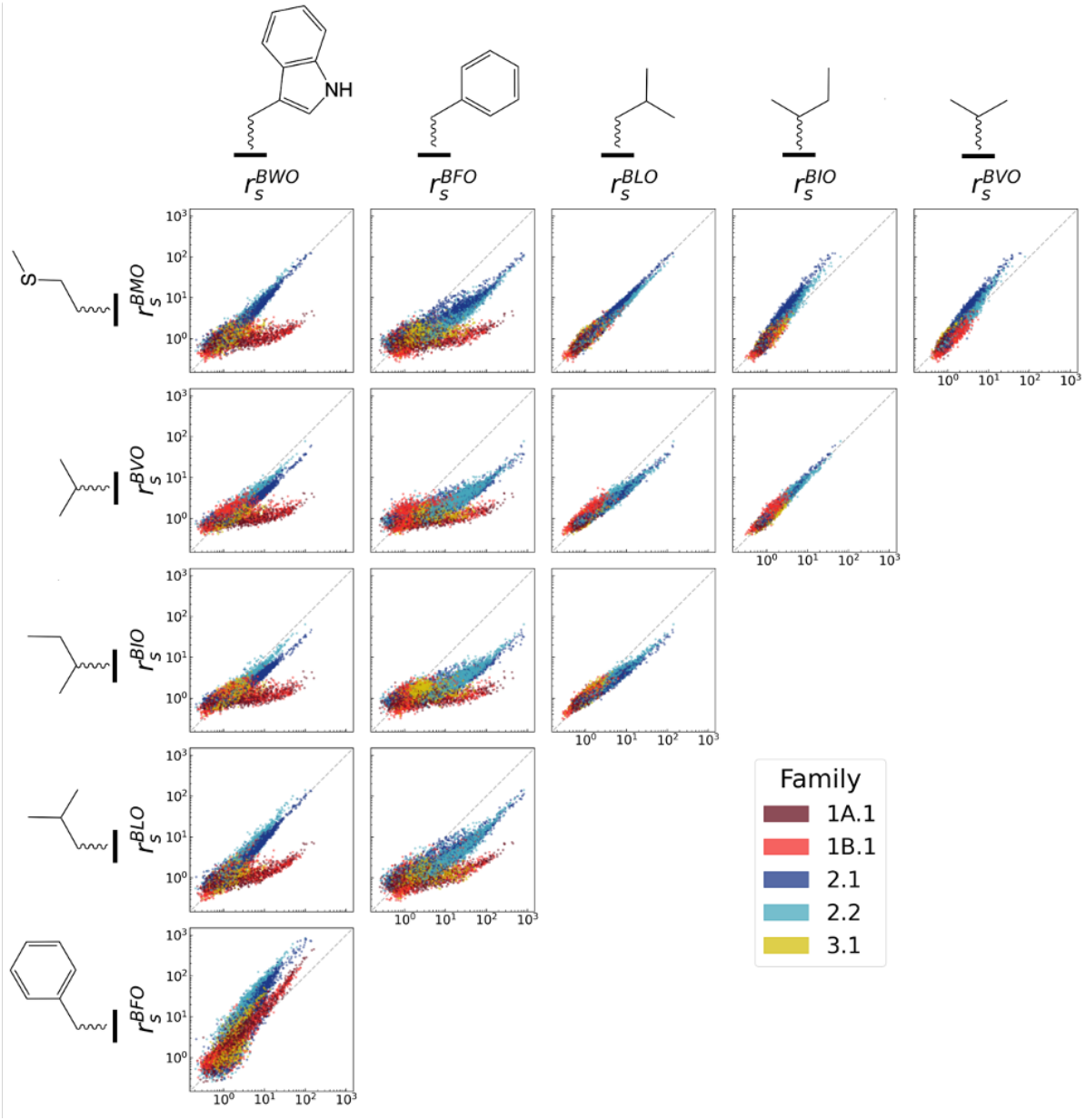

Figure 6.7: **Pairwise comparisons of ribozyme activity on different substrates.** Pairwise comparisons of catalytic enhancement (r_s) for individual ribozyme sequences with each BXO substrate. Dashed gray line indicates the identity line. Substrates are ordered by hydrophilicity.[207] See Figure 6.13 at the end of this chapter for error bars and mutant order for each family.

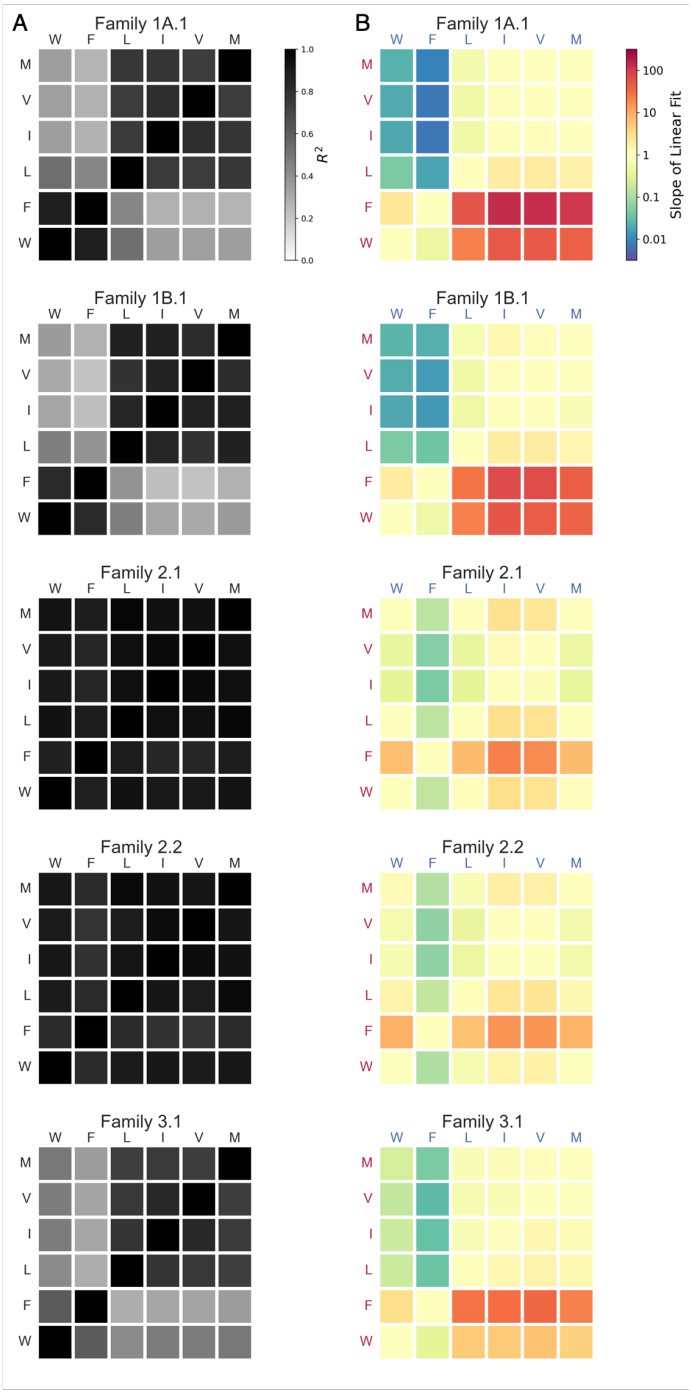

Figure 6.8: **Ribozyme correlations of activity and substrate preferences.** (A) Heat maps of coefficient of determination (R^2) for pairwise comparisons in Figure 6.7. (B) Heat maps for slopes of linear regression fits for pairwise comparisons in Figure 6.7. Slope >1 indicates a preference for the substrate on the y-axis; slope <1 indicates a preference for the substrate on the x-axis.

	Promiscuity Index		Aromatic Preference	
Family	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

Table 6.1: **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 all cases.

6.6 Substrate Specificity is Positively Correlated with Activity

To probe the relationship between catalytic activity and substrate specificity, two measures of specificity were used. First, as a general measure of substrate specificity for each sequence, we adapted the 'promiscuity index'[192] (Figure 6.9). This metric ($I_s = -\frac{1}{logN}\sum_{i=1}^{N}\frac{r_i}{\sum_{j=1}^{N} r_j} log \frac{r_i}{\sum_{j=1}^{N} r_j}$) is a normalized entropy which describes the evenness of rates across different substrates. The promiscuity index I_s ranges from 0 to 1, such that sequences that are completely promiscuous have $I_s = 1$ and sequences completely specific to one substrate have $I_s = 0$. Promiscuity was observed to decrease as overall activity increased for all families (Figure 6.9).

Second, since ribozymes in some families displayed preferential activity with aromatic amino acids compared to non-aromatic amino acids, the relative preference for aromatic substrates was calculated as $\frac{(r_s^{BWO}+r_s^{BFO})}{\sum_X r_s^{BXO}}$. 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 (Figure 6.10). Both the aromatic preference and the promiscuity index showed that the total activity of a ribozyme was positively correlated with specificity (negatively correlated with promiscuity

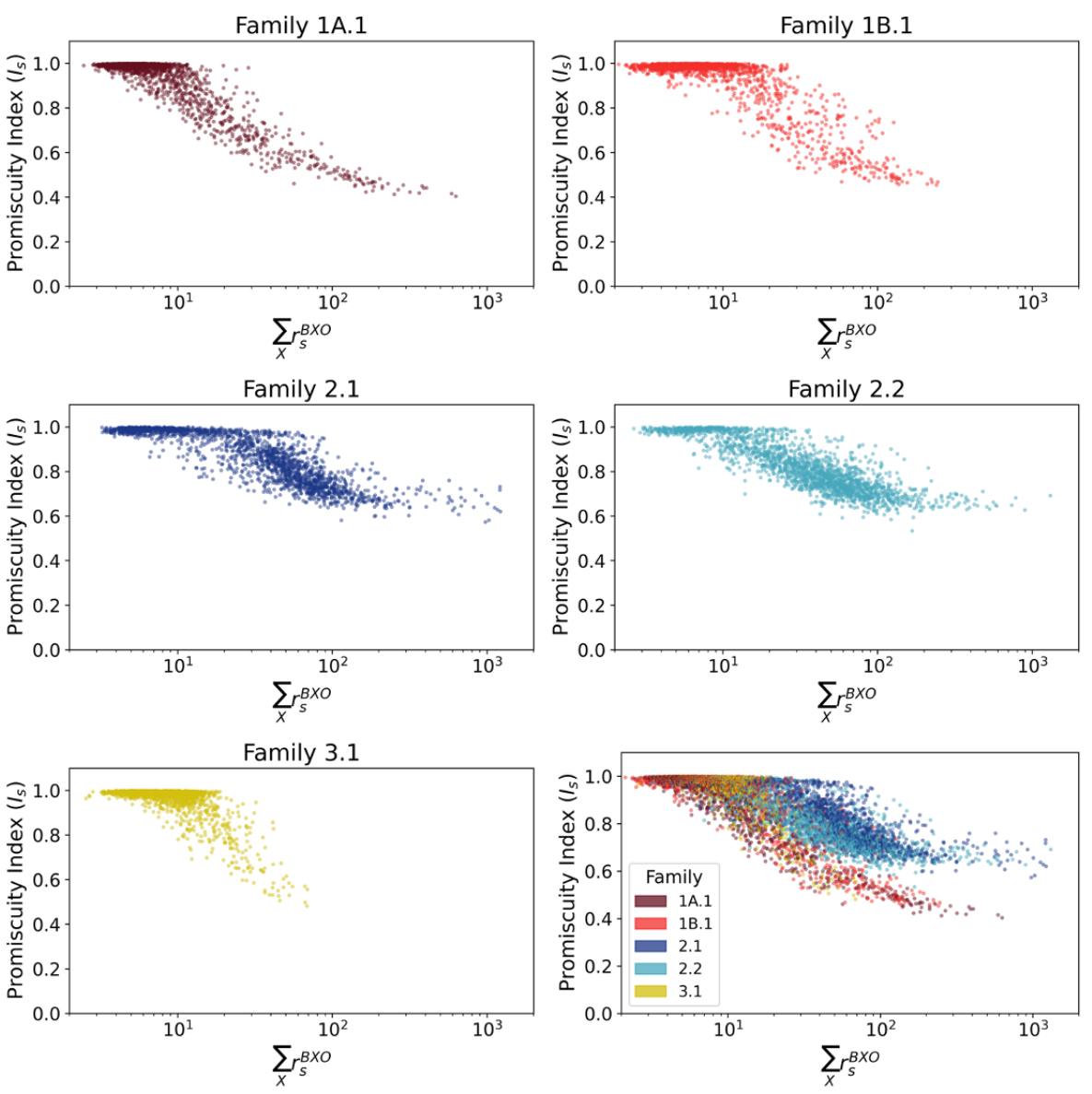

Figure 6.9: **Relationship between activity and promiscuity.** Promiscuity index (I_s) values for each sequence as a function of total activity (sum of activities with all tested substrates). Shown are each family individually and a composite of all families. The general trend indicates that promiscuity decreases (specificity increases) as overall activity increases.

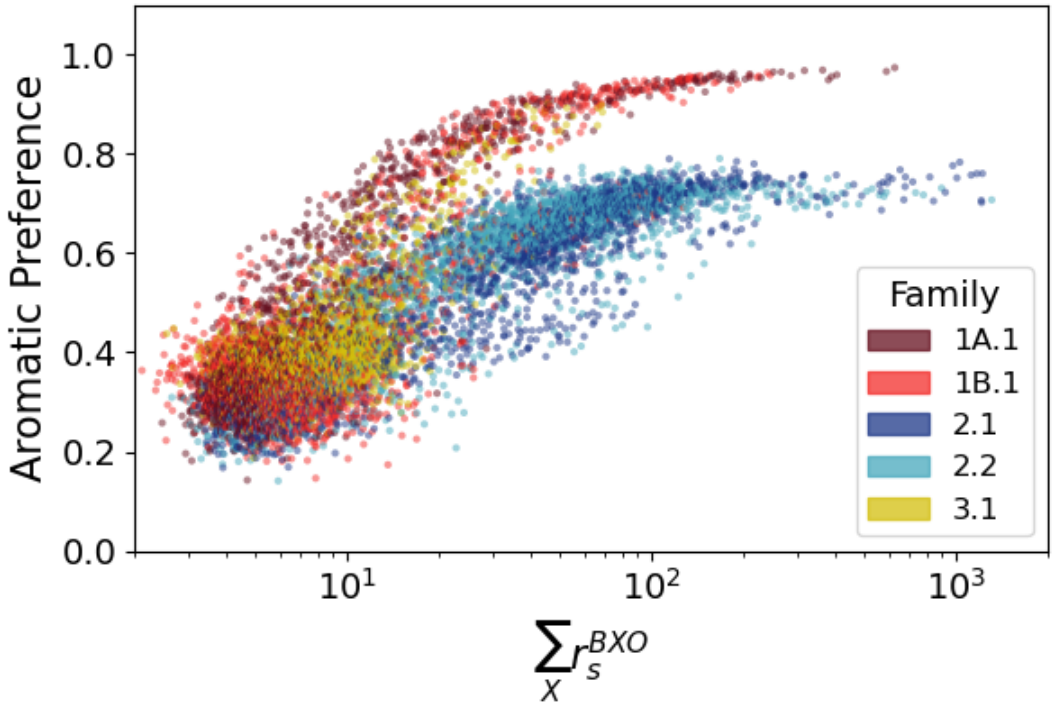

Figure 6.10: **Relationship between activity and aromatic preference.** Observed preference for aromatic substrates, as determined by the ratio of the sum of activities on BWO and BFO to the sum of activities on all tested substrates (BXO) (aromatic preference ratio). Increasing preference can be observed for increasing activity.

index and positively correlated with aromatic preference; Table 6.1).

6.7 Abundance of Opportunities for Co-Option for Alternative Substrates

To quantify the frequency of sequences with potential for co-option, sequences were categorized 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

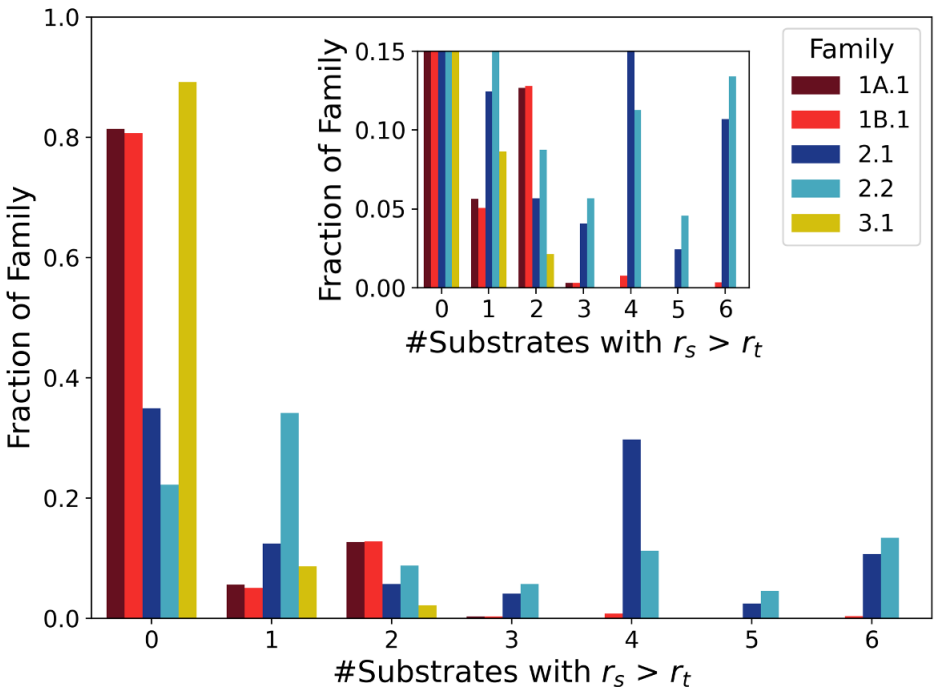

Figure 6.11: **Ribozyme sequences with co-option potential.** The frequency distribution of the fraction of unique sequences in each family (y-axis) that is active on a given number of substrates (x-axis). Activity on 2 or more substrates indicates potential for co-option. While Motif 2 sequences (Families 2.1 and 2.2) show a higher abundance of sequences active on more substrates, all families possess some co-option potential. Inset shows an enlargement of the low y-value region of the plot.

(Figure 4.10, Table 4.4), so values of $r_s > 5$ are statistically significantly greater than the normalized background rate. Second, increasing the rate 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, ribozymes that were active on more than one substrate were considered capable of co-option.

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 yielded abundant opportunities for co-option, with most sequences being active with at least two substrates (1029 sequences in Family 2.1; 853 sequences in Family 2.2), and many active with all six tested substrates (Figure 6.11). In contrast, Families 1A.1, 1B.1,

and 3.1, which contain more inactive sequences and generally preferred aromatic amino acids, yielded fewer co-option opportunities, with most sequences accepting one or zero substrates. Of sequences capable of co-option 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).

6.8 Optimization of Co-Opted Function 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 an optimally active sequence on a given substrate through evolution over the fitness landscape, 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) (Figure 6.12). 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-preferring ribozymes.

6.9 Discussion

The genetic code is an ideal platform for studying co-option in ribozyme evolution, as aminoacylations by the 20 biogenic amino acids represent naturally distinct functions.

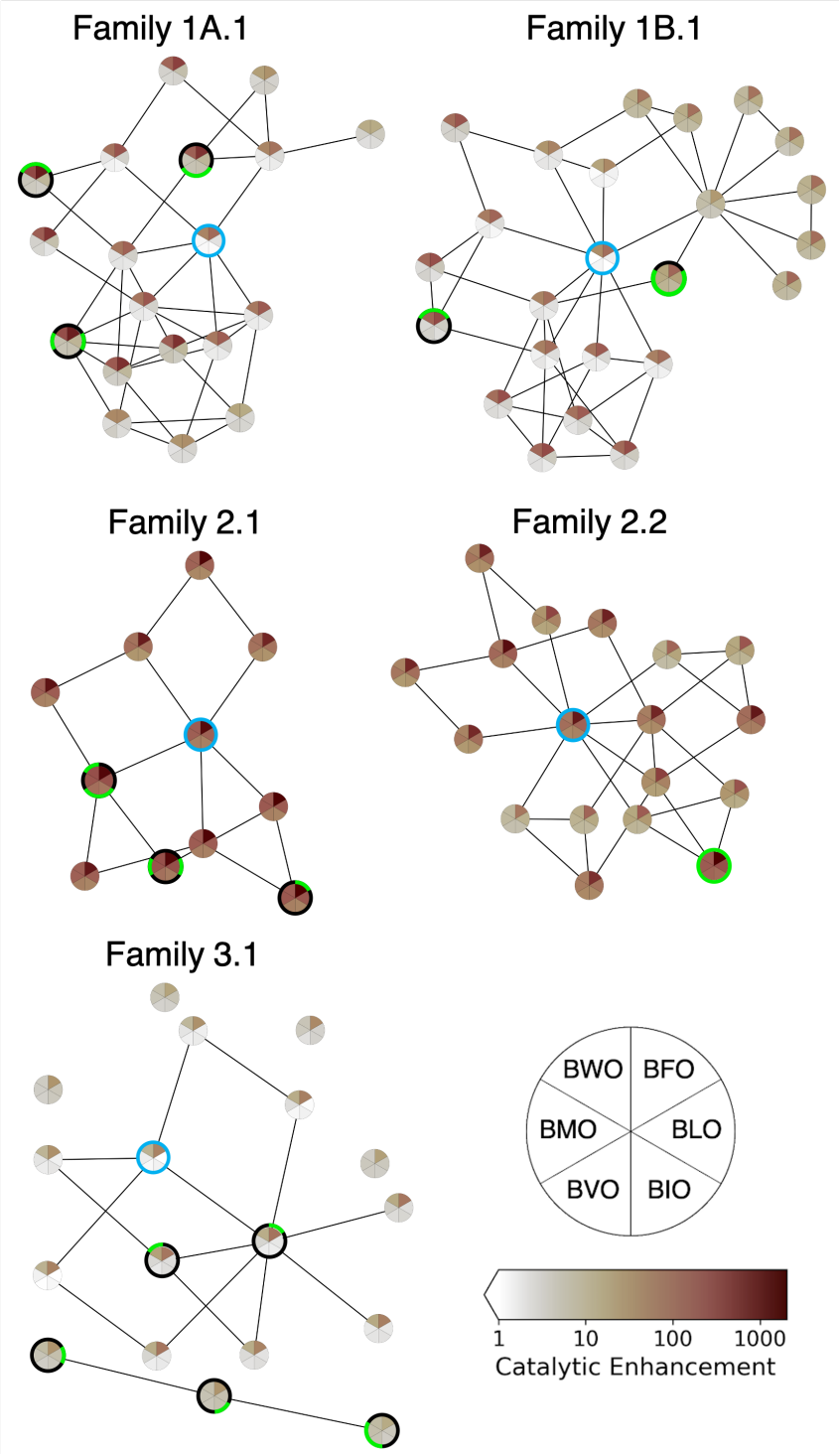

Figure 6.12: **Evolutionary pathways for optimization from potential co-option points on the fitness landscape.**

Figure 6.12: **Evolutionary pathways for optimization from potential co-option points on the fitness landscape.** 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 $d = 1$.

The genetic code itself is thought to have been established during the RNA World, in which ribozymes catalyzed aminoacylation.[361–363,382] Here we determined the activities of self-aminoacylating ribozyme families with several activated amino acid substrates. These ribozymes were originally discovered by exhaustive *in vitro* selection over sequence space (21 nt random region flanked by constant regions),[200] and thus their properties are expected to be a reasonable model for self-aminoacylating ribozymes. Each tested family contained dozens or hundreds of ribozyme 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 be co-opted and evolved to adopt other amino acid substrates. In an RNA World scenario, this process could be 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. 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 biophysical 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.[286,366,369] However, an alternative view is that this trait emerged as a by-product during the initial expansion of the genetic code.[12,371,372] 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 would be similar to the ancestral versions.[383] 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.[12,373] Using our experimental system of self-aminoacylating RNAs, we found that all ribozymes showed preferences for certain biophysical features, being par-

ticularly sensitive to aromaticity and branching in the side chain. Thus, co-option of these ribozymes would produce an association between these biophysical features and the RNA sequence, possibly including the primitive anticodon region. While the self-aminoacylating ribozymes studied here are a model system and not expected to recapitulate the evolution of the existing standard genetic code, these results illustrate the feasibility of the general principle that ribozyme co-option to incorporate new amino acid substrates would lead to error minimization as a by-product of expansion of the genetic code.

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,[136,145,158,159,384–386] 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.[145,149] The evolutionary consequence of the positive activity-specificity correlation is 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 the 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.[377–381] Interestingly, the aromatic residues were generally preferred by all ribozyme families. While the original

selection employed a tyrosine analog, an analogous selection using the leucine analog did not yield new ribozymes, indicating that this preference may be intrinsic. 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*.[206] 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. Such emergent phenomena have been argued to be critical complements to natural selection during the origin of translation.[387,388] Like the spandrels of St. Mark's Cathedral, architectural by-products that acquired important aesthetic value,[389] error minimization and specificity may have originated as mechanistic by-products of how the genetic code emerged, to later become invaluable features of the modern genetic code.

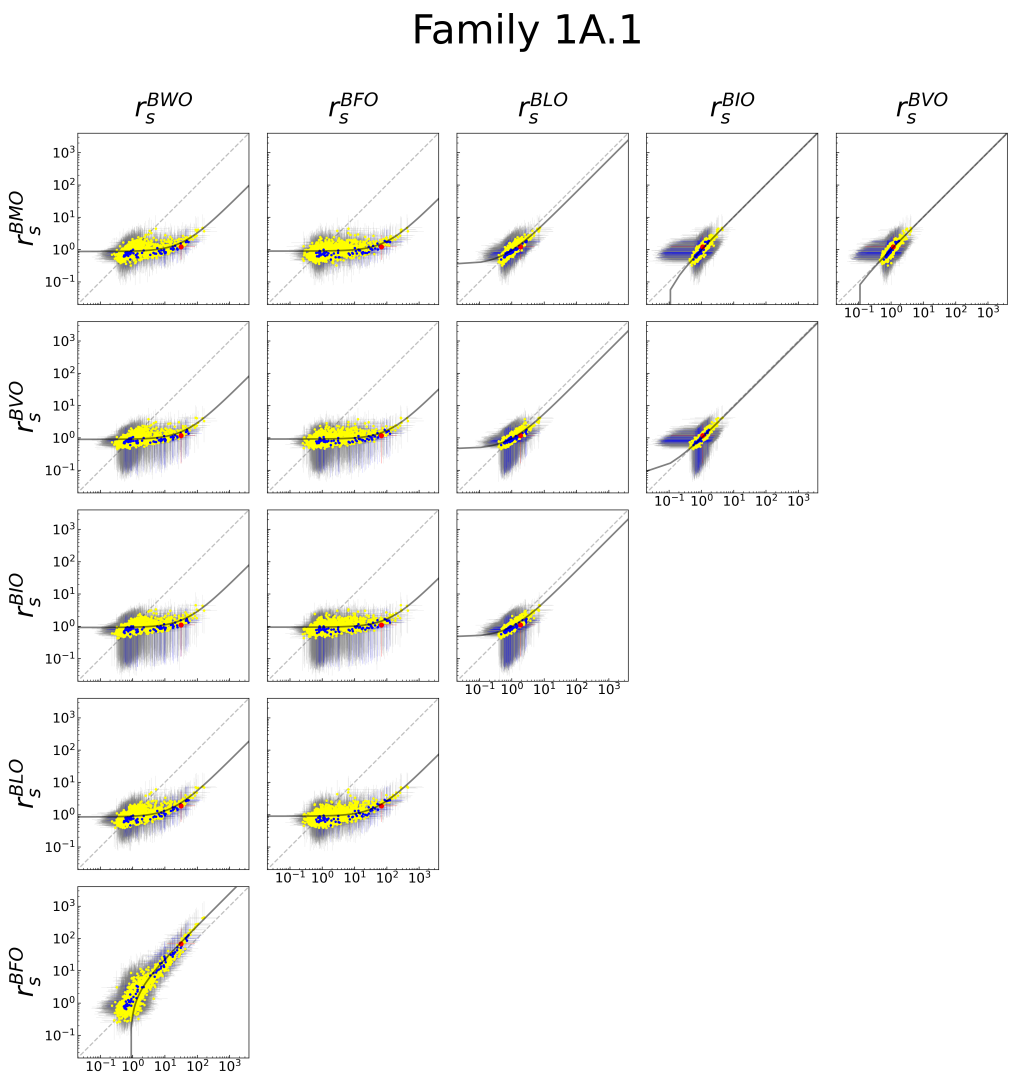

Figure 6.13: **Pairwise comparisons of ribozyme activity on different substrates.** Wild-type sequences are shown in red, single-mutants are shown in blue, and double-mutants are shown in yellow. Dashed gray line indicates line of identity. Black lines indicate linear regression fits used to calculate R^2 values and slopes in Figure 6.8. The same data are also plotted in Figure 6.7, but here the families are plotted separately, with mutant order and error bars (95% confidence interval) indicated. 95% confidence intervals of r_s were calculated from confidence intervals of $k_s A_s$ with normalization by the constant $k_0 A_0$.

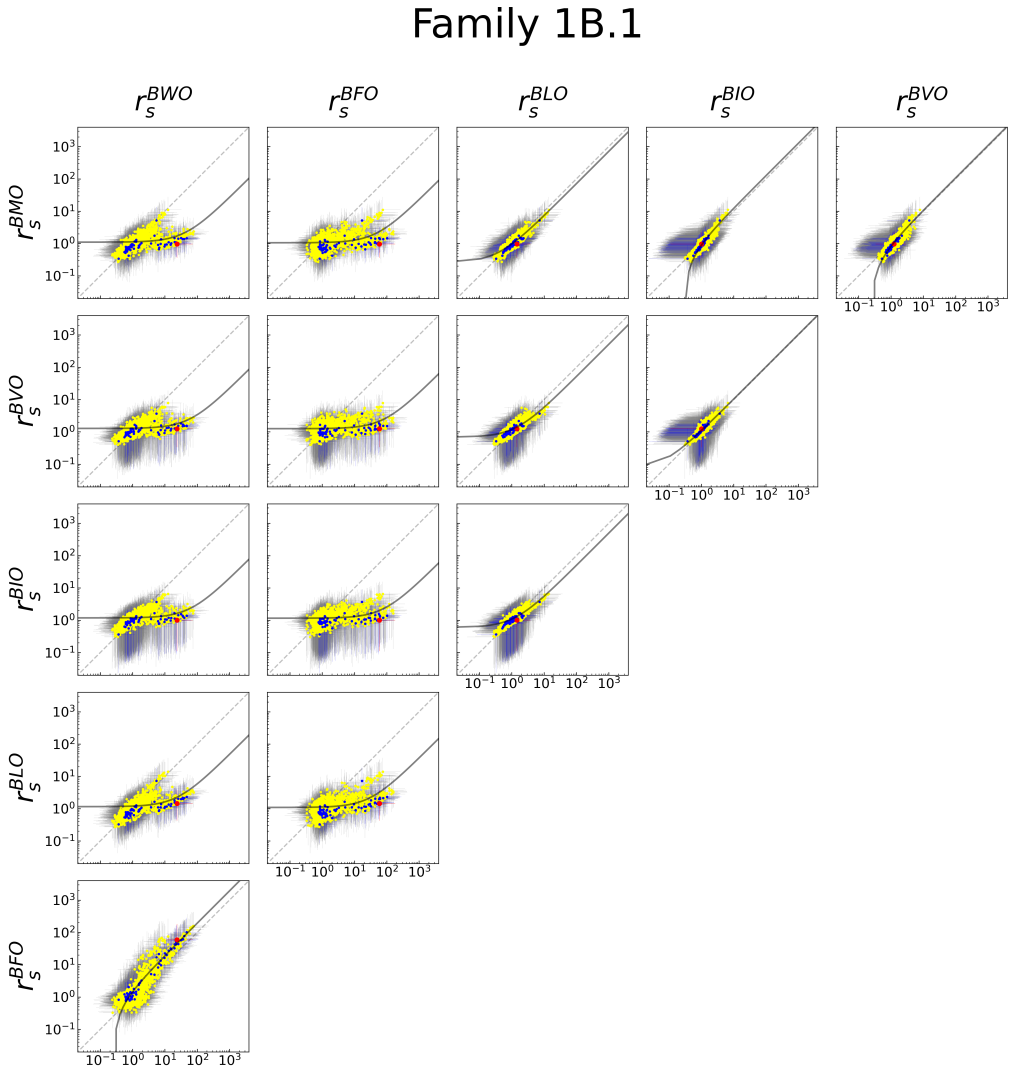

Figure 6.13: **Pairwise comparisons of ribozyme activity on different substrates.**

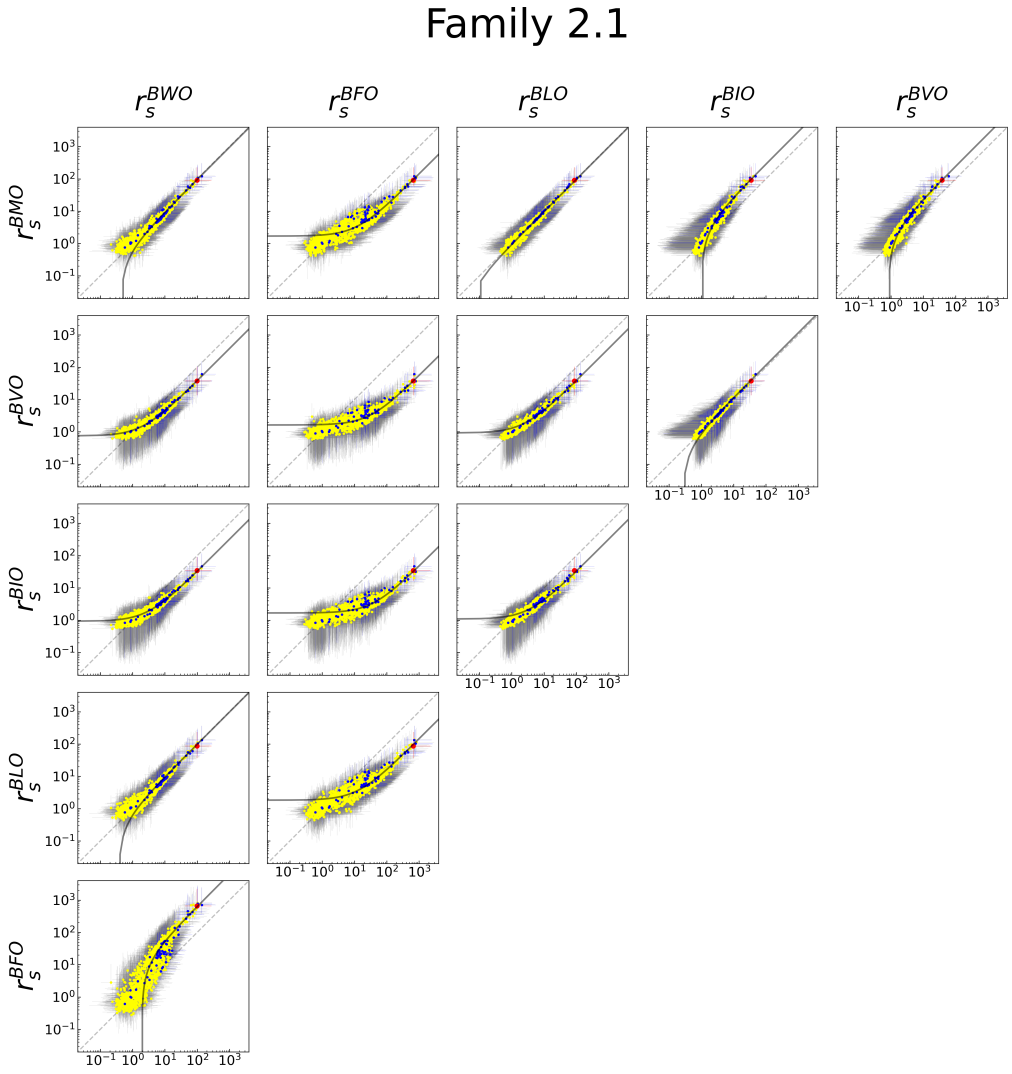

Figure 6.13: **Pairwise comparisons of ribozyme activity on different substrates.**

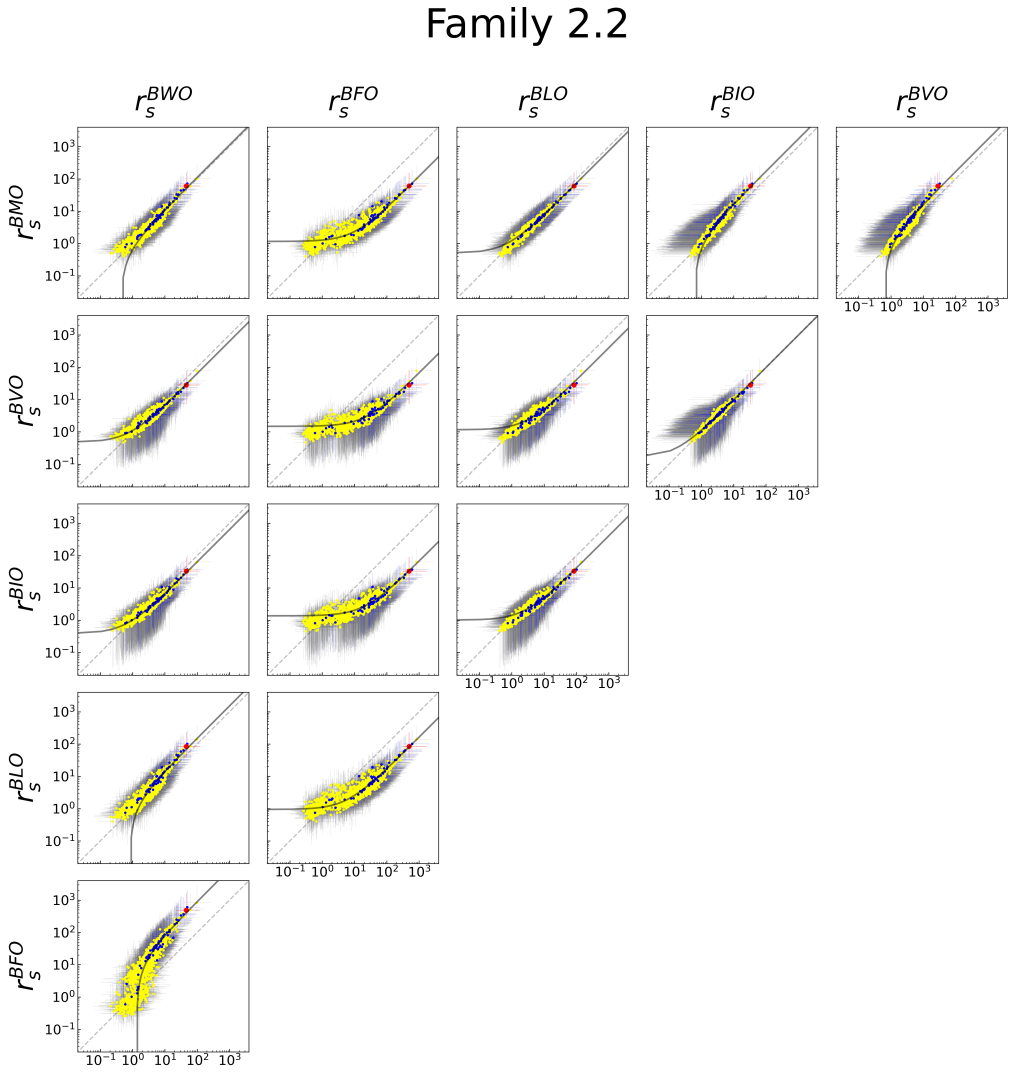

Figure 6.13: **Pairwise comparisons of ribozyme activity on different substrates.**

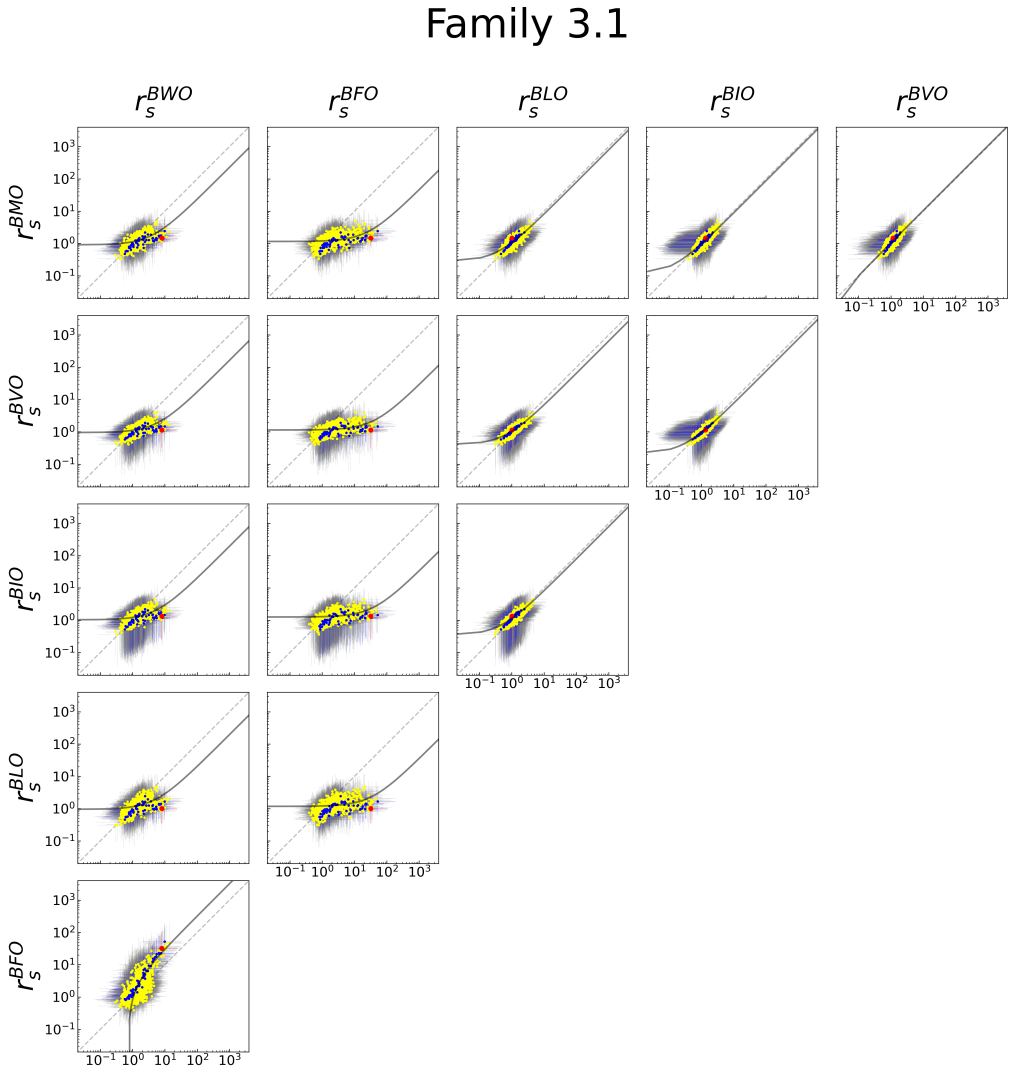

Figure 6.13: **Pairwise comparisons of ribozyme activity on different substrates.**

Chapter 7

Concluding Remarks

The research presented here describes the fundamental potential for a catalytic function, aminoacylation, to arise from a random collection of sequences, evolve higher fitness, and become co-opted for function with new substrates. Self-aminoacylating ribozymes were identified through *in vitro* selection from full coverage of sequence space and characterized using a massively parallel kinetic assay. Three major sequence motifs were identified on the landscape and analysis of evolutionary pathways revealed that, while local optimization within a ribozyme family would be possible, optimization of activity over the entire landscape would be frustrated by large valleys of low activity. The sequence motifs associated with each peak represent different solutions for catalysis, so the inability to traverse the landscape globally corresponds to an inability to restructure the ribozyme without losing activity. Five families representing the three sequence motifs were further investigated to measure their activity on six alternative substrates and were observed to possess high levels of co-optability. Related ribozymes exhibited similar biophysical substrate preferences and ribozyme activity was positively correlated with specificity.

The frustrated nature of the evolutionary landscape suggests that chance emergence

of a ribozyme motif would be more important than optimization by natural selection in determining evolutionary outcomes. In contrast, co-option of existing ribozyme motifs for new functions occurs quite readily, as a single functional motif can be recycled for use with many substrates. Indeed, promiscuity is considered one of the central tenets of evolutionary innovation.[182] The field of directed evolution has similarly relied on the co-option of promiscuous enzymes to produce new functions, with the greatest challenges often in optimizing, rather than identifying, a desired activity.[138] In many cases, specificity may in fact be the trait that requires selective pressure to attain.[390] Aminoacyl-tRNA synthetases, for example, often rely on extensive proofreading mechanisms to maintain the integrity of the genetic code.[360] Furthermore, promiscuity at the origin of life may not have been limited to ribozymes. Recent evolutionary analysis of another essential component of the translation machinery, elongation factor-Tu, suggests that early proteins may have exhibited similar substrate promiscuity, later evolving into more specialized forms.[391] While many different ribozyme activities have been identified, the range of catalytic RNA chemistries are yet to be fully explored. Testing the limits of co-optability through further exploration of substrate space may reveal the full extent of possible functions which could arise from the chance emergence of a single catalytic function. For example, catalytic promiscuity in aminoacylation ribozymes may have even made these the first biomolecules to catalyze the formation of polypeptides.[208,209]

The positive correlation between activity and specificity observed for all tested ribozyme families indicates that selection for increased activity also leads to increased specificity. However, the specificity identified in these ribozymes was not the result of selection for highly selective catalysts, but due to innate preferences for certain substrates by ribozymes in each family. In fact, the high degree of correlation between the landscapes of ribozymes with BYO, BFO, and BLO suggest instead that high selectivity for any one substrate may be unattainable in this system. While purifying selection for

highly specific ribozymes might produce different results, the results obtained in this work may be more relevant to a complex, dynamic prebiotic milieu. This type of environment would have been necessary for ribonucleic acid polymers to emerge, but also would have produced myriad other compounds. The aromatic amino acids (tyrosine, phenylalanine, tryptophan, and histidine) are thought to be among the most recent additions to the genetic code.[377–381] It is possible that a high affinity for these side chains by RNA was a factor in incorporating these particular amino acids, out of any possible R group that might have resulted from prebiotic chemistries, into the select 20 biogenic amino acids of the universal genetic code. Based on the results of this research, it is likely that the emergence of a single aminoacylation ribozyme may be sufficient for generating a diverse repertoire of aminoacyl-RNAs available for the emergence of protein translation. The type of selective pressures that would have resulted in aminoacyl-RNAs prior to the evolution of a translation system are unclear, but it's possible that, given the limited chemical complexity of RNA, aminoacyl groups may have initially served as a form of covalent cofactor, expanding the catalytic potential of these molecules. In this 'prebiotic soup', any new chemical compounds would likely have become substrates for promiscuous catalysts, creating a cycle of ever-expanding reactants and products. The most harmoniously interconnected of these reactions, contained by membranous vesicles, could then have established an early metabolism. It's possible, then, that promiscuous ribozymes were a key driver in the origin of life.

The original selection which identified these ribozymes selected only for reaction with BYO. While many other factors were implicitly selected for (e.g., RNA stability, reverse-transcription efficiency, etc.), many beneficial properties were observed which were not evenly distributed across the landscape. Traits like co-optability, specificity, and mutational robustness were present in varying degrees on the landscape. Given the apparent challenges for evolution in a prebiotic setting, the existence of these types of unintendend

properties would be immensely beneficial for the emergence of complex systems. This work suggests that the error minimization featured in the organization of the genetic code may have been one of these fortuitous characteristics. The further exploration of these types of properties may ultimately reveal the extent to which the origin of life was serendipitous or inevitable. While much remains to be discovered about how living systems evolve, this work, along with that of many others, may one day help us understand where we come from and where we are going.

Bibliography

[1] Bartel, D. P.; Szostak, J. W. *Science* **1993**, *261*, 1411–1418.

[2] Arnold, F. H. *Nat Biotechnol* **1998**, *16*, 617–8.

[3] Mushegian, A. R.; Koonin, E. V. *Proc Natl Acad Sci U S A* **1996**, *93*, 10268–73.

[4] Woese, C. *Proc Natl Acad Sci U S A* **1998**, *95*, 6854–9.

[5] Koonin, E. V. *Nat Rev Microbiol* **2003**, *1*, 127–36.

[6] Miller, S. L. *Science* **1953**, *117*, 528–9.

[7] Trifonov, E. N. *J Biomol Struct Dyn* **2011**, *29*, 259–66.

[8] Gilbert, W. *Nature* **1986**, *319*, 618.

[9] Pressman, A.; Blanco, C.; Chen, I. A. *Curr. Biol.* **2015**, *25*, R953–963.

[10] Joyce, G. F.; Szostak, J. W. *Cold Spring Harb Perspect Biol* **2018**, *10*.

[11] Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920–30.

[12] Koonin, E. V.; Novozhilov, A. S. *Annual Review of Genetics, Vol 51* **2017**, *51*, 45–62.

[13] Blanco, C.; Janzen, E.; Pressman, A.; Saha, R.; Chen, I. A. *Annu. Rev. Biophys.* **2019**, *48*, 1–18.

[14] Luksza, M.; Lassig, M. *Nature* **2014**, *507*, 57–61.

[15] Wright, S. *Genetics* **1931**, *16*, 97–159.

[16] Smith, J. M. *Nature* **1970**, *225*, 563–4.

[17] Kauffman, S.; Levin, S. *J Theor Biol* **1987**, *128*, 11–45.

[18] Wright, S. *Proceedings of the Sixth International Congress of Genetics* **1932**, *1*, 356–366.

[19] Pinheiro, V. B.; Holliger, P. *Trends Biotechnol* **2014**, *32*, 321–8.

[20] de Visser, J. A.; Krug, J. *Nat Rev Genet* **2014**, *15*, 480–90.

[21] Obolski, U.; Ram, Y.; Hadany, L. *Rep Prog Phys* **2018**, *81*, 012602.

[22] Aita, T.; Husimi, Y. *J Theor Biol* **1996**, *182*, 469–85.

[23] Gillespie, J. H. *The American Naturalist* **1983**, *121*, 691–708.

[24] Poelwijk, F. J.; Tanase-Nicola, S.; Kiviet, D. J.; Tans, S. J. *J Theor Biol* **2011**, *272*, 141–4.

[25] Katilius, E.; Flores, C.; Woodbury, N. W. *Nucleic Acids Res* **2007**, *35*, 7626–35.

[26] Kauffman, S. A.; Weinberger, E. D. *J Theor Biol* **1989**, *141*, 211–45.

[27] Kingman, J. F. C. *Journal of Applied Probability* **1978**, *15*, 1–12.

[28] Perelson, A. S.; Macken, C. A. *Proc Natl Acad Sci U S A* **1995**, *92*, 9657–61.

[29] Geard, N.; Wiles, J.; Hallinan, J.; Tonkes, B.; Skellett, B. A comparison of neutral landscapes - NK, NKp and NKq. Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on. pp 205–210.

[30] Hayashi, Y.; Aita, T.; Toyota, H.; Husimi, Y.; Urabe, I.; Yomo, T. *PLoS One* **2006**, *1*, e96.

[31] Rowe, W.; Platt, M.; Wedge, D. C.; Day, P. J.; Kell, D. B.; Knowles, J. *J R Soc Interface* **2010**, *7*, 397–408.

[32] Franke, J.; Klozer, A.; de Visser, J. A.; Krug, J. *PLoS Comput Biol* **2011**, *7*, e1002134.

[33] Ferretti, L.; Schmiegelt, B.; Weinreich, D.; Yamauchi, A.; Kobayashi, Y.; Tajima, F.; Achaz, G. *J Theor Biol* **2016**, *396*, 132–43.

[34] Aita, T.; Uchiyama, H.; Inaoka, T.; Nakajima, M.; Kokubo, T.; Husimi, Y. *Biopolymers* **2000**, *54*, 64–79.

[35] Neidhart, J.; Szendro, I. G.; Krug, J. *Genetics* **2014**, *198*, 699–721.

[36] Kun, A.; Szathmary, E. *Life (Basel, Switzerland)* **2015**, *5*, 1497–517.

[37] Fontana, W.; Konings, D. A.; Stadler, P. F.; Schuster, P. *Biopolymers* **1993**, *33*, 1389–404.

[38] Tacker, M.; Fontana, W.; Stadler, P.; Schuster, P. *European Biophysics Journal* **1994**, *23*, 29–38.

[39] Huynen, M. A. *Journal of Molecular Evolution* **1996**,

[40] Huynen, M. A.; Stadler, P. F.; Fontana, W. *Proceedings of the National Academy of Sciences of the United States of America* **1996**, *93*, 397–401.

[41] Fontana, W.; Schuster, P. *Science (New York, N.Y.)* **1998**, *280*, 1451–5.

[42] Gavrilets, S. *Fitness Landscapes and the Origin of Species (MPB-41)*; Princeton University Press, 2004.

[43] Gavrilets, S. *J Hered* **2014**, *105 Suppl 1*, 743–55.

[44] Stadler, P. F.; Happel, R. *Journal of Mathematical Biology* **1999**, *38*, 435–478.

[45] Derrida, B. *Physical Review B* **1981**, *24*, 2613–2626.

[46] Schultes, E. A.; Bartel, D. P. *Science* **2000**, *289*, 448–452.

[47] Held, D. M.; Greathouse, S. T.; Agrawal, A.; Burke, D. H. *J Mol Evol* **2003**, *57*, 299–308.

[48] Curtis, E. A.; Bartel, D. P. *Nat Struct Mol Biol* **2005**, *12*, 994–1000.

[49] Hayashi, Y.; Sakata, H.; Makino, Y.; Urabe, I.; Yomo, T. *J Mol Evol* **2003**, *56*, 162–8.

[50] Jimenez, J. I.; Xulvi-Brunet, R.; Campbell, G. W.; Turk-MacLeod, R.; Chen, I. A. *Proc Natl Acad Sci U S A* **2013**, *110*, 14984–9.

[51] Petrie, K. L.; Joyce, G. F. *J Mol Evol* **2014**, *79*, 75–90.

[52] Sanchez-Luque, F. J.; Stich, M.; Manrubia, S.; Briones, C.; Berzal-Herranz, A. *Sci Rep* **2014**, *4*, 6242.

[53] Fischer, N. O.; Tok, J. B. H.; Tarasow, T. M. *PLoS ONE* **2008**, *3*, e2720–e2720.

[54] Knight, C. G.; Platt, M.; Rowe, W.; Wedge, D. C.; Khan, F.; Day, P. J.; McShea, A.; Knowles, J.; Kell, D. B. *Nucleic Acids Res* **2009**, *37*, e6.

[55] Athavale, S. S.; Spicer, B.; Chen, I. A. *Current opinion in chemical biology* **2014**, *22*, 35–39.

[56] Pitt, J. N.; Ferre-D'Amare, A. R. *Science* **2010**, *330*, 376–9.

[57] Humphrey, W.; Dalke, A.; Schulten, K. *J Mol Graph* **1996**, *14*, 33–8, 27–8.

[58] Otwinowski, J.; Plotkin, J. B. *Proceedings of the National Academy of Sciences* **2014**, *111*, E2301–E2309.

[59] Wu, N. C.; Dai, L.; Olson, C. A.; Lloyd-Smith, J. O.; Sun, R. *Elife* **2016**, *5*.

[60] Pressman, A.; Moretti, J. E.; Campbell, G. W.; Muller, U. F.; Chen, I. A. *Nucleic Acids Res* **2017**, *45*, 10922.

[61] Kobori, S.; Yokobayashi, Y. *Angewandte Chemie* **2016**,

[62] Jalali-Yazdi, F.; Lai, L. H.; Takahashi, T. T.; Roberts, R. W. *Angew Chem Int Ed Engl* **2016**, *55*, 4007–10.

[63] Dhamodharan, V.; Kobori, S.; Yokobayashi, Y. *ACS Chem Biol* **2017**, *12*, 2940–2945.

[64] Aguilar-Rodríguez, J.; Payne, J. L.; Wagner, A. *Nature ecology & evolution* **2017**, *1*, 0045.

[65] Le, D. D.; Shimko, T. C.; Aditham, A. K.; Keys, A. M.; Longwell, S. A.; Orenstein, Y.; Fordyce, P. M. *Proceedings of the National Academy of Sciences* **2018**, 201715888.

[66] Eschenmoser, A. *Science* **1999**, *284*, 2118–24.

[67] Rogers, J.; Joyce, G. F. *Nature* **1999**, *402*, 323–325.

[68] Rogers, J.; Joyce, G. F. *RNA (New York, N.Y.)* **2001**, *7*, 395–404.

[69] Reader, J. S.; Joyce, G. F. *Nature* **2002**, *420*, 841–844.

[70] Sefah, K.; Yang, Z.; Bradley, K. M.; Hoshika, S.; Jiménez, E.; Zhang, L.; Zhu, G.; Shanker, S.; Yu, F.; Turek, D.; Tan, W.; Benner, S. A. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 1449–54.

[71] Zhang, L. et al. *Journal of the American Chemical Society* **2015**, *137*, 6734–6737.

[72] Kimoto, M.; Yamashige, R.; Matsunaga, K.-i.; Yokoyama, S.; Hirao, I. *Nature Biotechnology* **2013**, *31*, 453–457.

[73] Gawande, B. N.; Rohloff, J. C.; Carter, J. D.; Von Carlowitz, I.; Zhang, C.; Schneider, D. J.; Janjic, N. *Proc Natl Acad Sci U S A* **2017**, *114*.

[74] Tolle, F.; Brändle, G. M.; Matzner, D.; Mayer, G. *Angewandte Chemie International Edition* **2015**, *54*, 10971–10974.

[75] Silverman, S. K. *Trends Biochem Sci* **2016**, *41*, 595–609.

[76] Walsh, R.; DeRosa, M. C. *Biochem Biophys Res Commun* **2009**, *388*, 732–5.

[77] Paul, N.; Springsteen, G.; Joyce, G. F. *Chem Biol* **2006**, *13*, 329–38.

[78] Pinheiro, V. B.; Taylor, A. I.; Cozens, C.; Abramov, M.; Renders, M.; Zhang, S.; Chaput, J. C.; Wengel, J.; Peak-Chew, S. Y.; McLaughlin, S. H.; Herdewijn, P.; Holliger, P. *Science* **2012**, *336*, 341–4.

[79] Yu, H.; Zhang, S.; Chaput, J. C. *Nat Chem* **2012**, *4*, 183–7.

[80] Taylor, A. I.; Pinheiro, V. B.; Smola, M. J.; Morgunov, A. S.; Peak-Chew, S.; Cozens, C.; Weeks, K. M.; Herdewijn, P.; Holliger, P. *Nature* **2015**, *69*, 208–215.

[81] Volk, D. E.; Yang, X.; Fennewald, S. M.; King, D. J.; Bassett, S. E.; Venkitachalam, S.; Herzog, N.; Luxon, B. A.; Gorenstein, D. G. *Bioorganic Chemistry* **2002**, *30*, 396–419.

[82] Abeydeera, N. D. et al. *Nucleic acids research* **2016**, *44*, 8052–64.

[83] Culbertson, M. C.; Temburnikar, K. W.; Sau, S. P.; Liao, J. Y.; Bala, S.; Chaput, J. C. *Bioorg Med Chem Lett* **2016**, *26*, 2418–2421.

[84] Pieken, W. A.; Olsen, D. B.; Benseler, F.; Aurup, H.; Eckstein, F. *Science (New York, N.Y.)* **1991**, *253*, 314–7.

[85] Thirunavukarasu, D.; Chen, T.; Liu, Z.; Hongdilokkul, N.; Romesberg, F. E. *J Am Chem Soc* **2017**, *139*, 2892–2895.

[86] Dunn, M. R.; Jimenez, R. M.; Chaput, J. C. *Nature Reviews Chemistry* **2017**, *1*, 0076.

[87] Ni, S.; Yao, H.; Wang, L.; Lu, J.; Jiang, F.; Lu, A.; Zhang, G. *Int J Mol Sci* **2017**, *18*.

[88] Rothlisberger, P.; Hollenstein, M. *Adv Drug Deliv Rev* **2018**,

[89] Buenrostro, J. D.; Araya, C. L.; Chircus, L. M.; Layton, C. J.; Chang, H. Y.; Snyder, M. P.; Greenleaf, W. J. *Nature Biotechnology* **2014**, *32*, 562–568.

[90] Puchta, O.; Cseke, B.; Czaja, H.; Tollervey, D.; Sanguinetti, G.; Kudla, G. *Science* **2016**, *352*, 840–844.

[91] Li, C.; Qian, W.; Maclean, C. J.; Zhang, J. *Science* **2016**, *352*, 837–840.

[92] Malyshev, D. A.; Dhami, K.; Lavergne, T.; Chen, T.; Dai, N.; Foster, J. M.; Correa, J., I. R.; Romesberg, F. E. *Nature* **2014**, *509*, 385–8.

[93] Araya, C. L.; Fowler, D. M. *Trends Biotechnol* **2011**, *29*, 435–42.

[94] Phillips, A. M.; Gonzalez, L. O.; Nekongo, E. E.; Ponomarenko, A. I.; McHugh, S. M.; Butty, V. L.; Levine, S. S.; Lin, Y. S.; Mirny, L. A.; Shoulders, M. D. *Elife* **2017**, *6*.

[95] Fowler, D. M.; Fields, S. *Nat Methods* **2014**, *11*, 801–7.

[96] Starita, L. M.; Fields, S. *Cold Spring Harb Protoc* **2015**, *2015*, 711–4.

[97] Fowler, D. M.; Araya, C. L.; Fleishman, S. J.; Kellogg, E. H.; Stephany, J. J.; Baker, D.; Fields, S. *Nat Methods* **2010**, *7*, 741–6.

[98] Sarkisyan, K. S. et al. *Nature* **2016**, *533*, 397–401.

[99] Hietpas, R. T.; Jensen, J. D.; Bolon, D. N. *Proc Natl Acad Sci U S A* **2011**, *108*, 7896–901.

[100] Whitehead, T. A.; Chevalier, A.; Song, Y.; Dreyfus, C.; Fleishman, S. J.; De Mattos, C.; Myers, C. A.; Kamisetty, H.; Blair, P.; Wilson, I. A.; Baker, D. *Nat Biotechnol* **2012**, *30*, 543–8.

[101] Weinreich, D. M.; Lan, Y.; Jaffe, J.; Heckendorn, R. B. *Journal of Statistical Physics* **2018**,

[102] Szendro, I. G.; Schenk, M. F.; Franke, J.; Krug, J.; De Visser, J. A. G. *J Stat Mech-Theory E* **2013**, *2013*, P01005.

[103] Mimi, G.; Loana, A.; Roland, W. *Angew. Chem., Int. Ed.* **2017**, *56*, 2302–2306.

[104] Attwater, J.; Wochner, A.; Holliger, P. *Nat. Chem.* **2013**, *5*, 1011.

[105] Frommer, J.; Appel, B.; Müller, S. *Curr. Opin. Biotechnol.* **2015**, *31*, 35–41.

[106] Schuabb, C.; Kumar, N.; Pataraia, S.; Marx, D.; Winter, R. *Nat. Commun.* **2017**, *8*, 14661.

[107] Blount, Z. D.; Borland, C. Z.; Lenski, R. E. *Proc Natl Acad Sci U S A* **2008**, *105*, 7899–906.

[108] Melnikov, A.; Rogov, P.; Wang, L.; Gnirke, A.; Mikkelsen, T. S. *Nucleic Acids Res* **2014**, *42*, e112.

[109] Arroyo-Curras, N.; Dauphin-Ducharme, P.; Ortega, G.; Ploense, K. L.; Kippin, T. E.; Plaxco, K. W. *ACS Sens* **2018**, *3*, 360–366.

[110] Saha, R.; Pohorille, A.; Chen, I. A. *Orig. Life Evol. Biosph.* **2015**, *44*, 319–324.

[111] Rivas, G.; Minton, A. P. *Trends Biochem. Sci.* **2016**, *41*, 970–981.

[112] Mimi, G.; Christoph, H.; Satyajit, P.; Loana, A.; Gabriele, S.; Roland, W. *ChemPhysChem* **2017**, *18*, 2951–2972.

[113] Daher, M.; Widom, J. R.; Tay, W.; Walter, N. G. *J. Mol. Biol.* **2018**, *430*, 509–523.

[114] Saha, R.; Verbanic, S.; Chen, I. A. *Nature Communications* **2018**, *9*, 2313.

[115] Lee, H.-T.; Kilburn, D.; Behrouzi, R.; Briber, R. M.; Woodson, S. A. *Nucleic Acids Res.* **2015**, *43*, 1170–1176.

[116] Rode, A. B.; Endoh, T.; Sugimoto, N. *Angew Chem Int Ed Engl* **2018**, *57*, 6868–6872.

[117] Vaidya, N.; Manapat, M. L.; Chen, I. A.; Xulvi-Brunet, R.; Hayden, E. J.; Lehman, N. *Nature* **2012**, *491*, 72.

[118] Filteau, M.; Hamel, V.; Pouliot, M.-C.; Gagnon-Arsenault, I.; Dubé, A. K.; Landry, C. R. *Mol. Syst. Biol.* **2015**, *11*, 832.

[119] Ho, W.-C.; Zhang, J. *Nat. Commun.* **2018**, *9*, 350.

[120] Hietpas, R. T.; Bank, C.; Jensen, J. D.; Bolon, D. N. A. *Evolution* **2013**, *67*, 3512–3522.

[121] Ota, N.; Kurahashi, R.; Sano, S.; Takano, K. *Biochimie* **2018**, *150*, 100–109.

[122] Cervera, H.; Lalic, J.; Elena, S. F. *J Virol* **2016**,

[123] Li, C.; Zhang, J. *Nat Ecol Evol* **2018**, *2*, 1025–1032.

[124] Domingo, J.; Diss, G.; Lehner, B. *Nature* **2018**, *558*, 117–121.

[125] Qian, W.; Ma, D.; Xiao, C.; Wang, Z.; Zhang, J. *Cell Rep.* **2012**, *2*, 1399–1410.

[126] Baird, N. J.; Inglese, J.; Ferré-D'Amaré, A. R. *Nature Communications* **2015**, *6*, 8898–8898.

[127] Janzen, E.; Blanco, C.; Peng, H.; Kenchel, J.; Chen, I. A. *Chem Rev* **2020**,

[128] Robertson, M. P.; Joyce, G. F. *Cold Spring Harb. Perspect. Biol.* **2012**, *4*.

[129] Woese, C. R. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *54*, 1546–1552.

[130] Crick, F. H. *J. Mol. Biol.* **1968**, *38*, 367–379.

[131] Orgel, L. E. *J. Mol. Biol.* **1968**, *38*, 381–393.

[132] Ellington, A. D.; Szostak, J. W. *Nature* **1990**, *346*, 818–822.

[133] Robertson, D. L.; Joyce, G. F. *Nature* **1990**, *344*, 467–468.

[134] Ellington, A. D.; Chen, X.; Robertson, M.; Syrett, A. *Int. J. Biochem. Cell Biol.* **2009**, *41*, 254–265.

[135] Jensen, R. A. *Annu. Rev. Microbiol.* **1976**, *30*, 409–425.

[136] Khersonsky, O.; Tawfik, D. S. *Annu. Rev. Biochem.* **2010**, *79*, 471–505.

[137] Schultes, E.; Bartel, D. P. *Science* **2000**, *289*.

[138] Arnold, F. H. *Angew Chem Int Ed Engl* **2018**, *57*, 4143–4148.

[139] Seffernick, J. L.; de Souza, M. L.; Sadowsky, M. J.; Wackett, L. P. *J. Bacteriol.* **2001**, *183*, 2405–2410.

[140] Cornish-Bowden, A.; Cardenas, M. L. *J. Phys. Chem. B* **2010**, *114*, 16209–16213.

[141] Peracchi, A. *Trends Biochem. Sci.* **2018**, *43*, 984–996.

[142] Yagisawa, S. *Biochem J.* **1995**, *308*, 305–311.

[143] Hopfield, J. J. *Proc. Natl. Acad. Sci. U. S. A.* **1974**, *71*, 4135–4139.

[144] Leu, K.; Obermayer, B.; Rajamani, S.; Gerland, U.; Chen, I. A. *Nucleic Acids Res.* **2011**, *39*, 8135–8147.

[145] Tawfik, D. S. *Curr. Opin. Chem. Biol.* **2014**, *21*, 73–80.

[146] Tokuriki, N.; Jackson, C. J.; Afriat-Jurnou, L.; Wyganowski, K. T.; Tang, R.; Tawfik, D. S. *Nat. Commun.* **2012**, *3*, 1257.

[147] Brustad, E. M.; Arnold, F. H. *Curr. Opin. Chem. Biol.* **2011**, *15*, 201–210.

[148] Carlson, J. C.; Badran, A. H.; Guggiana-Nilo, D. A.; Liu, D. R. *Nat. Chem. Biol.* **2014**, *10*, 216–222.

[149] Beard, W. A.; Shock, D. D.; Vande Berg, B. J.; Wilson, S. H. *J. Biol. Chem.* **2002**, *277*, 47393–47398.

[150] Ram Prasad, B.; Warshel, A. *Proteins* **2011**, *79*, 2900–2919.

[151] Caglayan, M.; Bilgin, N. *Biochimie* **2012**, *94*, 1968–1973.

[152] Azpurua, J.; Ke, Z.; Chen, I. X.; Zhang, Q.; Ermolenko, D. N.; Zhang, Z. D.; Gorbunova, V.; Seluanov, A. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17350–17355.

[153] Banerjee, K.; Kolomeisky, A. B.; Igoshin, O. A. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 5183–5188.

[154] Ninio, J. *Biochimie* **1975**, *57*, 587–595.

[155] Murugan, A.; Huse, D. A.; Leibler, S. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 12034–12039.

[156] Schomburg, I.; Jeske, L.; Ulbrich, M.; Placzek, S.; Chang, A.; Schomburg, D. *J. Biotechnol.* **2017**, *261*, 194–206.

[157] Tcherkez, G. G.; Farquhar, G. D.; Andrews, T. J. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 7246–7251.

[158] Savir, Y.; Noor, E.; Milo, R.; Tlusty, T. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 3475–3480.

[159] Flamholz, A. I.; Prywes, N.; Moran, U.; Davidi, D.; Bar-On, Y. M.; Oltrogge, L. M.; Alves, R.; Savage, D.; Milo, R. *Biochemistry* **2019**, *58*, 3365–3376.

[160] O'Brien, P. J.; Herschlag, D. *Chem. Biol.* **1999**, *6*, R91–R105.

[161] Khersonsky, O.; Roodveldt, C.; Tawfik, D. S. *Curr. Opin. Chem. Biol.* **2006**, *10*, 498–508.

[162] Hult, K.; Berglund, P. *Trends Biotechnol.* **2007**, *25*, 231–238.

[163] Brevet, A.; Plateau, P.; Cirakoglu, B.; Pailliez, J. P.; Blanquet, S. *J. Biol. Chem.* **1982**, *257*, 14613–14615.

[164] Copley, S. D. *Curr. Opin. Struct. Biol.* **2017**, *47*, 167–175.

[165] Newton, M. S.; Arcus, V. L.; Gerth, M. L.; Patrick, W. M. *Curr. Opin. Struct. Biol.* **2018**, *48*, 110–116.

[166] James, L. C.; Tawfik, D. S. *Trends Biochem. Sci.* **2003**, *28*, 361–368.

[167] Seffernick, J. L.; Johnson, G.; Sadowsky, M. J.; Wackett, L. P. *Appl. Environ. Microbiol.* **2000**, *66*, 4247–4252.

[168] Seibert, C. M.; Raushel, F. M. *Biochemistry* **2005**, *44*, 6383–6391.

[169] Scott, C.; Jackson, C. J.; Coppin, C. W.; Mourant, R. G.; Hilton, M. E.; Sutherland, T. D.; Russell, R. J.; Oakeshott, J. G. *Appl. Environ. Microbiol.* **2009**, *75*, 2184–2191.

[170] Griswold, K. E.; Aiyappan, N. S.; Iverson, B. L.; Georgiou, G. *J. Mol. Biol.* **2006**, *364*, 400–410.

[171] Bernhardt, R. *J. Biotechnol.* **2006**, *124*, 128–145.

[172] Ekroos, M.; Sjogren, T. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13682–13687.

[173] Bernhardt, R.; Urlacher, V. B. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 6185–6203.

[174] Steele, C. L.; Crock, J.; Bohlmann, J.; Croteau, R. *J. Biol. Chem.* **1998**, *273*, 2078–2089.

[175] Hammer, S. C.; Syren, P. O.; Seitz, M.; Nestl, B. M.; Hauer, B. *Curr. Opin. Chem. Biol.* **2013**, *17*, 293–300.

[176] Pilkington, S. J.; Dalton, H. *Methods in Enzymology* **1990**, *188*, 181–190.

[177] Dillingham, M. S.; Kowalczykowski, S. C. *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 642–671.

[178] Lovett, S. T. *EcoSal Plus* **2011**, *4*.

[179] Cheng, X. Y.; Huang, W. J.; Hu, S. C.; Zhang, H. L.; Wang, H.; Zhang, J. X.; Lin, H. H.; Chen, Y. Z.; Zou, Q.; Ji, Z. L. *PLoS One* **2012**, *7*, e38979.

[180] Copley, S. D. *Curr. Opin. Chem. Biol.* **2003**, *7*, 265–272.

[181] Jeffery, C. J. *Trends Genet.* **2003**, *19*, 415–417.

[182] Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Mc, Q. G. S.; Roodveldt, C.; Tawfik, D. S. *Nat. Genet.* **2005**, *37*, 73–76.

[183] James, L. C.; Tawfik, D. S. *Protein Sci.* **2001**, *10*, 2600–2607.

[184] James, L. C.; Roversi, P.; Tawfik, D. S. *Science* **2003**, *299*, 1362–1367.

[185] Yang, K.; Metcalf, W. W. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7919–7924.

[186] Kaltenbach, M.; Emond, S.; Hollfelder, F.; Tokuriki, N. *PLoS Genet* **2016**, *12*, e1006305.

[187] Fasan, R.; Meharenna, Y. T.; Snow, C. D.; Poulos, T. L.; Arnold, F. H. *J. Mol. Biol.* **2008**, *383*, 1069–1080.

[188] Ycas, M. *J. Theor. Biol.* **1974**, *44*, 145–160.

[189] Amitai, G.; Gupta, R. D.; Tawfik, D. S. *HFSP J.* **2007**, *1*, 67–78.

[190] Wroe, R.; Chan, H. S.; Bornberg-Bauer, E. *HFSP J.* **2007**, *1*, 79–87.

[191] Espinosa-Cantu, A.; Ascencio, D.; Barona-Gomez, F.; DeLuna, A. *Front. Genet.* **2015**, *6*, 227.

[192] Nath, A.; Atkins, W. M. *Biochemistry* **2008**, *47*, 157–166.

[193] Nath, A.; Zientek, M. A.; Burke, B. J.; Jiang, Y.; Atkins, W. M. *Drug Metab. Dispos.* **2010**, *38*, 2195–2203.

[194] Chakraborty, S.; Asgeirsson, B.; Rao, B. J. *PLoS One* **2012**, *7*, e49313.

[195] Chakraborty, S.; Rao, B. J. *PLoS One* **2012**, *7*, e32011.

[196] Honaker, M. T.; Acchione, M.; Zhang, W.; Mannervik, B.; Atkins, W. M. *J. Biol. Chem.* **2013**, *288*, 18599–18611.

[197] Illangasekare, M.; Sanchez, G.; Nickles, T.; Yarus, M. *Science* **1995**, *267*, 643–647.

[198] Illangasekare, M.; Yarus, M. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 5470–5475.

[199] Li, N.; Huang, F. *Biochemistry* **2005**, *44*, 4582–4590.

[200] Pressman, A. D.; Liu, Z.; Janzen, E.; Blanco, C.; Muller, U. F.; Joyce, G. F.; Pascal, R.; Chen, I. A. *J. Am. Chem. Soc.* **2019**, *141*, 6213–6223.

[201] Ohuchi, M.; Murakami, H.; Suga, H. *Curr. Opin. Chem. Biol.* **2007**, *11*, 537–542.

[202] Pang, Y. L.; Poruri, K.; Martinis, S. A. *Wiley Interdiscip. Rev. RNA* **2014**, *5*, 461–480.

[203] Saito, H.; Kourouklis, D.; Suga, H. *EMBO J.* **2001**, *20*, 1797–1806.

[204] Xin, Y.; Li, W.; First, E. A. *Biochemistry* **2000**, *39*, 340–347.

[205] Francklyn, C. S.; First, E. A.; Perona, J. J.; Hou, Y. M. *Methods* **2008**, *44*, 100–118.

[206] Blanco, C.; Bayas, M.; Yan, F.; Chen, I. A. *Curr. Biol.* **2018**, *28*, 526–537.

[207] Hopp, T. P.; Woods, K. R. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 3824–3828.

[208] Illangasekare, M.; Yarus, M. *RNA* **1999**, *5*, 1482–1489.

[209] Turk, R. M.; Chumachenko, N. V.; Yarus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 4585–4589.

[210] Saito, H.; Suga, H. *J. Am. Chem. Soc.* **2001**, *123*, 7178–7179.

[211] Saito, H.; Watanabe, K.; Suga, H. *RNA* **2001**, *7*, 1867–1878.

[212] Murakami, H.; Ohta, A.; Ashigai, H.; Suga, H. *Nat. Methods* **2006**, *3*, 357–359.

[213] Ohta, A.; Murakami, H.; Higashimura, E.; Suga, H. *Chem. Biol.* **2007**, *14*, 1315–1322.

[214] Fujino, T.; Goto, Y.; Suga, H.; Murakami, H. *J. Am. Chem. Soc.* **2013**, *135*, 1830–1837.

[215] Fujino, T.; Goto, Y.; Suga, H.; Murakami, H. *J. Am. Chem. Soc.* **2016**, *138*, 1962–1969.

[216] Katoh, T.; Suga, H. *Nucleic Acids Res.* **2019**, *47*, e54.

[217] Ferre-D'Amare, A. R.; Scott, W. G. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a003574.

[218] Prody, G. A.; Bakos, J. T.; Buzayan, J. M.; Schneider, I. R.; Bruening, G. *Science* **1986**, *231*, 1577–1580.

[219] Uhlenbeck, O. C. *Nature* **1987**, *328*, 596–600.

[220] Haseloff, J.; Gerlach, W. L. *Nature* **1988**, *334*, 585–591.

[221] Jeffries, A. C.; Symons, R. H. *Nucleic Acids Res.* **1989**, *17*, 1371–1377.

[222] Odai, O.; Hiroaki, H.; Tanaka, T.; Uesugi, S. *Nucleosides & Nucleotides* **1994**, *13*, 1569–1579.

[223] Clouet-D'Orval, B.; Uhlenbeck, O. C. *RNA* **1996**, *2*, 483–491.

[224] Ruffner, D. E.; Stormo, G. D.; Uhlenbeck, O. C. *Biochemistry* **1990**, *29*, 10695–10702.

[225] Perriman, R.; Delves, A.; Gerlach, W. L. *Gene* **1992**, *113*, 157–163.

[226] Zoumadakis, M.; Tabler, M. *Nucleic Acids Res.* **1995**, *23*, 1192–1196.

[227] Hertel, K. J.; Herschlag, D.; Uhlenbeck, O. C. *EMBO Journal* **1996**, *15*, 3751–3757.

[228] Shimayama, T.; Nishikawa, S.; Taira, K. *Biochemistry* **1995**, *34*, 3649–3654.

[229] Herschlag, D. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 6921–6925.

[230] Koizumi, M.; Ohtsuka, E. *Biochemistry* **1991**, *30*, 5145–5150.

[231] Huang, F.; Yarus, M. *Biochemistry* **1997**, *36*, 6557–6563.

[232] Zaher, H. S.; Watkins, R. A.; Unrau, P. J. *RNA* **2006**, *12*, 1949–1958.

[233] Huang, F.; Yarus, M. *Biochemistry* **1997**, *36*, 14107–14119.

[234] Huang, F.; Yarus, M. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 8965–8969.

[235] Huang, F.; Yarus, M. *J. Mol. Biol.* **1998**, *284*, 255–267.

[236] Zaher, H. S.; Unrau, P. J. *J. Am. Chem. Soc.* **2006**, *128*, 13894–13900.

[237] Ekland, E. H.; Szostak, J. W.; Bartel, D. P. *Science* **1995**, *269*, 364–370.

[238] Ekland, E. H.; Bartel, D. P. *Nucleic Acids Res* **1995**, *23*, 3231–8.

[239] Ekland, E. H.; Bartel, D. P. *Nature* **1996**, *382*, 373–376.

[240] Shechner, D. M.; Grant, R. A.; Bagby, S. C.; Koldobskaya, Y.; Piccirilli, J. A.; Bartel, D. P. *Science* **2009**, *326*, 1271–1275.

[241] Johnston, W. K.; Unrau, P. J.; Lawrence, M. S.; Glasner, M. E.; Bartel, D. P. *Science* **2001**, *292*, 1319–1325.

[242] Wang, Q. S.; Cheng, L. K.; Unrau, P. J. *RNA* **2011**, *17*, 469–477.

[243] Wochner, A.; Attwater, J.; Coulson, A.; Holliger, P. *Science* **2011**, *332*, 209–212.

[244] Horning, D. P.; Joyce, G. F. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 9786–9791.

[245] Eigen, M. *Naturwissenschaften* **1971**, *58*, 465–523.

[246] Rajamani, S.; Ichida, J. K.; Antal, T.; Treco, D. A.; Leu, K.; Nowak, M. A.; Szostak, J. W.; Chen, I. A. *J. Am. Chem. Soc.* **2010**, *132*, 5880–5885.

[247] Lawrence, M. S.; Bartel, D. P. *Biochemistry* **2003**, *42*, 8748–8755.

[248] Zaher, H. S.; Unrau, P. J. *RNA* **2007**, *13*, 1017–1026.

[249] Attwater, J.; Raguram, A.; Morgunov, A. S.; Gianni, E.; Holliger, P. *Elife* **2018**, *7*, e35255.

[250] Garbesi, A.; Capobianco, M. L.; Coloma, F. P.; Maffini, M.; Niccolai, D.; Tondelli, L. *Nucleosides & Nucleotides* **1998**, *17*, 1275–1287.

[251] Garbesi, A.; Capobianco, M. L.; Colonna, F. P.; Tondelli, L.; Arcamone, F.; Manzini, G.; Hilbers, C. W.; Aelen, J. M.; Blommers, M. J. *Nucleic Acids Res.* **1993**, *21*, 4159–4165.

[252] Sczepanski, J. T.; Joyce, G. F. *Nature* **2014**, *515*, 440–442.

[253] Joyce, G. F.; Visser, G. M.; van Boeckel, C. A.; van Boom, J. H.; Orgel, L. E.; van Westrenen, J. *Nature* **1984**, *310*, 602–604.

[254] Attwater, J.; Tagami, S.; Kimoto, M.; Butler, K.; Kool, E. T.; Wengel, J.; Herdewijn, P.; Hirao, I.; Holliger, P. *Chemical Science* **2013**, *4*, 2804–2814.

[255] Samanta, B.; Joyce, G. F. *Elife* **2017**, *6*, e31153.

[256] Horning, D. P.; Bala, S.; Chaput, J. C.; Joyce, G. F. *ACS Synth. Biol.* **2019**, *8*, 955–961.

[257] Forconi, M.; Herschlag, D. *J. Am. Chem. Soc.* **2005**, *127*, 6160–6161.

[258] Huang, F.; Yang, Z.; Yarus, M. *Chem. Biol.* **1998**, *5*, 669–678.

[259] Lau, M. W.; Unrau, P. J. *Chem. Biol.* **2009**, *16*, 815–825.

[260] Lau, M. W.; Cadieux, K. E.; Unrau, P. J. *J. Am. Chem. Soc.* **2004**, *126*, 15686–15693.

[261] Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920–930.

[262] Green, R.; Noller, H. F. *Annu. Rev. Biochem.* **1997**, *66*, 679–716.

[263] Ramakrishnan, V. *Cell* **2002**, *108*, 557–572.

[264] Fox, G. E. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a003483.

[265] Ling, J.; Reynolds, N.; Ibba, M. *Annu. Rev. Microbiol.* **2009**, *63*, 61–78.

[266] Ogle, J. M.; Ramakrishnan, V. *Annu. Rev. Biochem.* **2005**, *74*, 129–177.

[267] Kuncha, S. K.; Kruparani, S. P.; Sankaranarayanan, R. *J. Biol. Chem.* **2019**, *294*, 16535–16548.

[268] Lee, Z. M.; Bussema, r., C.; Schmidt, T. M. *Nucleic Acids Res.* **2009**, *37*, D489–493.

[269] Selmer, M.; Dunham, C. M.; Murphy, F. V. t.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. *Science* **2006**, *313*, 1935–1942.

[270] Monro, R. E.; Cerna, J.; Marcker, K. A. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *61*, 1042–1049.

[271] Ruan, X. L.; Li, S.; Geng, P.; Zeng, X. T.; Yu, G. Z.; Meng, X. Y.; Gao, Q. P.; Ao, X. B. *Med. Sci. Monit.* **2015**, *21*, 3048–3053.

[272] Liu, C. C.; Schultz, P. G. *Annu. Rev. Biochem.* **2010**, *79*, 413–444.

[273] Mohler, K.; Ibba, M. *Nat. Microbiol.* **2017**, *2*, 17117.

[274] Ogle, J. M.; Murphy, F. V.; Tarry, M. J.; Ramakrishnan, V. *Cell* **2002**, *111*, 721–732.

[275] Moazed, D.; Noller, H. F. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 3725–8.

[276] Nathans, D. *Fed. Proc.* **1964**, *23*, 984–989.

[277] Yarmolinsky, M. B.; Haba, G. L. *Proc. Natl. Acad. Sci. U. S. A.* **1959**, *45*, 1721–1729.

[278] Bain, J. D.; Diala, E. S.; Glabe, C. G.; Dix, T. A.; Chamberlin, A. R. *J. Am .Chem. Soc.* **1989**, *111*, 8013–8014.

[279] Noren, C. J.; Anthony-Cahill, S. J.; Griffith, M. C.; Schultz, P. G. *Science* **1989**, *244*, 182–188.

[280] Dougherty, D. *Curr. Opin. Chem. Biol.* **2000**, *4*, 645–652.

[281] Hohsaka, T.; Sisido, M. *Curr. Opin. Chem. Biol.* **2002**, *6*, 809–815.

[282] Neumann, H.; Wang, K.; Davis, L.; Garcia-Alai, M.; Chin, J. W. *Nature* **2010**, *464*, 441–444.

[283] Wang, K.; Neumann, H.; Peak-Chew, S. Y.; Chin, J. W. *Nat. Biotechnol.* **2007**, *25*, 770–777.

[284] Xie, J.; Schultz, P. G. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7*, 775–782.

[285] Jukes, T. H.; Osawa, S. *Experientia* **1990**, *46*, 1117–1126.

[286] Freeland, S. J.; Hurst, L. D. *J. Mol. Evol.* **1998**, *47*, 238–248.

[287] Lohse, P. A.; Szostak, J. W. *Nature* **1996**, *381*, 442–444.

[288] Zhang, B.; Cech, T. R. *Nature* **1997**, *390*, 96–100.

[289] Ouellette, R. J.; Rawn, J. D. *Organic Chemistry : Structure, Mechanism, Synthesis*, second edition. ed.; Academic Press: London ; San Diego, CA, 2018; pp viii, 1047 pages.

[290] Seelig, B.; Jaschke, A. *Chem. Biol.* **1999**, *6*, 167–176.

[291] Tarasow, T. M.; Tarasow, S. L.; Eaton, B. E. *Nature* **1997**, *389*, 54–57.

[292] Seelig, B.; Keiper, S.; Stuhlmann, F.; Jaschke, A. *Angew. Chem. Int. Ed. Engl.* **2000**, *39*, 4576–4579.

[293] Stuhlmann, F.; Jaschke, A. *J. Am. Chem. Soc.* **2002**, *124*, 3238–3244.

[294] Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. *Science* **2010**, *329*, 309–313.

[295] Carothers, J. M.; Oestreich, S. C.; Szostak, J. W. *J. Am. Chem. Soc.* **2006**, *128*, 7929–37.

[296] Buckley, B. R.; Page, P. C. B.; McKee, V. *Synlett* **2011**, *2011*, 1399–1402.

[297] Bulman Page, P. C.; Buckley, B. R.; Rassias, G. A.; Blacker, A. J. *European journal of organic chemistry* **2006**, *2006*, 803–813.

[298] Xulvi-Brunet, R.; Campbell, G. W.; Rajamani, S.; Jiménez, J. I.; Chen, I. A. *Methods* **2016**, *106*, 86–96.

[299] Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res* **2004**, *14*, 1188–90.

[300] Blanco, C.; Verbanic, S.; Seelig, B.; Chen, I. A. *Journal of Molecular Evolution* **2020**, *88*, 477–481.

[301] Shen, Y.; Pressman, A.; Janzen, E.; Chen, I. A. *Nucleic Acids Res* **2021**,

[302] Lai, Y.-C.; Liu, Z.; Chen, I. A. *Proc. Natl. Acad. Sci. USA* **2021**, *in press*.

[303] Bolger, A. M.; Lohse, M.; Usadel, B. *Bioinformatics* **2014**, *30*, 2114–20.

[304] Masella, A. P.; Bartram, A. K.; Truszkowski, J. M.; Brown, D. G.; Neufeld, J. D. *BMC Bioinformatics* **2012**, *13*, 31.

[305] Zhang, J.; Kobert, K.; Flouri, T.; Stamatakis, A. *Bioinformatics* **2014**, *30*, 614–20.

[306] Aita, T.; Husimi, Y. *J Theor Biol* **1998**, *193*, 383–405.

[307] Woese, C. R.; Dugre, D. H.; Dugre, S. A.; Kondo, M.; Saxinger, W. C. *Cold Spring Harb Symp Quant Biol* **1966**, *31*, 723–36.

[308] Szostak, J. W.; Bartel, D. P.; Luisi, P. L. *Nature* **2001**, *409*, 387–90.

[309] He, X.; Liu, L. *Science* **2016**, *352*, 769–770.

[310] Zhou, Q.; Xia, X.; Luo, Z.; Liang, H.; Shakhnovich, E. *J Chem Theory Comput* **2015**, *11*, 5939–46.

[311] Zhou, Q.; Sun, X.; Xia, X.; Fan, Z.; Luo, Z.; Zhao, S.; Shakhnovich, E.; Liang, H. *J Phys Chem Lett* **2017**, *8*, 407–414.

[312] Padiolleau-Lefevre, S.; Ben Naya, R.; Shahsavarian, M. A.; Friboulet, A.; Avalle, B. *Biotechnol Lett* **2014**, *36*, 1369–79.

[313] Guenther, U. P.; Yandek, L. E.; Niland, C. N.; Campbell, F. E.; Anderson, D.; Anderson, V. E.; Harris, M. E.; Jankowsky, E. *Nature* **2013**, *502*, 385–8.

[314] Kobori, S.; Nomura, Y.; Miu, A.; Yokobayashi, Y. *Nucleic Acids Res* **2015**, *43*, e85.

[315] Denny, S. K.; Bisaria, N.; Yesselman, J. D.; Das, R.; Herschlag, D.; Greenleaf, W. J. *Cell* **2018**, *174*, 377–390 e20.

[316] de Duve, C. *Nature* **1988**, *333*, 117–8.

[317] Lee, N.; Bessho, Y.; Wei, K.; Szostak, J. W.; Suga, H. *Nat Struct Biol* **2000**, *7*, 28–33.

[318] Chumachenko, N. V.; Novikov, Y.; Yarus, M. *J Am Chem Soc* **2009**, *131*, 5257–63.

[319] Leman, L.; Orgel, L.; Ghadiri, M. R. *Science* **2004**, *306*, 283–6.

[320] Danger, G.; Boiteau, L.; Cottet, H.; Pascal, R. *J Am Chem Soc* **2006**, *128*, 7412–3.

[321] Danger, G.; Michaut, A.; Bucchi, M.; Boiteau, L.; Canal, J.; Plasson, R.; Pascal, R. *Angew Chem Int Ed Engl* **2013**, *52*, 611–4.

[322] Danger, G.; Plasson, R.; Pascal, R. *Chem Soc Rev* **2012**, *41*, 5416–29.

[323] Biron, J. P.; Parkes, A. L.; Pascal, R.; Sutherland, J. D. *Angew Chem Int Ed Engl* **2005**, *44*, 6731–4.

[324] Leman, L. J.; Orgel, L. E.; Ghadiri, M. R. *J Am Chem Soc* **2006**, *128*, 20–1.

[325] Liu, Z.; Beaufils, D.; Rossi, J. C.; Pascal, R. *Sci Rep* **2014**, *4*, 7440.

[326] Liu, Z.; Rigger, L.; Rossi, J. C.; Sutherland, J. D.; Pascal, R. *Chemistry* **2016**, *22*, 14940–14949.

[327] Liu, Z. W.; Hanson, C.; Ajram, G.; Boiteau, L.; Rossi, J. C.; Danger, G.; Pascal, R. *Synlett* **2017**, *28*, 73–77.

[328] Ebhardt, H. A.; Unrau, P. J. *RNA* **2009**, *15*, 724–731.

[329] Zuker, M. *Nucleic Acids Res* **2003**, *31*, 3406–15.

[330] Lorsch, J. R.; Bartel, D. P.; Szostak, J. W. *Nucleic Acids Res* **1995**, *23*, 2811–4.

[331] Wilkinson, K. A.; Merino, E. J.; Weeks, K. M. *Nat Protoc* **2006**, *1*, 1610–6.

[332] Bershtein, S.; Mu, W.; Serohijos, A. W.; Zhou, J.; Shakhnovich, E. I. *Mol Cell* **2013**, *49*, 133–44.

[333] Bershtein, S.; Serohijos, A. W.; Bhattacharyya, S.; Manhart, M.; Choi, J. M.; Mu, W.; Zhou, J.; Shakhnovich, E. I. *PLoS Genet* **2015**, *11*, e1005612.

[334] Rodrigues, J. V.; Bershtein, S.; Li, A.; Lozovsky, E. R.; Hartl, D. L.; Shakhnovich, E. I. *Proc Natl Acad Sci U S A* **2016**, *113*, E1470–8.

[335] Louie, R. H. Y.; Kaczorowski, K. J.; Barton, J. P.; Chakraborty, A. K.; McKay, M. R. *Proc Natl Acad Sci U S A* **2018**, *115*, E564–E573.

[336] Otwinowski, J.; McCandlish, D. M.; Plotkin, J. B. *Proc Natl Acad Sci U S A* **2018**, *115*, E7550–E7558.

[337] Schuster, P.; Fontana, W.; Stadler, P. F.; Hofacker, I. L. *Proc Biol Sci* **1994**, *255*, 279–84.

[338] Lawrence, M. S.; Bartel, D. P. *RNA* **2005**, *11*, 1173–80.

[339] Lau, M. W.; Ferre-D'Amare, A. R. *Molecules* **2016**, *21*.

[340] Gravner, J.; Pitman, D.; Gavrilets, S. *J Theor Biol* **2007**, *248*, 627–45.

[341] Popovic, M.; Fliss, P. S.; Ditzler, M. A. *Nucleic Acids Res* **2015**, *43*, 7070–82.

[342] Moessner, R.; Ramirez, A. *Physics Today* **2006**, *59*, 24.

[343] Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. *Q Rev Biophys* **2014**, *47*, 285–363.

[344] Di Silvio, E.; Brunori, M.; Gianni, S. *Angew Chem Int Ed Engl* **2015**, *54*, 10867–9.

[345] Kluber, A.; Burt, T. A.; Clementi, C. *Proc Natl Acad Sci U S A* **2018**, *115*, 9234–9239.

[346] Anderson, P. W. *Proc Natl Acad Sci U S A* **1983**, *80*, 3386–90.

[347] Amitrano, C.; Peliti, L.; Saber, M. *J Mol Evol* **1989**, *29*, 513–525.

[348] Sasai, M.; Wolynes, P. G. *Proc Natl Acad Sci U S A* **2003**, *100*, 2374–9.

[349] Marshall, C. *Australian Journal of Zoology* **2014**, *62*, 3–17.

[350] Wolf, Y. I.; Katsnelson, M. I.; Koonin, E. V. *Proc Natl Acad Sci U S A* **2018**, *115*, E8678–E8687.

[351] Bendixsen, D. P.; Ostman, B.; Hayden, E. J. *J Mol Evol* **2017**, *85*, 159–168.

[352] Weinreich, D. M.; Watson, R. A.; Chao, L. *Evolution* **2005**, *59*, 1165–1174.

[353] Curtis, E. A.; Bartel, D. P. *RNA* **2013**, *19*, 1116–28.

[354] Mutschler, H.; Taylor, A. I.; Porebski, B. T.; Lightowlers, A.; Houlihan, G.; Abramov, M.; Herdewijn, P.; Holliger, P. *Elife* **2018**, *7*.

[355] Smail, B. A.; Clifton, B. E.; Mizuuchi, R.; Lehman, N. *RNA* **2019**,

[356] Gould, S. J.; Vrba, E. S. *Paleobiology* **1982**, *8*, 4–15.

[357] Voros, D.; Konnyu, B.; Czaran, T. *PLoS Comput Biol* **2021**, *17*, e1008634.

[358] Szathmary, E.; Smith, J. M. *Nature* **1995**, *374*, 227–32.

[359] Perona, J. J.; Hadd, A. *Biochemistry* **2012**, *51*, 8705–29.

[360] Tawfik, D. S.; Gruic-Sovulj, I. *FEBS J* **2020**, *287*, 1284–1305.

[361] Anantharaman, V.; Koonin, E. V.; Aravind, L. *Nucleic Acids Res* **2002**, *30*, 1427–64.

[362] Aravind, L.; Anantharaman, V.; Koonin, E. V. *Proteins* **2002**, *48*, 1–14.

[363] Aravind, L.; Mazumder, R.; Vasudevan, S.; Koonin, E. V. *Curr Opin Struct Biol* **2002**, *12*, 392–9.

[364] Fournier, G. P.; Andam, C. P.; Gogarten, J. P. *BMC Evol Biol* **2015**, *15*, 70.

[365] Haig, D.; Hurst, L. D. *Journal of Molecular Evolution* **1991**, *33*, 412–417.

[366] Goodarzi, H.; Nejad, H. A.; Torabi, N. *Biosystems* **2004**, *77*, 163–173.

[367] Zhu, W.; Freeland, S. *Journal of Theoretical Biology* **2006**, *239*, 63–70.

[368] Firnberg, E.; Ostermeier, M. *Nucleic Acids Research* **2013**, *41*, 7420–7428.

[369] Archetti, M. *J Mol Evol* **2004**, *59*, 258–66.

[370] Novozhilov, A. S.; Wolf, Y. I.; Koonin, E. V. *Biology Direct* **2007**, *2*.

[371] Massey, S. E. *Journal of Theoretical Biology* **2016**, *408*, 237–242.

[372] Attie, O.; Sulkow, B.; Di, C.; Qiu, W. G. *Plos One* **2019**, *14*.

[373] Wolf, Y. I.; Koonin, E. V. *Biology Direct* **2007**, *2*.

[374] Liu, Z.; Hanson, C.; Ajram, G.; Boiteau, L.; Rossi, J.-C.; Danger, G.; Pascal, R. *Synlett* **2017**, *28*, 73–77.

[375] Kobori, S.; Yokobayashi, Y. *Angew Chem Int Ed Engl* **2016**, *55*, 10354–7.

[376] Yokobayashi, Y. *Acc Chem Res* **2020**, *53*, 2903–2912.

[377] Trifonov, E. N. *J Biomol Struct Dyn* **2004**, *22*, 1–11.

[378] Zaia, D. A.; Zaia, C. T.; De Santana, H. *Orig Life Evol Biosph* **2008**, *38*, 469–88.

[379] Higgs, P. G.; Pudritz, R. E. *Astrobiology* **2009**, *9*, 483–90.

[380] Cleaves, n., H. J. *J Theor Biol* **2010**, *263*, 490–8.

[381] Longo, L. M.; Blaber, M. *Arch Biochem Biophys* **2012**, *526*, 16–21.

[382] Artymiuk, P. J.; Rice, D. W.; Poirrette, A. R.; Willet, P. *Nat Struct Biol* **1994**, *1*, 758–60.

[383] Pak, D.; Kim, Y.; Burton, Z. F. *Transcription* **2018**, *9*, 205–224.

[384] Mayr, H.; Ofial, A. R. *Angew Chem Int Ed Engl* **2006**, *45*, 1844–54.

[385] Johansson, M.; Zhang, J.; Ehrenberg, M. *Proc Natl Acad Sci U S A* **2012**, *109*, 131–6.

[386] Larson, M. H.; Zhou, J.; Kaplan, C. D.; Palangat, M.; Kornberg, R. D.; Landick, R.; Block, S. M. *Proc Natl Acad Sci U S A* **2012**, *109*, 6555–60.

[387] Lanier, K. A.; Petrov, A. S.; Williams, L. D. *J Mol Evol* **2017**, *85*, 8–13.

[388] Lanier, K. A.; Williams, L. D. *J Mol Evol* **2017**, *84*, 85–92.

[389] Gould, S. J.; Lewontin, R. C. *Proc R Soc Lond B Biol Sci* **1979**, *205*, 581–98.

[390] Kaltenbach, M.; Emond, S.; Hollfelder, F.; Tokuriki, N. *PLoS Genet* **2016**, *12*, e1006305.

[391] De Tarafder, A.; Parajuli, N. P.; Majumdar, S.; Kacar, B.; Sanyal, S. *Mol Biol Evol* **2021**,