**Title**
Measuring Biochemical Possibility Spaces in Evolutionary Engineering

**Permalink**
https://escholarship.org/uc/item/1p67z69w

**Author**
Pressman, Abe Daniel

**Publication Date**
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

# Measuring Biochemical Possibility Spaces in Evolutionary Engineering

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

Abe Daniel Pressman

Committee in charge:

Professor Irene Chen, Co-Chair

Professor Siddharth Dey, Co-Chair

Professor Michelle O'Malley

Professor Songi Han

March 2019

The dissertation of Abe Daniel Pressman is approved.

Michelle O’Malley

Songi Han

Irene Chen, Committee Co-Chair

Siddharth Dey, Committee Co-Chair

March 2019

Measuring Biochemical Possibility Spaces in Evolutionary Engineering

by

Abe Daniel Pressman

## Personal Acknowledgements

I wouldn’t have finished this project, when I did, without the assistance of an unusually large number of doctors. And unlike most graduate students, I can say that about both doctors of philosophy and medicine. It’s nice to be here.

I’ve also had quite a few friends during graduate school (perhaps to make up for the fact that y’all keep graduating), and at this point, the list of everyone who’s helped me out in some meaningful is rather long. Suffice to say, if you’re reading this, thank you. In the context of my dissertation, I should probably single out Elizabeth Decolvenaere and Greg Campbell as the two people most responsible for making sure I stayed in school. Over the past year, my mother has provided a great deal of support in a great many things, and I’ll always be thankful for that, even if she keeps mistakenly suggesting that I should take more vacations. And, of course, I wouldn’t be where I am without a ton of trust and assistance from my advisor Irene, who has always had the patience to sit and listen to a hundred of my ideas at a time and then expertly point out which two are actually good ideas and why.

Finally, I can occasionally display unusual skill at dropping, breaking, or otherwise screwing up expensive reagents and delicate experiments. A final thank you, then, to those of my coworkers who somehow put up with this during my first few years as a researcher. You know who you are.

VITA OF ABE DANIEL PRESMAN
March 2019

## Education

2012-2019 *Philisophiae Doctor* in Chemical Engineering (with Certificate in Bioengineering), University of California, Santa Barbara.

2008-2012 *Scientiae Baccalaureus* with Honors in Applied Math–Biology, Brown University.

## Publications from Work at UCSB

2019 **Pressman, A.**, Liu, Z., Janzen, E., Blanco, C., Muller, U.F., Joyce, G.F., Pascal, R, Chen, I.A. "Mapping a ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA." *In minor revision, J Am Chem Soc.*

2019 Blanco, C., Janzen, E., **Pressman, A.**, Saha, R., Chen, I.A. "Molecular Fitness Landscapes from High-Coverage Sequence Profiling." *Ann Rev Biophys*. (Solicited review)

2017 Gilmore, S.P., Henske, J.K., Sexton, J.A., Solomon, K.V., Seppälä, S., Yoo, J.I., Huyett, L.M., **Pressman, A.**, Cogan, J.Z., Kivenson, V., Peng, X. "Genomic analysis of methanogenic archaea reveals a shift towards energy conservation." *BMC genomics* **18**, 639.

2017 **Pressman, A.**, Moretti, J. E., Campbell, G. W., Muller, U.F., Chen, I.A. "Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences." *Nucleic Acids Res* **45**, 10922. (Breakthrough article)

2015 **Pressman, A.**, Blanco, C., Chen, I. A. **"**The RNA World as a model system to study the origin of life." *Curr Biol* **25**, R953-R963. (Solicited review)

## Planned Publications

2019 **Pressman, A.**, Chen, I.A. "Measuring fitness distributions through a generalized Fisher's theorem" (working title). *In final preparation.*

TBD Shen, Y., **Pressman, A.**, Chen, I.A., et al. Untitled paper on the methods and statistical limitations of the *k*-Seq/SCAPE system. *In early preparation.*

TBD **Pressman, A.**, Blanco, C., Janzen, E., Chen, I.A., et al. Untitled paper on epistasis in aminoacylation ribozymes. *In early preparation, awaiting final data set.*

## Presentations from Work at UCSB

2018 **Pressman, A.,** Janzen, E., Blanco, C., Müller, U.F., Chen, I.A. “Scape: Mapping Complete Fitness Landscapes for Engineered Gene Spaces” (presentation). *Synthetic Biology: Engineering, Evolution & Design (SEED)*, Scottsdale, AZ.

2017 **Pressman, A.,** Moretti, J.E., Campbell, G.W., Müller, U.F., Chen, I.A. **“**Estimating Ribozyme Kinetics from Analysis of *in vitro* Evolution” (presentation). *XVIIIth International Conference on the Origin of Life (International Society for the Study of the Origin of Life)*, San Diego, CA.

2017 **Pressman, A.,** Moretti, J.E., Campbell, G.W., Müller, U.F., Chen, I.A. **“**Estimating Ribozyme Kinetics from Analysis of *in vitro* Evolution” (presentation). *UC Chemistry Symposium*, Lake Arrowhead, CA.

2017 **Pressman, A.,** Liu, Z., Moretti, J.E., Janzen, E., Campbell, G.W., Müller, U.F., Pascal, R., Chen, I.A. “Estimating Ribozyme Kinetics from Analysis of *in vitro* Evolution” (poster). *Gordon Research Conference in Synthetic Biology*, Stowe, VT.

## Ambiguously-Prestigious Awards and Honors

2017 First Place, Dance Your Ph.D. Contest. (Role: Aerial dancer, storyboard consultant). Scherich, N, et al. “Representations of the Braid Groups.” *AAAS*.

2016 Personally insulted (for Pressman et al. 2015) on a creationist blog for the first time.

# ABSTRACT

Measuring Biochemical Possibility Spaces in Evolutionary Engineering

by

Abe Daniel Pressman

At the molecular level, artificial selection—controlling the forces of evolution to improve or design new biochemical functions— makes up one of our strongest tools for finding better biocatalysts, pharmaceuticals, and biosensors, as well as for studying the history and process of evolution itself. But fully harnessing evolution requires knowledge of the shape and dynamics of complete evolutionary spaces. Prior to this work, very little research existed comparing the real dynamics of artificial selection to any of the theoretical work that has been written to support it. By updating the classical theory of simple selections towards an engineering focus, and combining this with direct observations of direct evolving populations, my work has shown the first mathematical descriptions of how whole populations evolve during the selection of novel biocatalysts.

This work seeks to address the analysis of evolutionary fitness and chemical activity spaces at several levels. First, we offer a broad-ranging theoretical approach to mapping the distribution of fitness effects in any system under driven selection. Through both simulations and recent experimental data, we show that it is possible to estimate the initial distribution of fitness for nearly any selected population. In addition to potential applications in automated

gene engineering, this theoretical solution also makes it possible to approximate the overall distribution of any selectable chemical function across random molecular space, a necessary condition for theoretical optimization of nearly any *in vitro* selection.

Zooming in, we next develop tools to view an entire population of active catalysts and how it dynamically changes over the course of an entire selection. Working with a model selection for *de novo* RNA triphosphorylation catalysts, we develop a new high-throughput method to measure many active catalysts in parallel, building the first portrait of how tens of thousands of different functional molecules enrich or disappear over the course of an entire artificial selection. New heuristics for assessing the effectiveness of various activity-estimation methods allowed us to efficiently identify highly active ribozymes, as well as estimating catalytic activity without performing any additional experiments. We also present the first picture of non-ideality during a real selection, demonstrating that stochastic effects can be a powerful and quantifiable confounding factor on predicted selection dynamics. Finally, this analysis allows us to build the highest-resolution extant picture of a biocatalyst activity distribution, showing a catalytic activity that is log-normal, consistent with a mechanism for the emergence of activity as the product of many independent contributions.

Finally, we design our own model selection to investigate the evolution of a theoretical aminoacylase RNA whose existence may have been crucial to the origin of the genetic code. Using this system, we have developed techniques for Sequencing to determine Catalytic Activity Paired with Evolution (SCAPE), a comprehensive workflow that allows complete mapping of large, dynamic landscape of chemical activity. By measuring catalytic activity of millions of evolved biomolecules simultaneously, we pair kinetic variations with genetic sequence at single nucleotide resolution, building the first complete map of all

evolutionary pathways to an engineered function from anywhere in genetic space. The resulting map contains approximately six orders of magnitude more data than any previously-measured landscape of catalytic data, and suggests features of genetic epistasis and evolutionary ruggedness may be remarkably consistent across many unrelated biocatalysts with similar function. Our methods and results suggest general applicability to more complicated systems, as a viable alternative to the heuristic methods typically used to evaluate molecular selections, as well as validating a suite of capable tools for quantifying and optimizing the emergence of a wide range of evolvable biocatalytic functions.

# Table of Contents

**Personal Acknowledgements** ........ **iv**

**Vita of Abe Daniel Pressman** ........ **v**

**Abstract** ........ **vii**

**Table of Contents** ........ **x**

**Table of Figures** ........ **xii**

**Glossary of Terms, Abbreviations, and Symbols** ........ **xiii**

**1. Introduction: Directed evolution and fitness landscapes** ........ **1**

1.1. Motivation and Overview ........ 1

1.2. Functional biopolymers and where to find them ........ 4

1.3. Fitness distributions and fitness landscapes ........ 8

1.4. High-throughput measurements of evolutionary space ........ 13

1.5. The origin of life as a model system for engineering evolution ........ 17

**2. Novel theoretical approaches to studying fitness distributions** ........ **21**

2.1. Background: Mathematical models of simple evolutionary selection ........ 21

2.2. Theory: A further-generalized Fisher's Theorem can calculate initial fitness distributions ........ 24

2.3. Theory: Generalized Fisher's Theorem under discrete growth ........ 26

2.4. Theory: Selection under changing environmental conditions ........ 28

2.5. Theory: Deconvoluting fitness distributions and measurement noise ........ 31

2.6. Testing generalized Fisher's Theorem approach with simulated data ........ 33

2.7. Generalized Fisher's Theorem approach and real experimental data ........ 39

**3. Estimating biocatalyst kinetics from high-throughput selection data** ........ **44**

3.1. Background/Theory: Estimating chemical activity from selection enrichment ........ 44

3.2. Triphosphorylation selection shows surprising enrichment variability ........ 49

3.3. Theory: Heuristic evaluation of methods for estimating activity ........ 55

3.4. High-throughput activity estimation finds better ribozymes ........ 62

3.5. Consistency and noise in multiple-round selection enrichment ........ 66

3.6. Retrospective evolutionary analysis predicts an initial distribution of triphosphorylation activity ........ 71

3.7. Conclusion: A log-normal distribution of catalytic activity over random RNA space ........ 77

**4. Measuring ribozyme activity over an entire evolutionary space** ........ **81**

4.1. Background: Prior limitations of fitness landscape approaches ........ 81

4.2. Theory: SCAPE, *k*-Seq and “virtual arrays” ........ 83
4.3. Oxazolone aminoacylation selection results ........ 87
4.4. A large, quantitative fitness landscape and its evolutionary pathways ........ 93
4.5. Conclusion: SCAPE reveals a highly frustrated evolutionary network ........ 96
**5. Limitations of the *k*-Seq assay ........ 99**
5.1. The curious *k*-Seq conundrum for a TMP phosphorylase ........ 99
5.2. Improving *k*-Seq accuracy through library manipulation ........ 102
**6. Potential high-definition landscape topography ........ 105**
6.1. Background: Epistasis in real fitness landscapes ........ 105
6.2. Theory: Topography and ruggedness of fitness landscapes ........ 107
6.3. Possible epistatic analysis of a multi-peak ribozyme activity landscape ........ 110
6.4. Conclusion: Aminoacylation landscape may show surprising homogeneity ........ 115
**7. Future prospects and conclusions ........ 116**
7.1. Landscapes for molecular specificity and environmental conditions ........ 116
7.2. SCAPE and *k*-Seq tools ........ 118
7.3. Final Thoughts ........ 119
**Academic Acknowledgements ........ 121**
**Appendix A: Additional Protocols and Data ........ 122**
A.1. Data on tested sequences ........ 122
A.2. Identification of aminoacylation ribozyme sites for top motifs ........ 126
A.3. Epistatic correlation parameter for an aminoacylation landscape ........ 130
A.4. Refinements to the *k*-Seq calculation ........ 131
A.5. *k*-Seq protocols ........ 132
**References ........ 136**

## Table of Figures

Figure 1.1. General overview of *in vitro* selection protocols and their similarities ................. 7
Figure 1.2. General schematic of fitness landscape shapes and properties ........................... 13
Figure 2.1. Testing the general robustness of GFT fitness distribution predictions .............. 36
Figure 2.2. Using GFT approach to build initial fitness distributions of test cases ............... 37
Figure 2.2 continued. Using GFT approach to build initial fitness distributions of test cases 38
Figure 2.3. Testing GFT approach with TMP triphosphorylase selection ............................ 42
Figure 2.4. Testing GFT approach with oxazolone aminoacylase selection ......................... 43
Figure 3.1. Identified triphosphorylase families and clusters ................................................ 52
Figure 3.2. Enrichment correlation over clusters, across multiple rounds ............................ 53
Figure 3.2. continued: Enrichment correlation over sequences, across multiple rounds ........ 54
Figure 3.3. Fit correlation of multi-round fitness estimates .................................................. 61
Figure 3.4. Estimating catalytic rates for triphosphorylation ribozymes ............................... 64
Figure 3.5. Emergence of a notable mutation within a ribozyme family .............................. 65
Figure 3.6. Distribution of enrichment noise across TMPase selection ................................ 69
Figure 3.7. Fisher's Theorem analysis of TMP selection ..................................................... 70
Figure 3.8. Measured selection function for fitness distribution change in TMPase selection ........................................................................................................................ 74
Figure 3.9. TMPase fitness distribution and predicted distribution evolving across rounds .. 75
Figure 3.10. Testing back-prediction with different simulated fitness distributions ............. 76
Figure 3.11. An initial fitness distribution for TMPase ribozyme activity ............................ 80
Figure 4.1. *k*-Seq schematic .................................................................................................. 86
Figure 4.2. Results of aminoacylase selection ....................................................................... 90
Figure 4.3. Initial activity distribution for aminoacylation ribozyme space and families ...... 91
Figure 4.4. Comparing *k*-Seq to experimental activity measurements ................................. 91
Figure 4.5. The distribution of noise in *k*-Seq measurements ................................................ 92
Figure 4.6. Evolutionary pathways for oxazolone aminoacylase ribozymes ......................... 95
Figure 5.1. Kimitations of *k*-Seq analysis applied to TMP ribozyme selection .................. 102
Table 6.1. Explanation of epistasis terms used in this chapter ........................................... 109
Figure 6.2. Distribution of fitness effects across individual ribozyme families .................. 113
Figure 6.3. Distribution of epistatic effects across individual ribozyme families ............... 114
Table A.1. Activity of tested TMP ribozymes .................................................................... 122
Table A.2. Triphosphorylase families undergoing a notable mutational shift ..................... 123
Table A.3. Aminoacylation ribozymes chosen for gel-shift assay. ..................................... 124
Table A.4. Properties of pathways between aminoacylase ribozyme peaks ........................ 125
Figure A.5. Location of ribozyme oxazolone-aminoacylation sites ................................... 128
Figure A.5 continued. Location of ribozyme oxazolone-aminoacylation sites ................... 129
Figure A.6. Epistatic correlation may be consistent across aminoacylation ribozyme peaks ........................................................................................................................ 131

## Glossary of Terms, Abbreviations, and Symbols

As this is a highly-interdisciplinary work, covering a field filled with poorly-defined and overlapping terminology, this list errs towards the side of comprehensive. Where some terms may have multiple meanings in the literature, the specific definition used in this work is supplied.

**Abundance ($\alpha_{R,i}$).** The fraction of a population at selection round *R* made up entirely of sequence *i*.

**Activity landscape.** Similar to a fitness landscape, but with direct chemical activity used as the “phenotype” instead of evolutionary favorability. Can be converted into a fitness landscape, and vice-versa, only if the dynamics of the system are sufficiently understood.

**Alphabet.** The set of different monomers that can be used in the construction of a sequential biopolymer. RNA and DNA are each constructed from a library of four nucleotides, while proteins are polypeptides constructed from a library of twenty amino acids.

**Amino-oxazolone.** A prebiotically-plausible source of activated amino acids, used in the selection for oxazolone aminoacylation ribozymes described in Figure 1.1E and carried out and analyzed in Chapter 4.

**Aptamers.** RNA or DNA affinity reagent that binds to a specific ligand. While usually found through *in vitro selection*, examples of aptamers do exist in nature.

**Artificial selection.** The use of specific evolutionary pressure to evolve a molecule or organism towards one specific purpose.

**BTO.** Biotinylated methyl tyrosine oxazolone, the amino-oxazolone substrate used in the aminoacylation ribozyme selection described in Chapter 4.

**Catalytic ratio ($r_i$).** The catalytic enhancement of initial rate over an uncatalyzed sequence, calculated as $r_i = k_iA_i/k_0A_0$.

**Coverage.** In the context of *in vitro* selection, this refers to the expected number of copies expected to be present in the initial pool for any specific molecular sequence.

**Cumulant-generating function (CGF; $K_t(x)$).** A function derived from a variable’s PDF, such that its successive derivates at $s = 0$ give the cumulants of that PDF. Described in equation (2.7).

**Deep mutational scanning (DMS).** A technique for completely probing a local fitness space, in which all possible single- or double-mutants of a wild-type sequence are present in the same starting population.

**Directed evolution.** Artificial selection that involves a mutagenesis step, either at the beginning or during selection, to drive a molecule or organism across local fitness space and towards a functional optimum.

**Dissociation constant ($k_{D,i}$).** The ligand concentration of ligand at which half of an antibody or aptamer *i* is expected to bind. Lower value indicates higher binding affinity.

**Distribution of Fitness Effects (DFE).** A probability distribution describing the ratio of different fitness values present across an entire population of non-homogenous organisms mutated from a wild-type sequence. Effectively, the localized fitness distribution of variations to a core sequence.

**Enrichment noise.** The fraction by which a particular sequence's enrichment in one round differs from what is expected. **Abundance-dependent noise** follows a pattern of increasing noise (and enrichment unpredictability) at decreasing abundance; **abundance-independent noise** is the component of enrichment noise that appears constant in a selection system, regardless of sequence abundance.

**Epistasis ($\varepsilon_{a,b}$).** An interaction between multiple sites in a biopolymer sequence such that the effect of one mutation depends on the state of another. Epistasis as defined in this work is calculated as in equation (6.1), with specific classes of epistasis defined in Table 6.1.

**Epistatic correlation ($\gamma_d$).** The average correlation activity effects of single mutations between sequences at evolutionary distance *d* of each other, as described in Appendix A.3. A general ruggedness parameter.

**Estimated enrichment, fitness ($E_e$, $F_e$).** An estimate of $E_{R,i}$ or $F_i$, derived from a multi-round HTS fitting process, as described in section 3.3. $E_e$ values are expected to be rescaled by an unknown factor; calculating $F_e$ requires a kinetic fit from selection under multiple selection conditions in order to find the correct rescaling factor.

**Exponential distribution.** A probability distribution P(*x*) which decays exponentially. P(*x*) appears as a straight line on logarithmic-*y*-axis plot.

**Family.** A collection of sequences all within a certain evolutionary distance of each other, function as a suspected fitness/activity peak. In this work, families are grouped by similarity across their entire randomized sequence region; can be further split into sequence clusters.

**Fisher's Theorem (FT).** Hyperbolically termed the "Fundamental theorem of natural selection" by R.A. Fisher, Fisher's theorem describes how a population's fitness mean and variance interact over time, described in equation (2.3). Can be used as a test of how ideally the shape of a fitness distribution is evolving.

**Fitness (*x*).** A quantitative measure of evolutionary favorability. Traditionally, can refer to various specific measurements. In this work, especially the mathematics of Chapter 2, we use the term "Fitness" or the variable *x* to denote **reproductive fitness** — that is, an organism or molecule's overall rate of reproduction relative to a wild-type or uncatalyzed baseline.

**Fitness distribution ($p_t(x)$ or $p_R(x)$).** A probability distribution describing the ratio of different fitness values (*x*) present across an entire population at some point during a selection. Over the course of a selection, its fitness distribution continually changes as a function of time or generation. **Initial Fitness Distribution** describes the distribution of fitness at the start of a selection; in the case of selection for a *de novo* function from a randomized pool, it also represents the total distribution of chemical activity over the random biopolymer space in question.

**Fitness landscape.** A mapping of sequence space "genotype" to a fitness value "phenotype," describing evolutionary favorability of all possible sequences in an evolutionary system.

**Fitness peak.** A group of sequences with elevated fitness, centered around a local optimum, and physically close to each other in terms of evolutionary distance. The most favorable sequence in a peak is called the **peak center**.

**Generalized Fisher's Theorem (GFT).** The application of Fisher's Theorem to higher-order cumulants, described in equation (2.4). **GFT analysis** refers to applying a further-generalized form of this, described in equation (2.11), towards the prediction of initial fitness distributions.

**High-Throughput Sequencing (HTS).** Any of the many DNA sequencing techniques formerly called "Next-Generation"; typically capable of reading on the order of $10^6$-$10^8$ individual DNA sequences.

***k*-Seq.** The "kinetic sequencing" protocol used to quantify ribozyme kinetics through a "virtual array" approach at the endpoint of our aminoacylation ribozyme selection. The *k*-Seq methodology uses multiple rounds of parallel selection, under different chosen selection conditions, to build a model of chemical activity for every single sequence.

**Kinetic rate constant ($k_i$).** As used in this work, refers to the reaction rate of a catalytic first-order, self-modifying ribozyme or enzyme.

**Log-normal distribution.** A probability distribution P($x$) in which ln($x$) is normally distributed. P($x$) appears as a parabola on a log-log plot. In certain cases, can collapse into a Pareto distribution.

**Maximum activity constant ($A_i$).** The fraction of molecules of a ribozyme or enzyme $i$ expected to fold and function correctly.

**Mean fitness ($\mu(t)$ or $\mu(R)$).** The average fitness of all organisms after a period of growth corresponding to $t$ wild-type generations or $R$ fixed cycles of selection.

**Motif.** A pattern of conserved sites, repeated across many sequences in a selected pool. May be substantially shorter than the overall randomized region. In this work, refers to sequence patterns that may be found across multiple families; each motif is suspected to correspond to a specific reaction mechanism.

**mRNA Display.** A cell-free translation method that uses *in vitro* enzymes to couple synthesized polypeptides directly to their own mRNA, allowing proteins to be used as a pulldown tag for the genetic information that synthesized them. This, along with other similar methods, can be used to effectively implement SELEX experiments in protein-based systems.

**Notable mutations.** In this analysis, describes a sequence that "takes over" an evolving fitness peak, by displacing another sequence as the most prominent. Used here to subdivide families into smaller sequence clusters.

**Pareto distribution.** Also known as the **scale-free distribution**; a probability distribution P($x$) which displays self-similarity. P($x$) appears as a straight line on a log-log plot.

**Peptide aptamer.** Short, protein-based biopolymer selected to bind strongly to a single target. Functionally similar to antibodies, but often designed for more *in vitro* uses, and easier to select for.

**Phage display.** A common procedure for *in vitro* directed evolution of antibodies and other proteins. Bacteriophage populations are transformed with vectors expressing a specific protein, then separated by affinity to a specific target, in a process similar to (but with more complex biology and kinetics than) SELEX or mRNA display.

**Polymerase Chain Reaction (PCR).** A method for rapidly copying DNA sequences, using *in vitro* polymerase enzymes in a cell-free environment.

**Probability distribution function (PDF).** The distribution function of a random variable, integrating to 1 over the range of that variable.

**Reaction probability/fraction ($F_i$).** The fraction of all copies of sequence $i$ expected to undergo successful binding or chemical modification during the selection step of a SELEX or other similar *in vitro* selection experiment. Ranges between 0 and 1; also known as **selection fitness**, and exists as a linear rescaling of reproductive fitness $x$.

**Relative enrichment ($E_{R,i}$).** The rate at which a sequence $i$ increases in abundance from selection round $R$ to $R + 1$. Exists as equal to $F_i$ divided by the mean reaction probability of the population at round $R$.

**Reverse Transcription (RT).** A tool that uses specialized reverse transcriptase enzymes to convert RNA sequences into DNA sequences, optionally regenerating promoter sites through the use of overhanging primers. In **RT-PCR**, can be coupled with PCR and transcription to rapidly copy a population of RNA sequences.

**Ribozymes.** RNA polymer sequence that catalyzes a specific reaction. Related to **deoxyribozymes**, i.e. DNA catalysts, which are far rarer in both living organisms and laboratory settings.

**Round ($R$).** The number of selection cycles that have occurred, in an artificial selection with fixed cycles of selection and copying (such as any SELEX-type *in vitro* selections). $R$=0, or “Round 0,” refers to the initial population before selection begins.

**Ruggedness.** The property of an epistatic, uncorrelated landscape in which many small local peaks and valleys are present.

**SCAPE.** Sequencing to measure catalytic activity paired with *in vitro* evolution. A workflow combining high-coverage selection with an HTS-based screen, in order to identify and measure the chemical kinetics of all high-activity sequences present in an entire sequence space.

**Selection coefficient ($1 - x$).** An organism’s fitness advantage over the wild-type. In population genetics, this is often confusingly also referred to as fitness.

**SELEX.** Selection of Ligands by Exponential Enrichment: an *in vitro*, cell-free selection scheme originally used to select novel aptamers, where a population of sequences is separated via binding to a fixed substrate and the survivors are then exponentially copied through PCR.

**Sequence cluster.** In Chapter 3, describes a subdivision of some sequence families. Here, it refers to all sequences closer to a specific center than to any other centers, within a maximum distance.

**Sequence count ($n_{R,i}$).** Also referred to as **copy number**. The absolute number of copies of a sequence $i$ detected by HTS in a population at selection round $R$.

**Sequence space.** The set of possible variations of a biopolymer, forming a space of $N$ dimensions and $M^N$ complexity, where $N$ is the number of variable sites and $M$ is the size of the variable alphabet.

**Sequence.** In the context of templated biopolymers, a molecule's sequence is its specific ordering of biological monomers, which can be easily compared through sequencing or used to synthesize specific molecules. As nucleotides serve as a template to their own replication, an individual molecule's sequence can be preserved across many rounds of copying. In the case of *in vitro* selection, can also be used to refer to a specific molecular "genotype."

**Substrate concentration ([$S$]).** The concentration of ribozyme/enzyme substrate or antibody/aptamer ligand present in a reaction.

**Time ($t$).** In the context of a continuously growing population (Chapter 2), refers to the number of wild-type generations that have elapsed. In the context of ribozyme kinetics (Chapters 3-5), refers to the length of time a reaction is allowed to incubate.

**Trimetaphosphate (TMP).** A prebiotically-plausible triphosphate source, used in the selection for TMP triphosphorylation ribozymes described in Figure 1.1D, carried out by collaborators, and analyzed in Chapter 3.

**Uncatalyzed rate ($k_0A_0$).** In the case of selection from a random library, the initial reaction rate of a random, non-catalytic sequence.

**Uniform stepwise distribution.** A probability distribution P($x$) consisting of multiple equal-width "steps," each with a fixed height.

**Wild-type (wt).** The un-mutated version of an organism or sequence. In the case of local fitness effects or distribution, we use this term to refer to a central sequence, with fitness set to 1, whose activity is used as a baseline for comparison to other sequence variants.

# 1. Introduction: Directed evolution and fitness landscapes

*"Be sparing in publishing theory. It makes persons doubt your observations."*

– Charles Darwin, *Letter to John Scott, 6 June 1863*

## 1.1. Motivation and Overview

Over the past few decades, the use of evolution on-demand to design and improve specific functional biomolecules has become an important tool in medicine and the life sciences.[2-4] But as engineers find new ways to harness artificial selection, our understanding of the chemical and biological dynamics involved have lagged far behind. While the concept of controlling and directing evolution is not a new one, going back all the way to the invention of agriculture, it has radically changed in the 21st century. Methods like Polymerase Chain Reaction (PCR), high-throughput sequencing (HTS) and cell-free translation have allowed artificial selection to move from breeding populations of living organisms with specific traits to directly evolving populations of functional molecules. But much of our understanding of artificial selections at the molecular level is based on untested theory, or on math tailored to the study of traditional evolving organisms. This work aims to change that, by shining a light into what specific evolutionary spaces actually look like during controlled evolution of a single given biomolecule.

Artificial selection, as it is used by biochemical engineers, usually involves either single-celled organisms selected for a specific purpose or molecules selected entirely *in vitro*. Cell-free *in vitro* selections, which generally involve only a handful of biochemical reagents and catalysts, represent the simplest possible case of evolution. And while this simplicity makes such selections ideal model systems for studying evolutionary design, that does not mean they lack usefulness. In addition to improving the activity of many functional proteins, *in vitro* selection has discovered novel peptide and nucleotide biopolymers with useful properties for catalysis,[5-11] diagnostics,[12-15] targeted drug therapies,[3,16-19] and custom gene modulation.[5,20] Introducing mutation to selections in a controlled manner gives directed evolution, which can improve or change the functionality of a variety of biological molecules[10,20-24] and give insight into possible evolutionary pathways over the history of life[25,26] or the development of antibiotic resistance,[27,28] allowing such synthetic approaches to answer questions related to both traditional biochemistry and biochemical engineering. But selections often fail to produce molecules of improved activity, require more time than predicted to reach such a state, or simply behave erratically; existing theory does little to anticipate such difficulties or offer solutions. Research combining selection theory with actual observations of molecular populations evolving was virtually nonexistent my research began, and is still extremely sparse, even though physical measurements derived from such studies should be able improve predictions of selection success and optimize conditions for a particular set of selections.[25,29,30] The goal of my thesis research is **to 1) invent and improve methods for mapping out the fitness distributions and landscapes on which evolutions occur; 2) develop the first extremely large landscape of chemical activity for a biopolymer catalyst** (estimating the activity of a space with ~$10^{12}$ molecular variations vs previous work limited to around $10^4$ variations); and **3) use this and other data to find general insight into the evolvability of biomolecular catalysts**, as well as to find mathematical approximations necessary to update existing models of *in vitro* selection, such as the starting distribution and non-ideality effects that may be present in common selections. Broadly, our data also form a quantitative framework to **4) address how evolutionary**

**landscapes constrain *de novo* emergence of catalytic function, providing the first map and likely pathways for all possible routes by which a biocatalyst can evolve from random molecular space.**

The bulk of this work has been broken down into four chapters of this thesis (chapters 2, 3, 4, and 6) roughly representing a set of four first-author publications, two of which are complete or submitted and two of which I expect to see completed shortly after my PhD is finished. While not strictly presented in the order that research was carried out, these chapters are instead intended to create a narrative, and presented in order of decreasing mathematical abstraction and increasing experimental complexity.

To meaningfully accomplish what might otherwise be overly-broad goals, the experimental focus of my work is on two selections for specific RNA-based catalytic functions. These examples are chosen not for specific bioengineering applications but instead because they are reactions potentially important to the origin of life on earth: even basic knowledge about their evolutionary landscapes and possibilities provides valuable insight into the pathways and speed with which early life might have evolved. And early-earth ribozymes make attractive model systems for directed evolution, as their biology and kinetics are both extremely simple. But the analysis and tools involved in this report should be equally applicable to any self-modifying biocatalyst. With a few changes to the kinetics, described at relevant points in the text, the entire body of work should be easy to extrapolate to most aptamer, riboswitch, or cell-free protein selections, as well as to some cell-based selections of sufficient simplicity. Very little is known about the shapes of real, large fitness landscapes for any biochemical function, and while this research finds and describes one specific example, repeating the process for many different types of reactions may give a better understanding of what is “normal” for the selection and study of novel bio-reagents. To that end, one remaining goal of this project, which I hope to finish in the near future (with help from other researchers who will be taking over the project) is development of all involved analyses into a set of packaged software tools to distribute.

## 1.2. Functional biopolymers and where to find them

Roughly, most molecules targeted by evolution fall into one of four categories: antibodies, aptamers, enzymes, and ribozymes. (Section 3.1. presents simple equations for the kinetics of aptamer and first-order ribozyme selections, which may be seen as analogous to those for similar antibodies and aptamers selected through cell-free methods such as mRNA display. The mathematics for more complicated selections, while likely still tractable, fall beyond the scope of this work.)

Antibodies, a natural component of the vertebrate immune system, are highly specialized proteins diversified through a recombination process, with the goal of binding and recognizing one specific molecule. As a functional protein, antibodies consist of peptide polymers containing a long fixed structural region and several shorter, variable regions which can bind to their target. Due to their (often) high affinity and specificity, antibodies are common and popular biological stains and detector molecules, as well as making up a large and growing component of the pharmaceutical market.[2] Originally, all biologically relevant antibodies were generated by model organisms, whose natural immune systems often improve their own antibodies through an evolutionary process called antibody maturation; in the past several decades, technologies such as phage display have managed to hijack antibody evolution, allowing the improvement or *de novo* selection of antibodies without a host organism.[31] Antibodies are some of the most profitable biomolecules; for context, one single antibody pharmaceutical, Adalimumab, accounts for approximately $5 billion in yearly sales, and is a product of *in vitro* evolution.[2,32] Other, simpler binding proteins have also been selected, such as "peptide aptamers" in which a small polypeptide is partially or entirely randomized, then selected for binding activity; these are usually much smaller and far more flexible in their structural requirements than antibodies.[33] Antibody selections generally require expression by engineered cells, leading to selections that require a delicate living component; peptide aptamer selections often use cell-free translation machinery, but the selection schemes involved are still more complicated than those for nucleotide aptamers.

Enzymes, as protein-based catalysts, are perhaps the most functionally diverse of biological molecules, varying widely in many features, including post-transcriptional modifications or chaperoned folding. The latter can be difficult to preserve in evolutionary experiments, and so directed evolution of proteins often involves small changes or experimental schemes tailored towards a single molecule.[34] Enzyme selections are often fairly complex, and can vary widely in their reaction kinetics and chemical function. Still, fairly simple enzymes can often be modeled with basic first-order kinetics, and *in vitro* selection methods such as mRNA display can closely mirror simpler ribozyme selections.[34,35]

Aptamers and ribozymes, as the nucleotide-based analogs of antibodies and enzymes, provide slightly simpler examples of molecular evolution. DNA and RNA aptamers, as nucleotide polymers with a partially or completely-random sequence, need only to bind a fixed molecular target; a wide variety of washing and pulldown methods exist to separate bound sequences from unbound ones, with DNA polymerases and reverse transcriptases used to easily copy the surviving molecules. Unlike in proteins, the ability of functional nucleotides to template their own reproduction has led to *in vitro* selections with a minimal number of steps. And nucleic acids contain a much smaller polymer alphabet (4 monomers vs. peptides' 20 monomers), leading to fewer possible molecules to study, in a combinatorically-smaller possibility space, than for similarly-sized peptides. Aptamers were the first main products of Systematic Evolution of Ligands by Exponential Enrichment (SELEX), a general scheme where selection requires repeated cycles of two broad steps. In the selection step of SELEX, a large sequence pool is subjected to a winnowing process where aptamers that bind a target are retained while the rest are discarded; in the replication step, the remaining sequences are amplified to regenerate a full population through DNA replication (Figure 1.1).[36,37] In this way, higher-fitness sequences "enrich" themselves in the pool, growing in number from one round to the next, while lower-fitness sequences are selected against, decreasing in number and eventually disappearing entirely. SELEX methods have effectively served as a template for methods like mRNA display that have simplified certain peptide aptamer and enzyme selections.

And finally, ribozymes, as nucleic acid-based catalysts, likely served as the first functional biomolecules on earth, providing special relevance in the study of how *de novo* functions evolve.[38] Due to the wide repertoire of enzymes already available in nature, engineered ribozymes have seen fairly limited use, as selection for novel ribozyme function is often a riskier bet than optimizing an existing biocatalyst. While ribozyme selections can be as complex as those for equivalent enzymes, many benefit from modified SELEX methods, by attaching an affinity tag or nucleic acid primer sequence necessary for their own copying.[39] Thus, the primary use of ribozymes may be as targets of academic study: a multitude of chemical functions allows them to answer more questions about evolution than single-function aptamers, while ribozymes are often far easier to evolve and characterize than similar proteins. As some ribozymes can contain extremely small catalytic motifs,[40] these are also an ideal target system for studies that wish to analyze polymer configuration spaces of limited complexity.

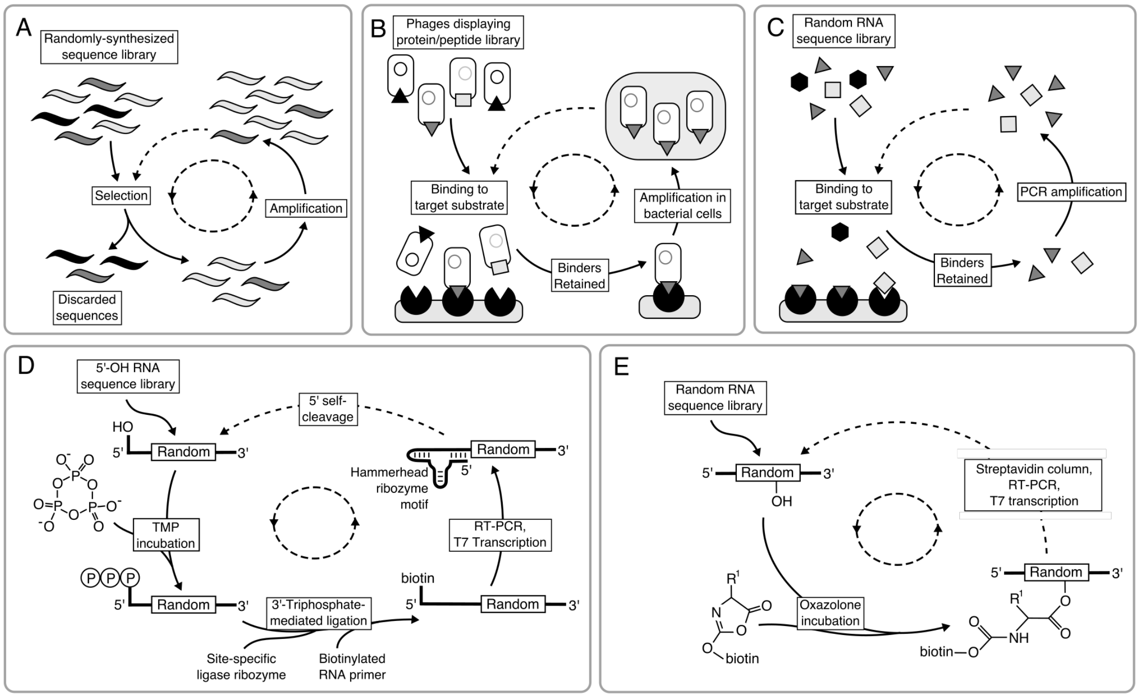

A
Randomly-synthesized sequence library
Selection
Amplification
Discarded sequences
B
Phages displaying protein/peptide library
Binding to target substrate
Amplification in bacterial cells
Binders Retained
C
Random RNA sequence library
Binding to target substrate
PCR amplification
Binders Retained
D
5'-OH RNA sequence library
HO
5'
Random
3'
5' self-cleavage
Hammerhead ribozyme motif
Random
3'
5'
TMP incubation
RT-PCR, T7 Transcription
P P P
5'
Random
3'
3'-Triphosphate-mediated ligation
biotin
5'
Random
3'
Site-specific ligase ribozyme
Biotinylated RNA primer
E
Random RNA sequence library
5'
Random
3'
OH
Streptavidin column, RT-PCR, T7 transcription
5'
Random
3'
Oxazolone incubation
R1
N
O
O
biotin
biotin
NH

**Figure 1.1. General overview of *in vitro* selection protocols and their similarities**
Schematic demonstrating the parallels between various methods of artificial selection mentioned in this report. **(A)** In the central scheme of in vitro selection, selections proceed through a cycle of selection and amplification steps. **(B)** In phage display techniques, proteins are expressed in a viral vector, selected for binding ability, and replicated within cells. **(C)** In traditional SELEX, aptamers or modified ribozymes are selected based on binding to a substrate, with DNA PCR or RT-PCR and translation used to replicate sequences. **(D)** In the TMP selection (Chapter 3), triphosphorylated ribozymes were a valid substrate for ligation to a biotinylated primer, allowing affinity-based partitioning of triphosphorylating sequences. **(E)** In the oxazolone aminoacylase selection, sequences covalently linked to an amino acid precursor display biotin, allowing a biotin-affinity Streptavidin column to isolate aminoacyl-RNA sequences displaying a biotin tag.

### 1.3. Fitness distributions and fitness landscapes

In principle, it would be possible to map out the trajectory of any directed evolution experiment if one knew how each new mutation would affect the activity of the targeted organism or molecule. (Such a prediction would be fundamentally probabilistic, and would also depend on experimental parameters such as population size and mutation rate). With such knowledge, it would be possible to design and optimize selection conditions for nearly any desired biomolecule. But this is fundamentally self-defeating: knowing the kinetics of every possible variant of a biocatalyst or affinity reagent would also make it possible to instantly select the best one for a given task, eliminating the need for evolutionary design or selection.

Of course, such perfect knowledge would also require kinetic testing of a staggering number of different molecules. For an RNA-based catalyst of 20 nucleotides, there are $4^{20} \approx 10^{12}$ possible monomer arrangements, each with unique chemical properties; for an enzyme of 20 amino acids, there are $20^{20} \approx 10^{26}$—simply synthesizing one copy of each possible 20-mer would result in approximately 500 metric tonnes of protein. Obviously, such knowledge quickly becomes impossible for all but the smallest functional biopolymers. Instead, we turn to approximate measurements of the shape of evolutionary spaces, which—based on the chosen level of abstraction—let us predict what evolution may look like when carried out for a specific function, or for related functions which we expect to evolve similarly. Broadly, these approximations can be characterized as fitness distributions, fitness landscapes, and epistatic approaches.

First, we must define some terms. Central to all of biology is “fitness,” an organism’s ability to survive and reproduce. Essentially, fitness is a single quantitative value—how well a selected molecule does its job and thus survives selection—that can be used as the "phenotype" of an artificially selected species. In classical evolutionary biology, fitness is a product of an organism's environment and community; in ideal artificial selection environments, survival is usually closely tied to its ability to carry out one specific chemical reaction or bind to a particular substrate. Unfortunately, in the study of artificial selection, the term "fitness" has been used to describe a range of specific measurements, from a species's direct activity to its rate of survival at the end of a selection.[41-43]This work specifically defines fitness as a species or sequence's reproductive rate relative to some standard, as will be discussed further in sections 2.1 and 3.1, but the mathematics of fitness landscapes can be applied to any definition that provides a single-variable measurement of molecular activity or effectiveness.

For selection on a single molecule or gene, every possible configuration of monomers corresponds to exactly one sequence. A "sequence" is the specific ordering of nucleic acids or amino acids that give proteins and functional nucleic acids their structure and chemical activity, and is easily represented by a string of characters. Often, directed evolution experiments only target one specific portion of a biomolecule for randomization and evolution, and sometimes a species's "sequence" refers specifically to this variable region. The set of possible variations of a biopolymer are referred to as its "sequence space," forming a space of $N$ dimensions and $M^N$ complexity, where $N$ is the number of variable sites and $M$ is the size of the variable "alphabet" ($M$=4 for most nucleic acid and =20 for most peptide selections).

As sequence space is unfathomably vast for all but the smallest biomolecules, abstractions are needed to study whole possible evolutionary spaces. The simplest of these is perhaps the distribution of fitness effects (DFE)[1], a probability distribution describing the ratio of different fitness values present across an entire population of non-homogenous organisms. In artificial selections, we can simplify this further, to a “fitness distribution,” which is simply the rate at which different fitness values appear across an evolving population. The fitness distribution at the start of a selection for *de novo* function is an “initial fitness distribution” the rate at which different fitness values appear across the entirety of a single functional molecule’s sequence space. Initial fitness distributions are necessary to optimize the parameters of *in vitro* selection[44], but prior to the start of this work remained a purely theoretical consideration. From a more chemical concern, fitness distributions can be converted into and back from distributions of chemical activity, if the kinetic relationship between fitness and activity is sufficiently understood. Advances over the past few years have allowed measurement over extremely small fitness and activity spaces, as described in section 1.4, but so far our only measures of any initial distributions over spaces larger than $10^6$ sequences come from estimation methods presented in chapters 3 and 4 of this work. Selection theorists, meanwhile, have proposed a variety of possible functions that may describe how catalytic function is distributed across biopolymer space; section 3.6 of this work describes several of these proposals, and how they line up with actual data from a large-scale measured distribution of chemical activity.

A more detailed way to study evolutionary spaces is fitness landscapes, which are the direct mapping of sequence space to fitness, taking the form of a single-valued function of *N* variables.[45,46] The mutation and selection components of evolution can effectively be seen as a random walk on a fitness landscape with a bias toward climbing hills of high fitness. [47] Mapping a fitness landscape perfectly means measuring the kinetics of every possible molecular variation, and has historically only been possible for extremely small sequence spaces. As discussed in the next section, developments in DNA sequencing technology have begun to slowly expand this capacity, but it is still possible to extrapolate features from partial or sparsely-sampled fitness landscapes. Typically, fitness distributions consist of one or many "peaks," each centered around a different high-activity chemical mechanism, with locally-maximal ribozymes surrounded by similar sequences whose fitness decreases with increasing sequence distance. (Distance, across sequence space, is usually measured as the number of single mutations or "edits" required to change one sequence into another). A general overview of the number and width of peaks in a given sequence space could quickly inform whether a selection over that space is likely to fail, and ways to offset such an outcome.[44,48] As with fitness distributions, fitness landscapes can theoretically be converted into activity landscapes, which describe the direct chemical kinetics of every sequence, allow understanding of fitness over a wider range of chemical conditions. The shape, steepness, and smoothness/ruggedness of such peaks in any such landscape and the shortest pathways between them are parameters that can be used to address classic questions such as: How repeatable are evolutionary outcomes? Do there exist neutral evolutionary pathways connecting different mechanisms with the same chemical activity?  Landscape approaches, though still in their infancy, are thus applicable to almost every evolutionary process ranging from the simplest possible life to complex functional aptamers and ribozymes.

Finally, a third major way in which evolutionary spaces are often described is epistatic analysis. Epistasis is a measure of the combinatorial effects of multiple mutations, while ruggedness is any general metric of local landscape topography, often based on summed epistasis data. Taken together, these form a set of measurable parameters which can describe the level of smoothness and interconnectivity of sites across an entire fitness landscape, or across individual peaks within a landscape. Knowledge of epistasis and ruggedness in a fitness landscape could be used to fine-tune parameters such as the mutation rate necessary for a given function to optimally evolve.[29,49,50] Such analysis is discussed further in Chapter 6—while it is not a primary concern in this report, our large fitness landscape approach may allow epistatic analysis at an unprecedented new level, as suggested by preliminary epistasis study. Currently, it is not known to what extent epistasis and ruggedness remain conserved or vary across different portions of any evolutionary landscape, but section 6.4 describes how the question of homogeneity in fitness landscape topography may be answered.

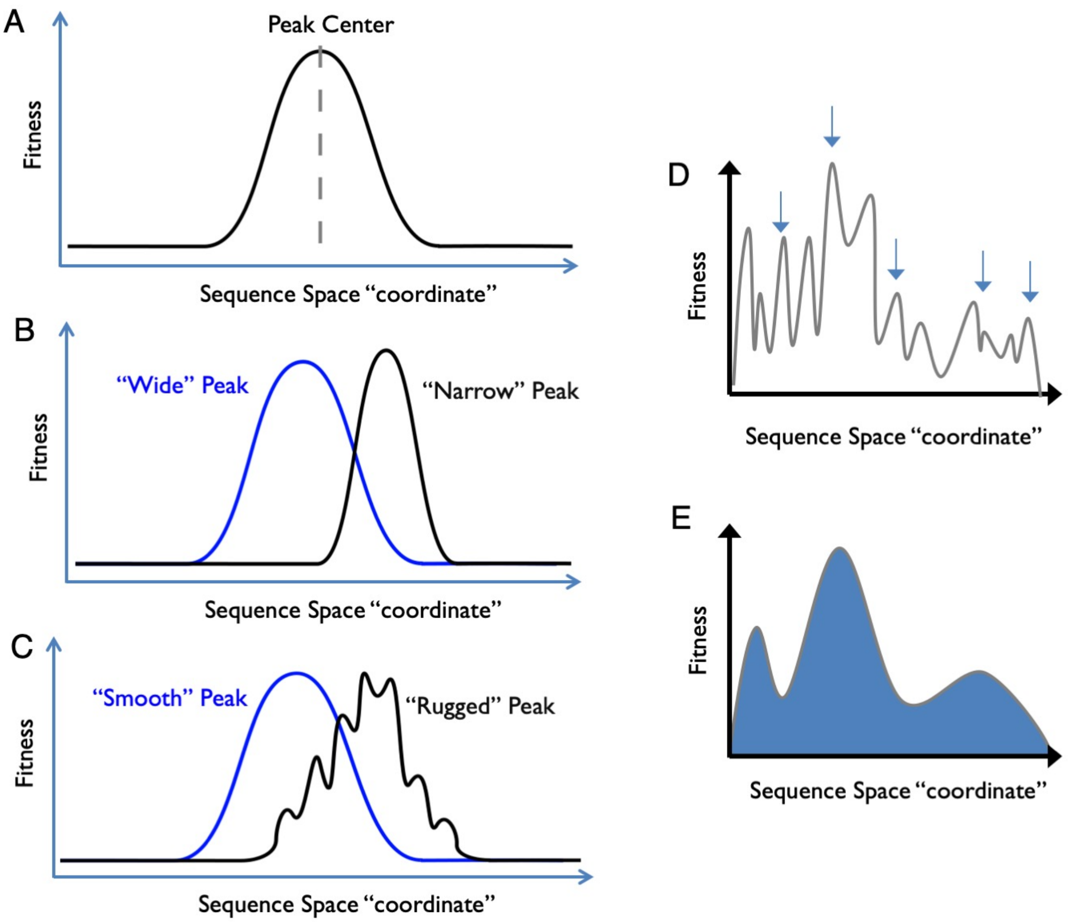

A
Peak Center
Fitness
Sequence Space "coordinate"
B
"Wide" Peak
"Narrow" Peak
Fitness
Sequence Space "coordinate"
C
"Smooth" Peak
"Rugged" Peak
Fitness
Sequence Space "coordinate"
D
Fitness
Sequence Space "coordinate"
E
Fitness
Sequence Space "coordinate"

**Figure 1.2. General schematic of fitness landscape shapes and properties**
Cartoon demonstrating the basic fitness landscape features discussed in this section. **(A)** A fitness "peak" consists of sequences close together along a theoretical sequence space pseudo-coordinate, usually defined as edit distance between polymer sequences. The peak center or local maximum is the sequence of highest fitness. **(B)** The "width" of a peak generally refers to how quickly fitness drops away with distance from the maximum. **(C)** "Ruggedness" refers to how rough the surface of a peak is; a rugged peak will show many smaller increases and decreases in fitness while moving along a straightforward trajectory. **(D)** In the approach described in Chapter 3, it is impossible to fully sample a large fitness landscape; instead, we can sample many points from a wide range of unrelated peaks, in order to get a general overview of the landscape's properties. **(E)** In the approach described in Chapter 4, we can generate a full fitness landscape by measuring the kinetics of all possible sequences (or at least all sequences with enhanced activity), but are limited to acting over smaller fitness landscapes.

## 1.4. High-throughput measurements of evolutionary space

To a scientist outside the field of artificial selection, the idea of measuring the kinetics of tens of millions of catalysts simultaneously might seem staggering, and yet in the study of fitness landscapes it remains an unfortunately-low restriction. To wit, the combinatorial nature of nucleic acids allows a large sequence space to be defined by a small number of nucleotides. Typical nucleic acid selection procedures allow a starting pool of $10^{14}$-$10^{15}$ randomized sequences; this provides 100x coverage for a 20-mer random nucleotide region, but less than one hundredth of a percent of sequence space for a 30-mer random region. Here, functional nucleic acids present certain advantages compared to more complicated protein selections, as an alphabet of only four nucleotides allows far higher coverage of random sequence libraries for *de novo* functional ribozymes and aptamers. But even the highest-throughput sequencing (HTS) methods, when used at the endpoint of a selection, can observe only on the order of $10^{7}$-$10^{9}$ sequences. While the depth of sequencing now available is far higher than even five years ago, it still constrains measurement to only the smallest or sparsest-sampled sequence spaces.

Thus, 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. measurement throughput. Predominantly *in silico* approaches have shown some usefulness here, such as in the effective generation of an anti-HIV aptamer landscape using only a small number of tested molecules,[51] but such study requires the assumption of an extremely simple relationship between fitness and sequence. Instead, the focus of most fitness landscape research has been on small, manageable landscapes.

Historically, the most direct measurements of functional biochemical landscapes come from microarray systems, which tile thousands of copies of each of many sequences on a slide for optical measurement of reagent production or substrate binding. Approximately $10^{5}$-$10^{6}$ sequences can be studied in reasonable copy number in a single microarray study, equivalent to full coverage of DNA/RNA sequence space with *N*=10. Nucleic acid microarrays have been used to investigate double and triple-mutational scans of aptamers,[52] used with rational truncation to investigate the importance of structural constraints on aptamer activity,[53] and combined with *in silico* approaches to interrogate putative high-activity regions of larger evolutionary spaces in array-based directed evolution.[24] 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.[54]

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).[42] 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.

It should be noted that sparse sampling of fitness landscapes limits the applicability of epistasis and ruggedness analysis. 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 but fully random sampling can also lead to patterns of inaccuracy in epistasis and ruggedness [56], and the prevalence of indirect evolutionary pathways that bypass local valleys [57] could lead to underestimates of evolvability if the explored space is too small.

To improve landscape coverage and interrogate larger sequence spaces, it is possible to overcome the limit of with *in vitro* selection – if selection can isolate nearly all of the high-activity sequences, complete mapping of an RNA fitness landscape becomes possible for short sequences.[43] As described in Section 4.1, such an approach has been limited in its ability to assign useful and accurate fitness values, a limitation which this work seeks to overcome.

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, a specific kinetic parameter, 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,[43,58] likely due to experimental noise and biases related to sequencing (e.g., PCR). When this research project was started, no alternative methods for using HTS to measure activity existed; in the past several years, parallel development to that described in sections 3 and 4 has led to new approaches that use HTS to perform direct activity screens, functioning as a "virtual array" (but limited to measuring far smaller evolutionary spaces).[59,60] The newest methods in optimizing biomolecules, including those described in this work, seem to be those that thus turn evolutionary selections directly into screening assays that directly measure chemical function.

While high-throughput techniques for measuring larger fitness distributions have occurred mostly in the study of functional nucleic acids, similar techniques are beginning to emerge in the study of protein fitness landscapes, suggesting that nucleic acids may continue to function as a useful model system for the study of more complicated selections. The study of protein fitness landscapes, which tend to focus on mutational analysis of existing proteins rather than *de novo* functions, has also transitioned to HTS. In a technique known as deep mutational scanning (DMS), the activity of a mutant library is linked to organismal (cell or virus) fitness by the survival of different genetic variants, similar to HTS analysis of different sequence's survival in *in vitro* nucleotide selections.[61-63] 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.[64] While such protein methods have generally been limited to screening basic variants of a wild-type protein, they should be equally viable for studying the endpoint of functional-protein selections, allowing the methods this work develops from studying ribozyme selections to be equally applicable in analysis of enzyme or even antibody evolution.

## 1.5. The origin of life as a model system for engineering evolution

Understanding how the earliest life arose is a fundamental problem of biology, and one fundamental to our understanding of what alternate forms life could take, either on another planet or through progress in bioengineering. Much of the progress that has been made in this area astrobiology has been due to a synthetic and mechanistic perspective on the origin of life. In short, while we may never be able to tell exactly how life began on Earth, engineering prebiotic analogs can tell us what pathways to the earliest organisms are possible (and which are less or more likely), as well as how such pathways might evolve.[38]

The earliest life on Earth likely consisted mostly or entirely of functional ribozymes, reacting and copying within a primordial soup,[65-67] and these prebiotic life-analogs show great potential as a model system for studying engineered evolution. Specifically, such ribozymes are small, with catalytic functions often carried out by only a few nucleotides,[40] leading to tractable sequence spaces that can still contain a diversity of active mechanisms and genetic motifs.[27,68] And unlike other toy biological systems, prebiotic ribozymes allow both the study of fundamental evolutionary steps *and* the mechanisms necessary to encode orthogonal information storage[69] or translation machinery[26,70-72] within engineered cells. Basic functional RNA studies have also provided significant insight into what larger fitness landscapes might look like, as extensive studies of the fitness landscapes of various functional RNAs have shown mostly isolated, sharp peaks with distinct structural motifs linked by generally unfavorable paths of mutation.[25,43,50]

As a particular focus of study, we choose aminoacylation ribozymes, which catalyze a simple[40,73] and important[74-76] function whose evolvability has been relatively unstudied.[77] To build proteins, the ribosomes of all life on earth require primed tRNAs, which are loaded with an activated amino acid by an aminoacylase. Before such proteins could evolve, however, early life would likely have required an RNA-based system for aminoacylating its tRNA analogs, making aminoacylation ribozymes a key step in the origin of life. It remains an open question whether ribozymes specific for different amino acids would evolve independently versus from a single promiscuous ancestor. The observed similarity among modern tRNAs has been suggested to be the result of diversification from a single ancestral ribozyme evolving specificity to a number of different amino acids.[78-80] Unfortunately, existing aminoacylation ribozymes have been selected primarily with AA-AMP precursors,[7,81] which are unlikely to have been present in an early earth environment as they break down quickly in the presence of $CO_2$;[82] prior to this work, no research had successfully identified a prebiotically-plausible aminoacylase ribozyme, leaving a major missing link in the evolution of the genetic code. Oxazolone-amino acids (Figure 1.1E), which can be prebiotically produced, show more reactivity than thioesters along with more stability than AMP-amino acids, and have been proposed as a realistic prebiotic source of activated amino acids.[83] Thus, oxazolone-based aminoacylation ribozymes were chosen as an ideal model system for studying the generation of full fitness landscapes as part of a *de novo* functional selection. The results of such landscape analysis of an oxazolone-based aminoacylation ribozyme system are presented and discussed in Chapter 4.

The other ribozyme system presented and described in this work is that of RNA capable of TMP-based self-triphosphorylation (further discussed in Chapter 3). This selection was previously carried out in Ulrich Müller's lab at UCSD, and thus does not represent new experimental work; however, it presented a useful data set for studying round-to-round enrichment of an evolving aptamer. TMP is a biological relevant molecule potentially produced through geological processes, while triphosphorylation of dephosphorylated nucleotide ends would be a potentially useful function in RNA world scenarios. Previously, the selection had identified and tested a number of individual ribozyme sequences through bacterial cloning and sequencing, but it was decided that HTS analysis might provide a better understanding of the range of kinetics achievable by such a ribozyme. It was also hoped that

HTS would do better at analyzing the shape of fitness space and the relative effects that small mutations might have on such a ribozyme's evolution, including whether or not it would be more likely to evolve through neutral mutation paths or individual beneficial mutations.

## 2. Novel theoretical approaches to studying fitness distributions

The work presented in this chapter primarily corresponds to Pressman and Chen, 2019, an in-progress publication for which analysis has been completed.

### 2.1. Background: Mathematical models of simple evolutionary selection

As discussed in section 1.3, fitness distributions—whether as the DFE of a population adapting to a new environment, the initial fitness distribution of a de novo catalytic function, or anything in between—are one of the most important parameters needed to optimize evolution, but are also extremely difficult to measure. Short of actually measuring the fitness of a million different organisms or sequences, there exists in the literature no way but guesswork to actually measure how fitness effects are distributed over a random molecular space. In this chapter, we describe a theoretical approach that can estimate fitness distributions without requiring more than a handful of fitness measurements (on the order of ten, rather than tens of thousands), as well as related novel mathematical approaches that may be directly useful in the optimization or automation of evolution.

Studying directed evolution at or near the molecular level means generally limiting ourselves to simple, single-celled organisms or entirely cell-free replication systems, which has the side benefit of greatly simplifying the necessary genetic models. In the simplest of these cases, we assume that mutation is not a constant process, but one introduced at a single point—either with a mutagenized wild-type population, a targeted molecule with a randomized region, or an entirely random molecule pool used to search for *de novo* function. We also assume our system reproduces asexually, which further lets us treat its genome as effectively haploid.

Overall, each unique genotype in our starting population has a fixed proportional reproductive fitness, which we call *x.* The system has a fixed carrying capacity; the actual value of this is arbitrary, but we assume that the system is only allowed to replicate up to a certain consistent maximum population. We can assign baseline fitness $x = 1$ to a "wild-type" organism, as under fixed capacity only species' relative fitness matters. (This simplifies the math, assigning a value of $x = 0$ to a species incapable of reproducing and $x = 2$ to one that reproduces twice as fast as the wild-type, a simplified concept of fitness that can be traced back all the way to the initial mathematical descriptions of evolution.[84] A sequence's selection coefficient *s*, as frequently used in population genetics, is thus $s = x – 1$).

In the initial state (time $t$=0, before selective pressure is applied), the fitness distribution of the population follows some function $p_0$(x). At time *t*, this function evolves as $p_t(x)$, such that

$$\frac{\partial p_t(x)}{\partial t} = p_t(x)\big(x - \mu(t)\big) \tag{2.1}$$

where $\mu(t) = \int_0^\infty x p_t(x) dx$ is the mean fitness x at time t. This evaluates as

$$p_t(x) = \frac{e^{xt} p_0(x)}{\int_0^\infty e^{xt} p_0(x) dx} \tag{2.2}$$

One useful tool here for studying evolving populations is Fisher's Theorem (FT), originally proposed in 1930 by its author as the "Fundamental theorem of natural selection"[85] and promptly ignored for its limited applicability to only simple asexual, haploid systems. Fisher's Theorem states that the change in average fitness of a population should equal the variance of sequence fitness in the population (in the case of reproduction falling into discrete rounds, this is normalized by mean fitness). In other words,

$$\frac{\partial \mu(t)}{\partial t} = \sigma^2(t) = \int_0^\infty p_t(x)[x - \mu(t)]^2 dx \tag{2.3}$$

where $\sigma^2(t)$ is the variance of x at time t.

It has been previously noted[86] that in a case with continuous reproduction (that is, higher fitness genotypes reproduce faster, live longer, or otherwise have correspondingly more reproductive cycles), FT can be generalized as the *n*th cumulants $\kappa_n(t)$ of the probability distribution at time $t$ $p_t(x)$ such that

$$\frac{\partial \kappa_n(t)}{\partial t} = \kappa_{n+1}(t) \tag{2.4}$$

As $\mu(t) = \kappa_1(t)$, we can equivalently say

$$\frac{\partial^n \mu(t)}{\partial t^n} = \kappa_{n+1}(t) \tag{2.5}$$

FT describes only the change in allele frequency driven by evolutionary selection, and does not account for genetic drift or other non-selective changes to a population's fitness distribution. Thus, it can be used as an effective test for the extent to which changes in a population are driven by selective pressure as opposed to other factors.[85,87,88]

In general, $\mu(t)$ is a fairly easy quantity to measure for most artificial selections,[1] as it is simply the average activity (either catalytic, binding, or uncapped growth) of an sample taken from the general population. As typically applied, this idea of a generalized Fisher's Theorem (GFT) approach is of limited use. Measuring the change in mean fitness over time can give additional data on how the variance, skewness, etc. of a fitness distribution is changing, but requires a large number of data points to provide a small amount of further interpretation.

## 2.2. Theory: A further-generalized Fisher's Theorem can calculate initial fitness distributions

Historically, a lack of attention paid to Fisher's Theorem and its general case have led to some rather interesting relations being overlooked, which my work seeks to address. If we assume that x is bounded—satisfied, as all species present will have finite positive fitness—the distribution $p_0(x)$ is uniquely defined by the set of its moments from n = 0 to infinity, which in turn can be uniquely defined by the set of its cumulants from n = 0 to infinity. In turn, assuming μ(t) is analytic, each potential trajectory of mean fitness can also be uniquely defined by its derivatives, which correspond to the same set of cumulants. Thus, any bounded initial distribution $p_0(x)$ uniquely corresponds to μ(t), its population mean fitness as a function of time. (Fitting μ(t) to predictions from a hypothetical $p_0(x)$ has some slight basis in the literature,[89] though only as a way to differentiate between two proposed initial fitness distributions and not as a way to specifically fit a starting distribution).

Here, though, things get even more interesting. For a given Probability Distribution (PDF) *f*(x) whose domain is $x \geq 0$, its cumulant-generating function (CGF) (if it exists) can be calculated as $\ln \mathcal{L}\{f\}(-s)$, where $\mathcal{L}\{f\}(-s)$ is the Laplace transform of *f*(*x*) with transform variable -*s*. Thus, a PDF of population fitness $p_t(x)$ has corresponding CGF $K_t(s)$ following

$$\mathrm{K_t(s)} = \ln \mathcal{L}\{\mathrm{p_t(x)}\}(-\mathrm{s}) \tag{2.6}$$

and so

$$\frac{\partial \mathrm{K_t(s)}}{\partial \mathrm{s}} = \frac{\mathcal{L}'\{\mathrm{p_t(x)}\}(-\mathrm{s})}{\mathcal{L}\{\mathrm{p_t(x)}\}(-\mathrm{s})} \tag{2.7}$$

One property of a CGF is that its derivatives, evaluated at zero, equal the cumulants of the original PDF. Thus, at time t, we have $p_t(x)$ with the corresponding $K_t(s)$ whose derivative at s = 0 is μ(t), or

$$\mu(t) = \frac{\partial K_t(s)}{\partial s}\bigg|_{s=0} = \frac{\mathcal{L}'\{p_t(x)\}(-s)}{\mathcal{L}\{p_t(x)\}(-s)}\bigg|_{s=0} \tag{2.8}$$

But we note that

$$\frac{\mathcal{L}'\{p_t(x)\}(-s)}{\mathcal{L}\{p_t(x)\}(-s)}\bigg|_{s=0} = \frac{\mathcal{L}'\left\{\frac{e^{xt}p_0(x)}{\int_0^\infty e^{xt}p_0(x)dx}\right\}(-s)}{\mathcal{L}\left\{\frac{e^{xt}p_0(x)}{\int_0^\infty e^{xt}p_0(x)dx}\right\}(-s)}\Bigg|_{s=0} \tag{2.9}$$

and by properties of the Laplace transform, $\mathcal{L}\{e^{xt}p_0(x)\}(-s) = \mathcal{L}\{p_0(x)\}(-s-t)$; thus, equation (2.9) is equivalent to

$$\frac{\mathcal{L}'\{p_0(x)\}(-s-t)}{\mathcal{L}\{p_0(x)\}(-s-t)} \bullet \frac{\int_0^\infty e^{xt}p_0(x)dx}{\int_0^\infty e^{xt}p_0(x)dx}\bigg|_{s=0} = \frac{\partial K_0(s+t)}{\partial(s+t)}\bigg|_{s=0} = \frac{\partial K_0(t)}{\partial(t)} = \frac{\partial K_0(s)}{\partial s}\bigg|_{s=t} \tag{2.10}$$

That is, mean fitness is equal to the derivative of the initial-time CGF evaluated at s = t.

By equations (2.5) and (2.10) above, we can then define a further-generalized FT as follows:

$$\kappa_n(t) = \frac{\partial^{n-1}\mu(t)}{\partial t^{n-1}} = \frac{\partial^n K_0(s)}{\partial s^n}\bigg|_{s=t} = \frac{\partial^n[\ln \mathcal{L}\{p_0(x)\}(-s)]}{\partial s^n}\bigg|_{s=t} \tag{2.11}$$

as well as the more obviously-useful relation

$$p_0(x) = \mathcal{L}^{-1}\left\{\exp\int \mu(t)dt\right\}(-t) \tag{2.12}$$

which allows us to calculate the starting distribution of evolutionary fitness (or the distribution at any point during a course of constant selective pressure) via the inverse Fourier transform of the exponent of mean fitness.

This is fairly significant, as we now have a way to convert a relatively easy-to-measure quantity (the evolution of mean fitness over the course of a selection) into an extremely difficult-to-measure parameter (the distribution of fitness at the beginning of a selection). In practice, the Inverse Laplace transforms are often difficult to calculate, proving extremely complicated for all but the simplest functions. Thus, it is not necessarily ideal to calculate $p_0(x)$ for an unknown evolutionary case simply by inverse Laplace transformation of the exponent of integrated mean fitness. Instead, a more practical approach is to compare the transformed quantity $\mathcal{L}\{p_0(x)\}(-s)$ to $\exp\int \mu(s)ds$ for a variety of proposed functional approximations of $p_0$(x). Laplace transformations are unique, and exist for a variety of common probability distributions, as well as for a stepwise-defined function with an arbitrary number of histogram bins.

Towards that end, however, $\mathcal{L}\{p_0(x)\}(-s)$ tends to be a rapidly-growing exponential or polynomial function for most common probability distributions; the extremely high resulting slopes of these curves make it hard to fit $\mathcal{L}\{p_0(x)\}(-s)$ to $\exp\int \mu(s)ds$. Instead, we consider log-derivative space. The mean-fitness growth curves $\mu(t)$ for a variety of initial distributions tend to be fairly distinct in their shape (Figure 2.2), making it easiest to fit $\mu(t)$ to $\left.\frac{\partial[\ln \mathcal{L}\{p_0(x)\}(-s)]}{\partial s}\right|_{s=t}$ for various proposed $p_0(x)$ functions.

### 2.3. Theory: Generalized Fisher's Theorem under discrete growth

The case of Fisher's theorem described so far in this work describes continuous growth, where organisms duplicate more or less continuously and randomly without specific generational pauses. In the case of growth that follows a generational cycle with a specific length, this is may not be a valid approximation. The continuous growth assumption involves higher-fitness organisms replicating faster; if instead all organisms replicate at the same rate, with fitness describing number or survival rate of offspring, then doubling x should produce a population increase of $2^t$ over t wild-type generations, while under the continuous assumption it produces a population increase of $e^{2t}$. In the case of selective pressure over such separable generations (such as is the case in many in vitro selections), a discrete-time approach is needed.

In the discrete-time case, we note that

$$p_R(x) = \frac{x^R p_0(x)}{\int_0^\infty x^R p_0(x)dx} \tag{2.13}$$

where n is the number of generational/selection "rounds" that have elapsed. In the discrete case, we consider the central moments $E[x^n]_R$ — that is, the nth central moment after R rounds of selection, and additionally consider that

$$E[x^n]_{R+1} = \frac{\int_0^\infty x^{R+n+1} p_0(x)dx}{\int_0^\infty x^{R+1} p_0(x)dx} = \frac{\int x^{R+n+1} p_0(x)dx}{\int_0^\infty x^R p_0(x)dx} \frac{\int_0^\infty x^R p_0(x)dx}{\int_0^\infty x^{R+1} p_0(x)dx} = \frac{E[x^{n+1}]_R}{\mu(R)} \tag{2.14}$$

so thus

$$E[x^n]_0 = \mu(0)[x^{n-1}]_1 = \mu(0)\mu(1)[x^{n-2}]_2 = \cdots = \prod_{k=0}^{n} \mu(k) \tag{2.15}$$

becomes our discrete-case analog to GFT, and the central moments (and standardized moments and cumulants) of $p_0(x)$ can be uniquely defined as products of $\mu(R)$ across multiple generations.

Unlike in the continuous case, there is no easy mathematical transformation to turn this $\mu(R)$ into $p_0(x)$, as the relevant properties of Laplace transforms no longer apply.

However, for a given proposed PDF for $p_0(x)$, we can still calculate a $\mu(R)$ trajectory and thus fit a variety of potential initial distributions to the observed increase in mean fitness. Instead of a Laplace transform, the equation for $\mu(R)$ resembles a Mellin transform, as follows:

$$\mu_R(x) = \frac{\int_0^\infty x^{R+1} p_0(x) dx}{\int_0^\infty x^R p_0(x) dx} \tag{2.16}$$

which we term “Discrete Method 1.”

But an easy workaround also exists to convert some cases from the discrete growth model here to the continuous one. Define $x' = e^{x-1}$; then

$$p_R(x') = \frac{e^{x'R} e^{-R} p_0(x')}{\int_0^\infty e^{x'R} e^{-R} p_0(x') dx} = \frac{e^{x'R} p_0(x')}{\int_0^\infty e^{x'R} p_0(x') dx} \tag{2.17}$$

and

$$p_0(x') = \mathcal{L}^{-1}\left\{\exp \int \langle x' \rangle(t) dt\right\}(-t) \tag{2.18}$$

From here, the math appears similar to the continuous case, with the caveat that we cannot directly measure $\langle x' \rangle(t)$ . However, if the domain of $p_0$ is relatively small, e.g. $0.5 < x < 1.5$ (or most species have fitness close to a wild-type of 1), we can assume that the first-order approximation $e^x \approx 1 + x$ is sufficiently valid that $\langle x' \rangle(t) \approx \langle x \rangle(t)$. Then we expect $\langle x \rangle$ to increase following the same curve as $\langle x' \rangle$, and

$$p_0(x) \approx \mathcal{L}^{-1}\{\exp \int \mu(R) dR\}(-R) \tag{2.19}$$

which we term “Discrete Method 2.”

### 2.4. Theory: Selection under changing environmental conditions

While we focus our GFT examples and analysis on cases of fixed selection/growth conditions, the same methods can be applied more complexly to a scenario in which conditions vary. The simplest example of this is that of a single change in selective pressure—how will scaling the effects of a selection condition up or down affect the distribution of fitness effects at a given time point? Here, we can assume two selection conditions, with respective (and related) fitness $x_1$ and $x_2$, such that $x_2 = f(x_1)$. Then, a starting distribution of fitness can be adjusted by variable transformation as follows:

$$p_{0,x_2}(x_2) = \frac{df^{-1}(x_2)}{dx_2} e^{f^{-1}(x_2)t_1} p_{0,x_1}\big(f^{-1}(x_2)\big) \tag{2.20}$$

As a slightly more complicated case, we can consider a selection in sequential phases, where the a biological function is selected for at one selective pressure and then in later generations at a different pressure. In the initial phase, from $t = t_0$ to $t_1$, we assign each species a fitness $x_1$; in the second phase, from $t = t_1$ to $t_2$, we assign fitness $x_2$, such that $x_2 = f(x_1)$. At $t < t_1$, we observe

$$p_{t,x1}(x_1) = \frac{e^{x_1 t} p_{0,x1}(x_1)}{\int_0^\infty e^{x_1 t} p_{0,x1}(x_1) dx_1} \tag{2.21}$$

at $t = t_1$, we see

$$p_{t1,x1}(x_1) = \frac{e^{x_1 t_1} p_{0,x1}(x_1)}{\int_0^\infty e^{x_1 t_1} p_0, x1(x_1) dx_1} \tag{2.22}$$

and by variable transformation,

$$p_{t1,x2}(x_2) = \frac{\frac{df^{-1}(x_2)}{dx_2} e^{f^{-1}(x_2)t_1} p_{0,x1}\big(f^{-1}(x_2)\big)}{\int_0^\infty \frac{df^{-1}(x_2)}{dx_2} e^{f^{-1}(x_2)t_1} p_{0,x1}\big(f^{-1}(x_2)\big) dx_2} \tag{2.23}$$

Then, at $t > t_1$, we have

$$p_{t,x2}(x_2) = \frac{e^{x_2(t-t_1)} p_{t1,x2}(x_2)}{\int_0^\infty e^{x_2(t-t_1)} p_{1,x2}(x_2) dx_2} \tag{2.24}$$

The same premise can be applied when investigating a selection performed multiple times under varying conditions, or when choosing a selective pressure for a selection with a known variation of biological parameters. Assume $p(x)$ represents the starting distribution of a biochemical parameter, such as the activity of a catalyst or the binding coefficient of a receptor or affinity molecule. Let $x = f^{-1}(y)$ be the fitness resulting from a gene with parameter value $y$ at a chosen concentration of substrate or ligand. Then $p_{0,x}(x) = \frac{df(x)}{dx} p_{0,y}(f(x))$, and

$$p(x) = \frac{\frac{df(x)}{dx} e^{xt} p_{0,y}\big(f(x)\big)}{\int_0^\infty \frac{df(x)}{dx} e^{xt} p_{0,y}\big(f(y)\big) dx} \tag{2.25}$$

While elaborate, this is tractable in simple cases. For instance, in the *in vitro* selection of an active first-order catalyst (or any other selection where fitness is tied primarily to the result of a simple catalytic activity), we can assume fitness $x$ follows $x = 1 - \exp(-K\, y)$, where $K$ equals substrate concentration times reaction time, and $y$ is catalytic rate. Then $y = -(\log(1\text{-}x))/K = \mathrm{f}(x)$, and $\frac{df(x)}{dx} = 1/[K(1-x)]$; thus,

$$p(x) = \frac{e^{xt} p_{0,y}\left(-\frac{\mathrm{Log}\,(1-\mathrm{x})}{\mathrm{K}}\right) / K(1-x)}{\int_0^\infty e^{xt} p_{0,y}\left(-\frac{\mathrm{Log}\,(1-\mathrm{x})}{\mathrm{K}}\right) / K(1-x) dx} \tag{2.26}$$

The case of simple aptamer or antibody selection, using Langmuir-model binding, follows x = [S]/(y + [S]), where [S] is substrate concentration and y is an individual gene's dissociation constant. Then $y = [S]\,(1-x)/x$, and $f'(x) = -1/x^2$, giving

$$p(x) = \frac{e^{xt} p_{0,y}\left(\frac{[\mathrm{S}](1-\mathrm{x})}{\mathrm{x}}\right) / x^2}{\int_0^\infty e^{xt} p_{0,y}\left(\frac{[\mathrm{S}](1-\mathrm{x})}{\mathrm{x}}\right) / x^2 dx} \tag{2.27}$$

Such considerations make it theoretically possible to optimize control of an automated or semi-automated genetic engineering system. Unfortunately, we are limited in our ability to treat evolution as analogous to a normal circuit component; while the evolution of mean fitness or the population of a specific species are time-invariant functions, neither behave in a linear fashion. Finding a way to present such optimization simply presents an intriguing further question, but one beyond the scope of this work.

## 2.5. Theory: Deconvoluting fitness distributions and measurement noise

As explained in chapters 3 and 4, the effect of stochastic noise in measuring activity of individual sequences can be significant, and thus such noise could theoretically present a challenge to accurately measuring underlying fitness distributions. Luckily, it is also a phenomenon that can easily be accounted for mathematically, as follows:

Assume that every species in a fitness distribution $p_0(x)$ can have its fitness approximately measured, but with random noise applied to the measurement. We further assume this noise of measurement is predictable and consistent across all individual genomes, appearing as a random variable with distribution *noise*($x$). Then, in attempting to measure $p_0(x)$, we can only actually observe a blurry version of the real distribution, with *blurring* following the convolution $p_{0,measured}(x) = (p_0 * noise)(x) = \int_0^z p_0(z)noise(x - z)dz$. In GFT analysis, it becomes surprisingly easy to remove this kind of noise effect through an interesting property of the Laplace transform: specifically, we note that

$$\mathcal{L}\{(a * b)(x)\}(-s) = \mathcal{L}\{a(x)\}(-s)\mathcal{L}\{b(x)\}(-s) \tag{2.28}$$

Thus, if our initial distribution is mistakenly measured as $(\mathrm{p_0} * \mathrm{noise})(\mathrm{x})$, this impacts our analysis as follows:

$$\kappa_n(t) = \frac{\partial^n \left[\ln \mathcal{L}\left\{p_{0,measured}(x)\right\}(-s)\right]}{\partial s^{\mathrm{n}}}\Bigg|_{s=t} = \frac{\partial^n \left[\ln\left(\mathcal{L}\{p_0(x)\}(-s)\mathcal{L}\{noise(x)\}(-s)\right)\right]}{\partial s^n}\Bigg|_{s=t}$$

$$= \frac{\partial^n [\ln \mathcal{L}\{p_0(x)\}(-s)]}{\partial s^n}\Bigg|_{s=t} + \frac{\partial^n [\ln \mathcal{L}\{noise(x)\}(-s)]}{\partial s^n}\Bigg|_{s=t} \tag{2.29}$$

That is, each predicted cumulant (mean, variance, etc.) of fitness will be off by a specific amount that depends only on the shape and magnitude of noise with which we measured $p_0$. If the approximate shape of this noise is known, it can be completely and easily subtracted out from our predictions of how the fitness distribution changes as the population evolves.

The simplest case occurs when we expect noise to follow a normal distribution: if *noise*($x$) is normally distributed with mean $m$ and variance $\sigma^2$, then

$$\mathcal{L}\{noise(x)\}(-s) = \exp\left(t\, m + \frac{1}{2}\sigma^2 t^2\right) \tag{2.30}$$

$$\kappa_{n,noise}(t) = \frac{\partial^n \left[t\mu + \frac{1}{2}\sigma^2 t^2\right]}{\partial s^n}\Bigg|_{s=t} \tag{2.31}$$

Thus, mean fitness $\mu(t)$ would be off from our prediction by a linear $m + \sigma^2 t$, our variance in fitness would be off by a constant error of $\sigma^2$, and all higher-order cumulants of the distribution would behave as expected.

Of course, if the shape of *noise*($x$) is known, we can also simply perform a deconvolution, such that $p_0$ is equal to $p_{0,measured}$ deconvoluted by *noise*($x$).

### 2.6. Testing generalized Fisher's Theorem approach with simulated data

To test the ability of the generalized Fisher's Theorem approach to fit initial distributions, simulations were conducted on a number of "test" distributions. Test distributions consisted of starting histograms generated from a normal distribution, log-normal distribution, Pareto distribution, and bimodal normal distribution, with the latter chosen specifically as a tricky distribution which we expected would be difficult to fit. Two sets were created of each of these four distributions. In the "low fitness" set, the initial distribution ranged over $0.75 \leq x \leq 1.25$, with mean $\mu(0) = 1$ and $p_0(1.25) \approx 10^{-6}$. In the "high fitness" set, the initial distributions were centered around $x = 1$, but $\mu(0)$ was slightly effected by cropping as distributions ranged over $0 \leq x \leq 6$. Fitness distributions evolved over 20 time intervals, with maximum $t = 100$ and 5 for the high and low fitness sets respectively.

First, we sought to quantify the effect of selection noise on mean fitness trajectories and the resulting initial fitness distributions. For our simulation set, at each time interval, each histogram bin was multiplied by a normally-distributed random variable with a mean of 1 and variable standard deviation; Figure 2.1A,B show the case of a low-fitness normal distribution evolving with 0%, 10%, and 50% enrichment noise per step. A large number of evolving distributions were similarly tested. In general, enrichment noise in the 10-20% range (similar to that expected in later rounds of the selection analyzed in Chapter 4) did not have a noticeable effect on initial distribution fit; enrichment noise in the 50% range (similar to that expected in the selection analyzed in Chapter 3) substantially interfered with initial distribution fit, providing predicted distributions with roughly the correct mean fitness but of incorrect shape and with.

Next, we sought to evaluate methods for fitting the mean fitness trajectories of populations that evolve following discrete replication cycles. The discrete case is expected to better mirror the reality in selections following an artificial replication step, such as many *in vitro* selections; to categorize all artificial selections, we need methods that address both discrete and continuous cases, as explained in section 2.3. From that section, we describe two methods for doing so. Discrete Method 1 simply evaluates a prediction based on the mathematics of the discrete case; this is more computationally-intensive, but still tractable for fits to simple distributions of only a few parameters. This method proved robust, showing little distortion of initial distribution fit over a range of cases (with examples provided in Figure 2.1 C,D). Discrete Method 2 treats the discrete case as identical to the continuous case, an assumption expected to hold for distributions covering only a narrow range of fitness values, which in simulations was demonstrated to be the case (Figure 2.1 E,F).

In order to evaluate the robustness of the GFT approach, rather than fitting initial distributions to the initial distribution, $\mu(t)$ curves were intentionally fit to two different curves, testing the approach's ability to fit distributions of unknown shape: the gamma distribution, a highly-flexible two-parameter distribution, and a stepwise distribution with 10 bins (whose large number of parameters made fitting somewhat difficult). Evolution occurred with 10% of randomness per time step. Gamma distribution fits were used to estimate a rough starting distribution for 10-bin distribution fits, which otherwise had difficulty converging. Overall, these simulations (Figure 2.2) suggest that this approach can fit at least a rough estimate of a variety of different simple distributions. They also show that fitting to an exact equation appears to be more successful when that equation is more similar to the actual initial distribution; and that fitting a multi-bin uniform distribution does a better job reproducing the evolutionarily-important upper tail and mean of a distribution, while potentially struggling to accurately gauge the space in between.

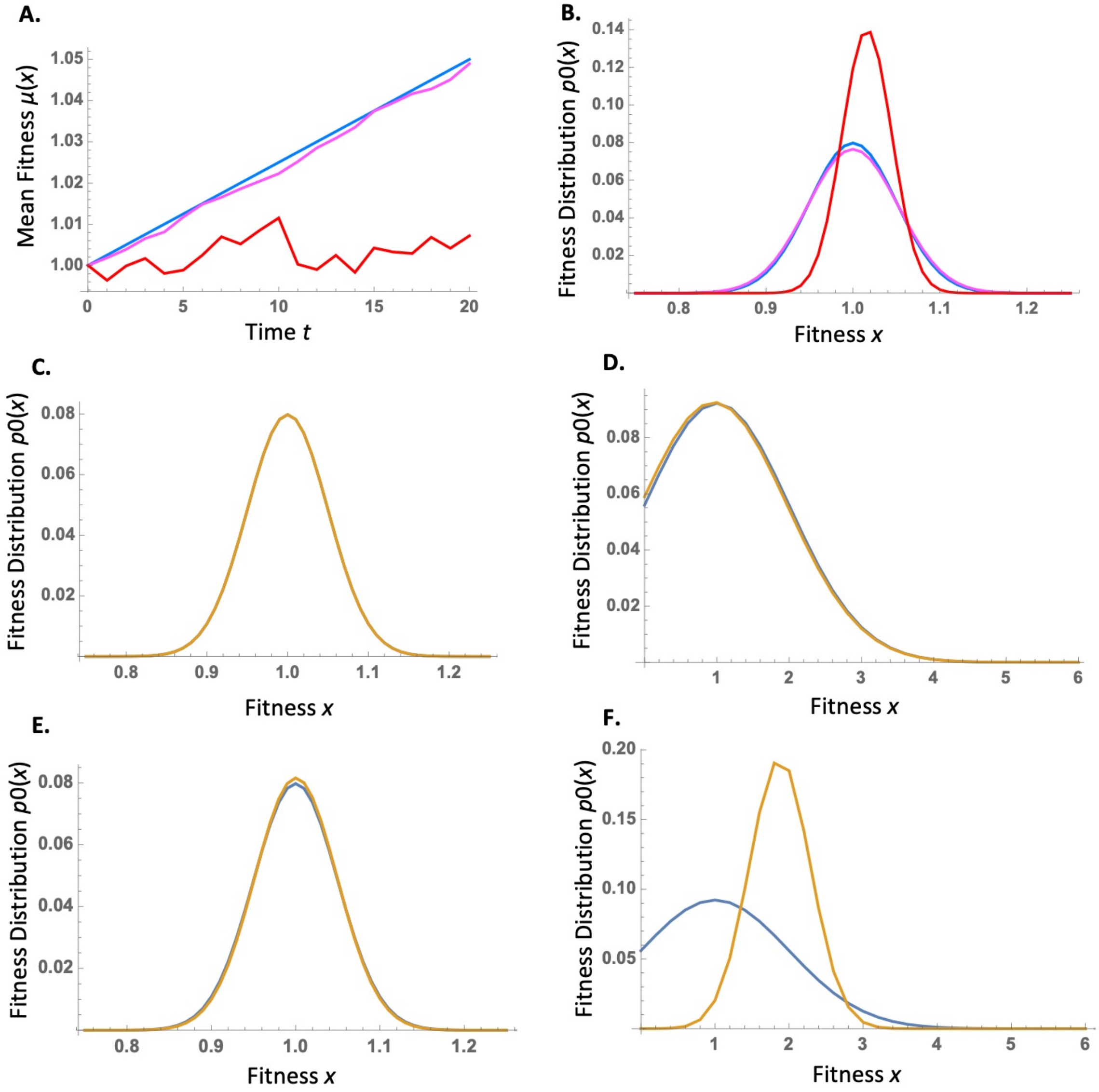


**Figure 2.1. Testing the general robustness of GFT fitness distribution predictions**
**(A-B)** Initial fitness distribution as a normal distribution centered around 1, with standard deviation of 0.5. Distribution was allowed to evolve with 0 error per time step (blue), 10% error (magenta), or 50% error (red). (A) denotes the resulting mean fitness trajectories over the first 20 time steps; (B) denotes an attempt to fit the final trajectories to a normal initial distribution by a GFT approach. **(C-D)** For the case of a fitness distribution evolving following discrete selection rounds (as should be the case with many in vitro selection methods), we test Discrete Method 1 in both low and high-fitness normal initial distributions. GFT-predicted distribution (orange) shows good agreement with real initial distribution. **(E-F)** As C-D, but we test Discrete Method 2 in both low and high-fitness normal initial distributions. GFT-predicted distribution (orange) shows good agreement with real initial distribution for a low-fitness initial distribution, but not for one with a wider fitness range, as predicted.

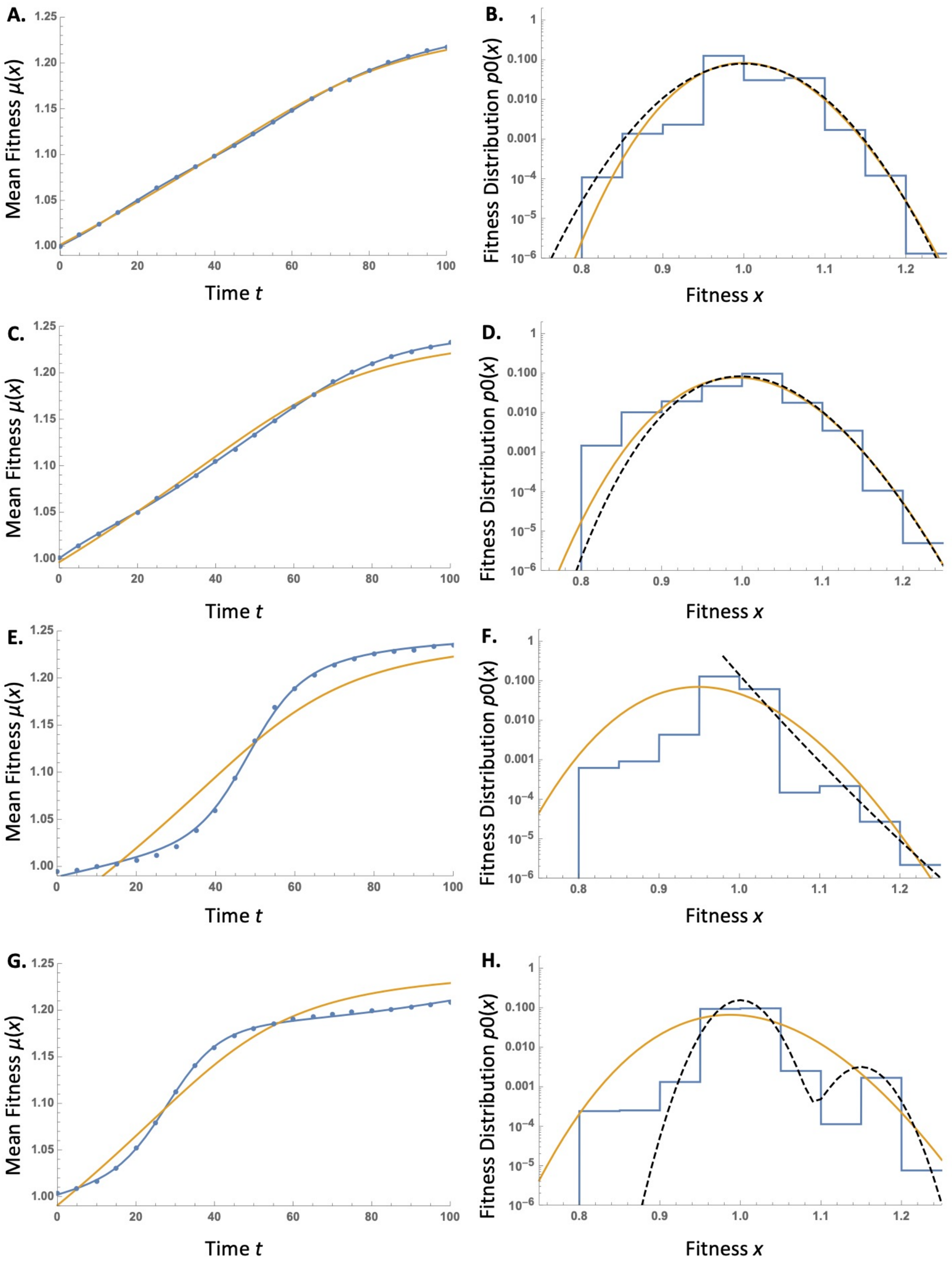


**Figure 2.2. Using GFT approach to build initial fitness distributions of test cases**

The evolution and fitness distributions for narrow-range "low fitness" **(A-B)** normal distribution, **(C-D)** log-normal distribution, **(E-F)** Pareto distribution, **(G-H)** bimodal normal distribution. **Left** column of panels shows the change over time of mean fitness; **right** column of panels shows original and fitted initial fitness distributions. Dots (left) and black dash (right) show original fitness distribution. Orange line is the result of a Gamma distribution fit, carried out with the standard Mathematica nonlinear curve-fitting tool. Black line is the fit to an initial distribution consisting of a 10-bin uniform distribution. Overall, the Gamma distribution has an easier time fitting more similarly-shaped distributions, but captures the general tail of fitness effects well; the 10-bin distribution shows similar difficulty, predicting the evolutionarily-significant mean and upper tail of initial distributions but tends to leave slight holes in the Pareto case.

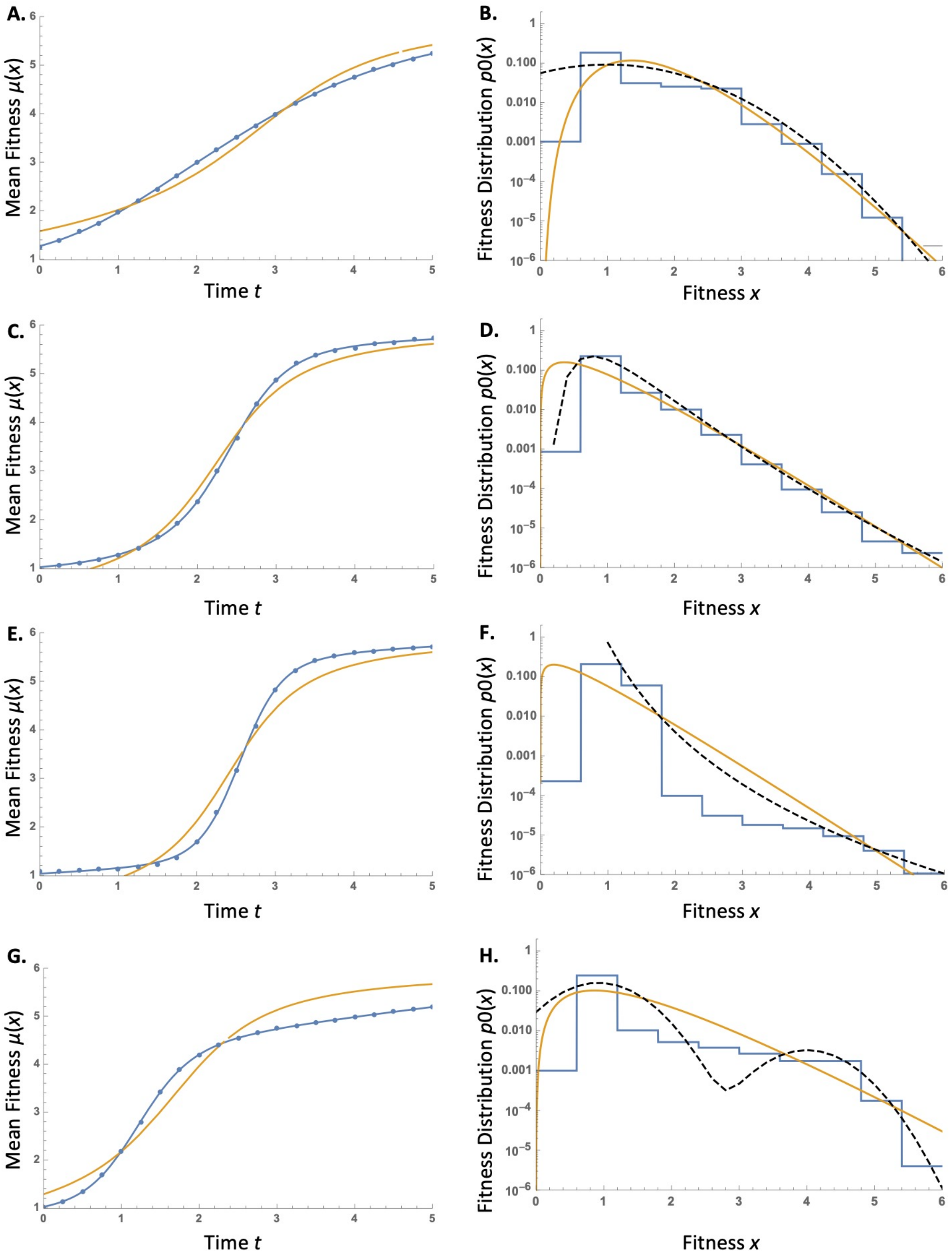


**Figure 2.2 continued. Using GFT approach to build initial fitness distributions of test cases**
The evolution and fitness distributions for wide-range "high fitness" **(A-B)** normal distribution, **(C-D)** log-normal distribution, **(E-F)** Pareto distribution, **(G-H)** bimodal normal distribution. **Left** column of panels shows the change over time of mean fitness; **right** column of panels shows original and fitted initial fitness distributions. Dots (left) and black dash (right) show original fitness distribution. Orange line is the result of a Gamma distribution fit, carried out with the standard Mathematica nonlinear curve-fitting tool. Black line is the fit to an initial distribution consisting of a 10-bin uniform distribution. Overall, fitting behaves similarly to the

low fitness case; once again the 10-bin distribution has difficulty with the middle region of the Pareto case but effectively captures the more-important tail and mean, showing the same issues with the bimodal case.

## 2.7. Generalized Fisher's Theorem approach and real experimental data

Actually testing the GFT approach to estimating real fitness distributions proved somewhat challenging due to a simple lack of existing data; very few experiments have been carried out which both a fitness distribution and mean fitness curve has been measured. As test cases, we use data from the work described later in this report, specificially from the selections analyzed in Chapter 3 and Chapter 4.

For both the TMP triphosphorylase ribozyme and oxazolone aminoacylase ribozyme selections, we chose the highest-abundance sequence as a control; using this sequence's enrichment ratio in various rounds, we were able to calculate very rough estimates of $\mu(R)$. These data are likely to be less accurate than actual chemical quantification of mean pool fitness at every round, as noise in the top sequence's enrichment may add noise to the mean fitness trajectory. But as the goal of this analysis is to estimate fitness distributions from available data we were curious of the extent to which this would still lead to an effective estimate of initial distribution. For comparison, initial fitness distributions under selection conditions were calculated as described in sections 3.6 and 4.3 (Figures 2.3A, 2.4A). Both distributions appeared roughly log-normal, so this distribution was used as an approximation; a Pareto distribution was also fit for the triphosphorylase data, as it appeared to resemble both distributions. We expected the vast majority of sequences in both selections to have no catalytic activity over the base reaction rate of RNA with their substrates. Thus, the actual distribution used for fitting consisted of a log-normal distribution added to a Dirac delta function of variable magnitude centered at $x = 1$.

In the case of the triphosphorylase selection (Figure 2.3), only the first five rounds of data were used, as after this point the selection was expected to be dominated by mutational effects. Here, the GFT-predicted distribution does not fit well with the HTS-estimated distribution of fitness values found in Chapter 3. Specifically, the distribution predicted from mean fitness shows a much shallower drop off in abundance with increasing fitness, especially at the upper end of the fitness distribution. It is notable that the best-fit mean fitness curve (Figure 2.3B) is still increasing at Round 5, despite the data suggesting that it should have leveled off by this point. This lack of better curve fit at high $\mu$, likely resulting from the selection's high enrichment noise propagating into noisy measurement of $\mu(R)$, may be responsible for the inconsistencies seen between the two distributions. We see that the middle range ($10 < x < 100$ or so) agrees fairly well with the HTS-predicted initial fitness distribution, suggesting that we may be able to roughly estimate $p_0(x)$ with some accuracy while $x$ falls in the range in which we have decent $\mu(R)$ values.

In the aminoacylation selection data (Figure 2.4), we have an interesting case. As the substrate concentration during selection was high enough to saturate many high-activity sequences, the distribution of fitness values is expected to be substantially different from the distribution of kinetic activity. (While we may expect the distribution of initial rates to follow a log-normal distribution, the shape of the selection fitness distribution is substantially different). Here, the GFT-estimated distribution does fit well with the HTS-estimated distribution of fitness values found in Chapter 4, for both a Pareto and log-normal fit. While the log-normal distribution appears a better fit overall, the Pareto fit is closer to the HTS-derived distribution at low fitness ($x < 10$ or so), which corresponds to the Pareto distribution better fitting the mean fitness curve at low mean fitness around R = 3 (Figure 2.4B). It is worth noting here that the aminoacylase selection was carried out with significantly less noise and more controlled conditions vs. the triphosphorylase selection, potentially indicating that keeping selection as close to ideal as possible may also improve the accuracy of GFT-estimated of fitness distributions.

All together, these data sets suggest that we may be able to roughly estimate $p_0(x)$ (and by extension $p_t(x)$ at any $t$) with enough accuracy to be insightful, for cases where $x$ falls in the range in which we have decent $\mu(R)$ values. The key limitations of a GFT approach to estimating fitness distributions appear to be A) obtaining an accurate time-course of mean fitness, as the population evolves over a relevant range of fitness values, B) choosing an initial distribution model fairly similar to the real distribution, with a tractable number of parameters, and possibly C) keeping the selection itself carefully controlled and consistent in its parameters. The full usefulness of this method is probably yet to be realized, as an experiment engineered to take these factors into account from the start may result in cleaner data and a better prediction. As the most obvious application of this analysis may be controlling automated gene and phage engineering, future experiments should perhaps be tailored to investigate such systems.

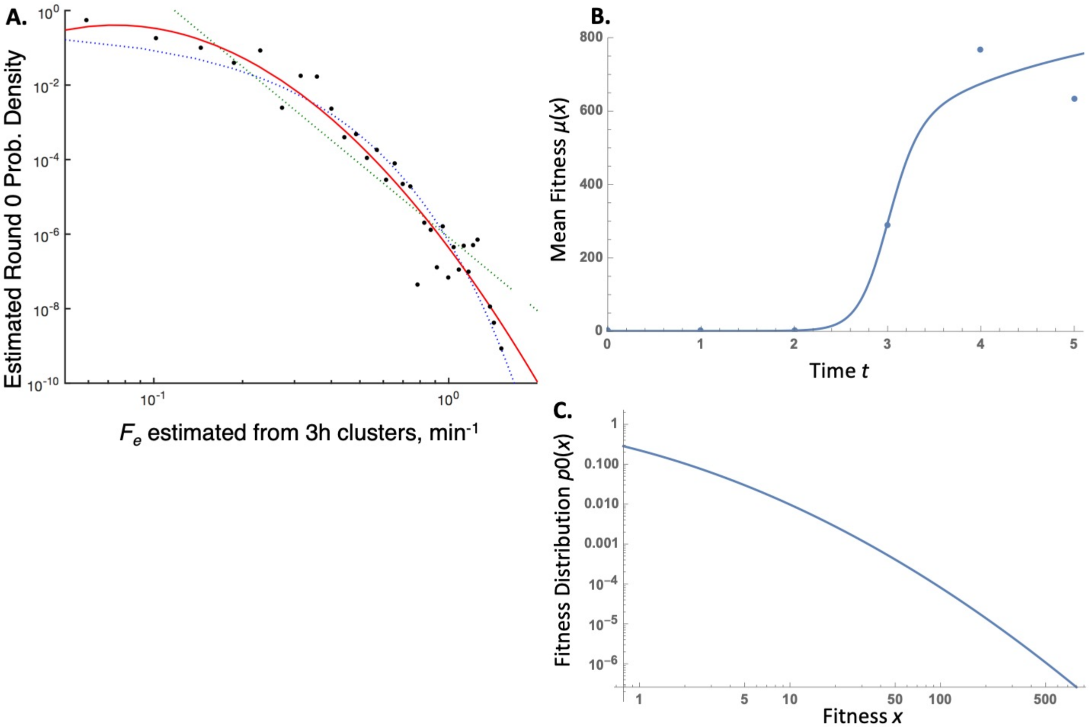


**Figure 2.3. Testing GFT approach with TMP triphosphorylase selection**
**(A)** Initial predicted fitness distribution for a selection for triphosphorylation ribozymes, found through analysis described in Chapter 3. **(B)** The mean fitness for the same population evolving over time. Points indicate mean fitness predicted from enrichment of the top sequence. Blue is a log-normal distribution fit, using Discrete Method 1. **(C)** The fit found in B, displayed as fitness distributions, and scaled to the same expected integrated curve as data shown in A. Assuming the highest $F_e$ sequences to roughly correlate to the highest expected value of x, we see little direct correlation with the curve in A, as the drop-off here is significantly lower.

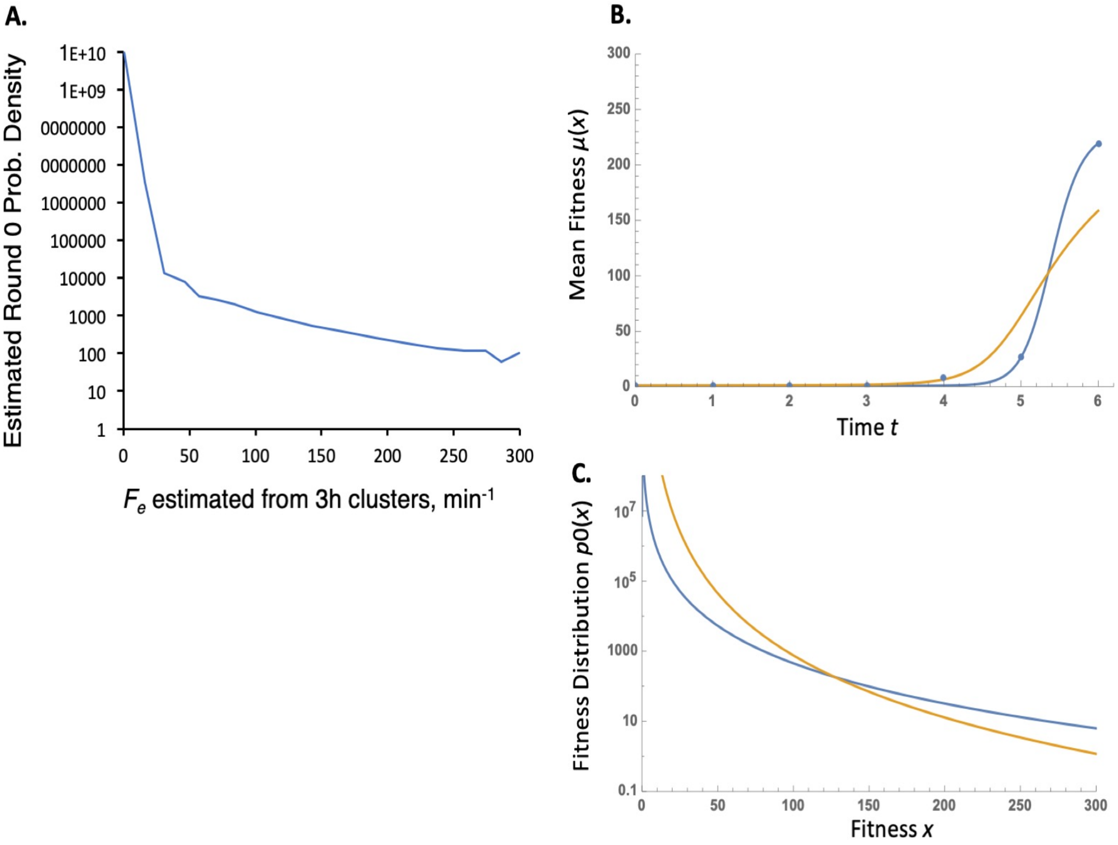


**Figure 2.4. Testing GFT approach with oxazolone aminoacylase selection**
**(A)** Initial predicted fitness distribution for a selection for aminoacylation ribozymes, found through analysis described in Chapter 4. **(B)** The mean fitness for the same population evolving over time. Points indicate mean fitness predicted from enrichment of the top sequence. Blue is a log-normal distribution fit and orange is a Pareto distribution fit, using Discrete Method 1. **(C)** The fit found in B, displayed as fitness distributions, and scaled to the same expected integrated curve as data shown in A. Blue is log-normal distribution, orange is Pareto. Assuming the highest $F_e$ sequences to roughly correlate to the highest expected value of x, we see little significant similarity to the curve in A, with log-normal distribution fitting much better at all but the lowest end, and Pareto distribution fitting better over this region.

# 3. Estimating biocatalyst kinetics from high-throughput selection data

Parts of this section were adapted from Pressman, A., Moretti, J. E., Campbell, G. W., Muller, U.F., Chen, I.A., *Nucleic Acids Res*, 45, Copyright 2017.[58] Reprinted with permission from Oxford University Press.

## 3.1. Background/Theory: Estimating chemical activity from selection enrichment

Traditionally, in vitro selections give little information on their inner workings, with no way to observe selection dynamics and endpoint cloning used to test sequence and function of only a small number of sequences per experiment. Virtually no research has actually compared the theory of directed evolution with actual observation of whole populations as they evolve, despite affordable high-throughput sequencing (HTS) making this a possibility. The work in this chapter focuses on analyzing and understanding HTS-selection data from many rounds of a single selection, using an enrichment-based approach that combines model evolutionary dynamics with analysis of sequences' informational uncertainty as it changes between rounds.

In addition to addressing the variability and non-ideality present in an actual selection, estimated chemical activity of a large number of sequences could also build a picture of an evolving fitness distribution. As discussed in section 1.3, knowing the distribution of catalytic activity over biopolymer sequence space is a necessary condition for optimizing and understanding the limits of artificial selection. [90,91] This may be especially important in nucleic acid selections (i.e. aptamers and ribozymes), because these experiments typically investigate larger and far more randomized sequence spaces. But as we discuss in section 1.4, actual measurements of fitness landscapes for any biochemical function are extremely limited; that same lack of experimental data extends even to observed fitness and activity distributions, which should in principle be simpler to measure. Where these data do exist, they tend to show very different arguments for how aptamer and ribozyme fitness/activity distributions might be shaped. Studies on selections from starting pools with different sequence complexity have suggested a power-law relation between pool size and aptamer affinity or ribozyme activity, though only a small number of measurement points are available.[90,91] In contrast, theoretical considerations have suggested a log-normal distribution of $k_D$ values (and a normal distribution of binding energies) in sequence space for most nucleic acid aptamers,[92,93] as well as a normal distribution of activation energies of RNA melting.[94]

Extracting distributions from *in vitro* evolution was historically hampered by the low throughput of sequencing data, though this concern has been somewhat addressed by increased reliance on HTS to analyze artificial selections. To estimate fitness, HTS analyses typically count sequences present at a selection endpoint, although more recent analyses use the relative enrichment of sequences before and after a final round,[95-97] or follow a specific ribozyme and its variants over several rounds.[98] Recent progress has also developed screening approaches that directly measure sequence activity via sequencing reactions carried out under a range of selection conditions, [59,60] as discussed in Chapter 4—but those approaches, both in our lab and others, had not yet been fully developed when the work described in this chapter was carried out.

As selection abundance can be a poorly predictor chemical activity,[43,58] especially in the case of selection with mutagenesis, we sought to investigate, and hopefully improve on, the ability to predict the activity of a population of ribozymes from multiple rounds of enrichment data. Critically, prior to the beginning of this work, little research existed comparing the real evolution of selected populations to what would be predicted by theory. In the work described in this chapter, we sought to directly investigate the consistency of a large and diverse ribozyme population's evolution. Rather than simply laying out our best guesses for ribozyme activity, we also developed a set of heuristic tools that we hoped could analyze how accurate any such prediction scheme actually is, as well as characterizing the amount of non-ideality "noise" present in a real selection.

In principle, the mathematics involved in estimating chemical activity from selection enrichment should be fairly simple. In practice, some of the terminology can vary in its specific definition, so we define a set of terms here for later clarification. We assume, at round *R* of a selection, a population of $N_R$ total sequences. An individual biopolymer sequence, *i*, occurs as a subpopulation of $n_{R,i}$ observations. Then we define the *abundance* $\alpha_{R,i}$ of sequence *i* at round *R* as

$$\alpha_{R,i} = \frac{n_{R,i}}{N_R} \tag{3.1}$$

Each individual molecule in a population under selection, as described in section 1.2, goes through two steps: a selection step, and a replication step. In the selection step, molecules are either discarded or retained based on their ability to bind a substrate or undergo a specific reaction. We assign each molecule has a probability of $F_i$ to carry out this necessary reaction and survive selection, where $F_i$ is a number between 0 and 1.

From a population of $n_{R,i}$ molecules, the number expected to survive is a binomially-distributed random variable with number of trials $n_{R,i}$ and success probability $F_i$; at sufficiently high *n*, this can be approximated by a normal distribution with mean $n_{R,i}F_i$ and variance $n_{R,i} F_i (1 - F_i)$. If we assume the replication step of each selection round regenerates the selected population to the same total number of sequences, with all molecules replicated at an equal rate, the abundance after one round of selection and replication should follow

$$\alpha_{R+1,i} = \alpha_{R,i} \frac{F_i}{\langle F_R \rangle} \tag{3.2}$$

where $\langle F_R \rangle$ is the average population reaction probability $F$ at round $R$. In comparison to the math described in section 2.1, it is clear that $F_i$ is somewhat analogous to sequence fitness $x$ but capped at a maximum value of 1, since no sequence can survive the selection step at greater abundance than it began with.

For a population of different sequences, it is relatively easy to estimate *relative* values of $F_i$ for each sequence. We define relative enrichment $E_{R,i}$ as the rate at which sequence $i$ increases in abundance from round $R$ to round $R + 1$, such that

$$E_{R,i} = \frac{F_i}{\langle F_R \rangle} = \frac{\alpha_{R+1,i}}{\alpha_{R,i}} \tag{3.3}$$

which can be easily calculated if $\langle F_R \rangle$ is known for the round in question. Theoretically, $\langle F_R \rangle$ values could be measured by comparison to a sequence of known activity or quantitative estimates of the sequence population surviving each selection step; we suggest several more complicated, but possibly also more accurate, methods to estimate it in the next section. Notably, if we compare two different rounds $R$ and $R'$, we expect to see

$$E_{R',i} = E_{R,i} \frac{\langle F_R \rangle}{\langle F_{R'} \rangle} \tag{3.4}$$

as $F_i$ is an unknown but constant "hidden value" for each sequence $i$; plotting $E_r$ against $E_R$ would be expected to give a strong linear correlation.

In the case of a first-order ribozyme selection, $F_i \approx A_i k_i [S] t$, for sufficiently low values of $A_i k_i[S]$t. This is the rationale by which $E_R$ has been used as a proxy for sequence activity; assuming most sequences have low enough activity under selection conditions to fall into this roughly linear regime.

If, however, observations are performed at multiple different $t$ or $[S]$ values, data points for first-order ribozyme catalysis can be fit to the equation

$$F_i = A_i\left(1 - \mathrm{e}^{\mathrm{k_i}[S]t}\right) \tag{3.5}$$

where, for a given sequence $i$, the reacted fraction $F_i$ indicates reacted fraction, $A_i$ is a maximum activity constant (which accounts for both sequence-dependent folding and stability, as well as any loss during recovery), $k_i$ is a sequence's catalytic rate, $[S]$ is substrate concentration, and $t$ is reaction time. For a cell-free *in vitro* enzyme selection, we would use the same math, with $A_i$ representing the rate of proper protein expression. In the case of an aptamer or *in vitro* antibody or peptide aptamer selection, we would instead use

$$F_i = A_i\left(\frac{[S]}{k_{D,i} + [S]}\right) \tag{3.6}$$

where $k_{D,i}$ is an individual biopolymer's dissociation constant (assuming substrate concentration is significantly greater than aptamer/antibody concentration). In a cell-based selection or more complicated catalytic mechanism, the equations become somewhat more specialized to a specific biochemical mechanism, but the same process can be used if there exists any predicted equation for the dependence of selection fitness on a varying substrate concentration or other varying reaction parameter.

## 3.2. Triphosphorylation selection shows surprising enrichment variability

Just what, exactly, happens in an actual in vitro selection? To answer this, we sought to examine an existing selection for which many rounds of sequence abundance data could be available, allowing us to track the enrichment of each of $10^4$-$10^6$ sequences over many rounds of selection. In this, we sought the first attempt to observe and compare compare population dynamics in real in vitro selection to that suggested by theory. We chose the selection of TMP triphosphorylation ribozymes, carried out by a collaborator (as described in section 1.5) for this investigation.

This ribozyme selection began with a random pool (N150), whose effective complexity (1.7 x $10^{14}$ starting sequences from a theoretically possible set of 2.0 x $10^{90}$) far exceeded the capacity of HTS. The selection thus illuminates many random 'pinpoints' in sequence space [68], giving a picture of how catalytic activity may be distributed across an extremely large and randomly-sampled RNA sequence space. Although the complexity of this pool prevents detailed mapping of the complete fitness landscape [99], our selection yielded hundreds of high-fitness sequence clusters, which provided data suitable for generating a potential overall probability distribution of fitness.

Ribozymes capable of self-triphosphorylating a 5'-OH end using a trimetaphosphate (TMP) substrate were selected as described in a previous study[99] Following incubation with 50 mM of trisodium TMP under buffered conditions, a ligase ribozyme was used to regenerate full-length sequences from triphosphorylated 5'-OH ends. Due to the dependence of selection on a successful ligation event, the constant $A_i$ described in equation (3.5) was expected to contain both a stability-related and ligase-favorability component, thus varying more widely between sequences than in a simple one-step ribozyme. The first four rounds of selection were carried out with 3-hour TMP incubation, while the pool at Round 4 was separated into two branches; the 3h branch continued these conditions, while the 5m branch underwent TMP incubation for only 5 minutes. Both branches, after round 4, were subjected to mutagenic PCR, causing the pool to undergo directed evolution in later rounds. All rounds of selection were sequenced via Illumina MySeq, with 1.6-4.8x$10^6$ sequences counted per round.

Sequences from each round were grouped by similarity using Chen lab tools, into unique ribozyme families, with the highest-abundance sequence in each family defined as the center. All families were separated by large edit distances from each other due to the long length of the random region, allowing unambiguous assignment of sequences to families. Some families displayed a shift in center sequence across rounds, typically consisting of 1-2 nucleotide mutations (termed 'notable' mutations). For some analyses, similarity to one of multiple variant centers was used to “split” families into two or more new smaller “clusters” of sequences, treated independently, as the notable mutations impacted fitness. Sequence families could not be reliably identified in Rounds 1 and 2 due to the large number of unique sequences, but 829 ribozyme families were identified in Round 3. These were gradually winnowed over subsequent rounds. Over a hundred unique families were present at the end of Round 8 of both branches of selection, and several of the major families were best analyzed after splitting into clusters based on the presence or absence of notable mutations (see below). The top 20 families comprised approximately 80% of the pool (Figure 3.1.A,B).

The presence of many ribozyme families indicated that a low-throughput approach would not be sufficient to identify the fittest sequences (Figure 3.1.C), and sequences from previous analysis of the selection (in which approximately 40 colonies selected from transformants from various selection rounds were Sanger sequenced and assayed for activity) belonged to some, but not all, high-abundance families. As described in Section 3.4, these previously-determined sequences were for the most part far less active than those chosen through HTS methods to have high activity.

One striking feature that emerged from calculated enrichment data was the sheer lack of correlation between most sequences' enrichment in any two rounds. As all sequences should increase/decrease at a rate directly proportional to a fixed constant (equation (3.3)), we expected sequences' enrichment values to be linearly correlated from one round to the next. Instead, when plotting all sequences' enrichment across two separate rounds of selection, all combinations of rounds produced a graph with effectively zero linear correlation; plotting the enrichment of entire sequence clusters gave only a slight correlation (figure 3.2). This suggests a significant variability in the round-to-round enrichment of individual sequences under *in vitro* selection, an effect far more significant than any previously reported, with the potential to throw into question single-round enrichment as a tool for estimating a sequence's selection fitness.

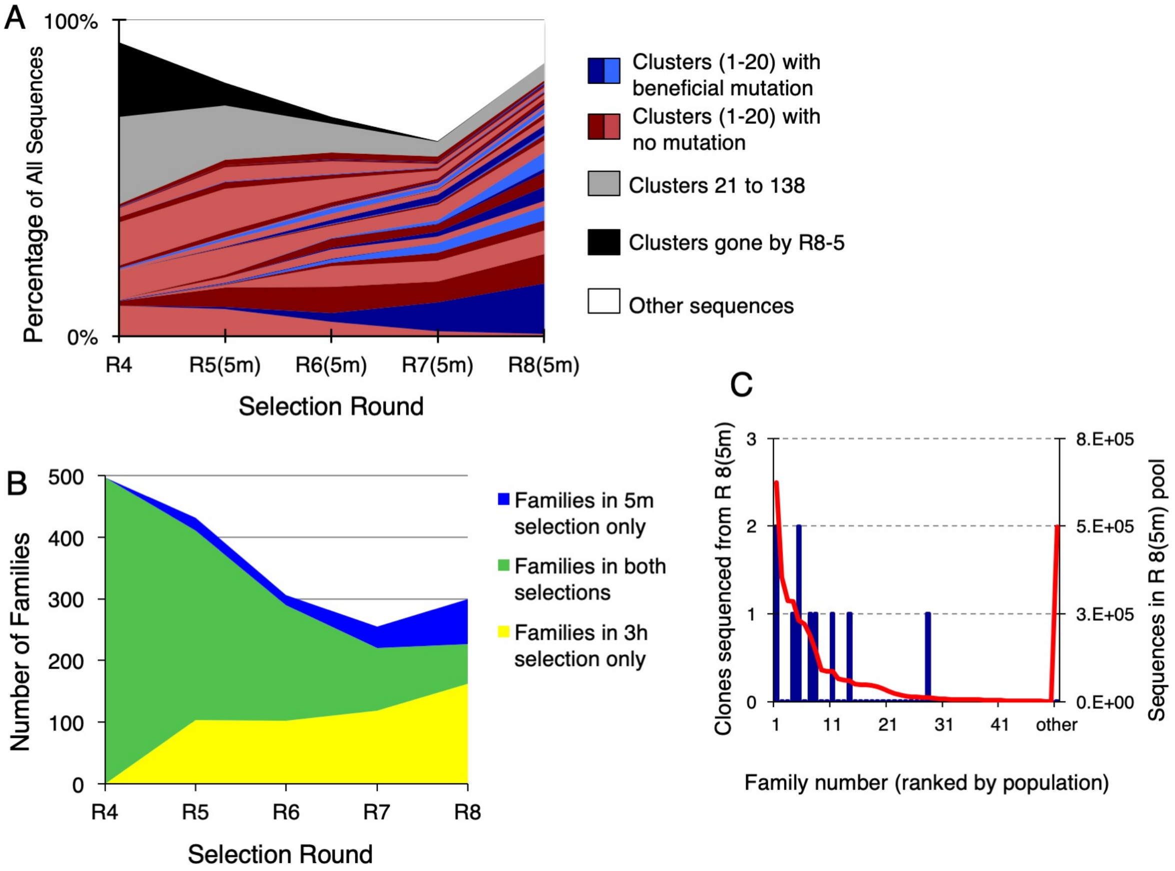


**Figure 3.1. Identified triphosphorylase families and clusters**

**(A)** The abundance of clusters over time in the 5m branch of selection: red areas represent clusters close to the original centers of the top 20 families of Round 8(5m), blue areas represent new clusters whose central sequence diverged from the original family center. **(B)** The 5m and 3h selection branches were separated after Round 4; some sequence families remained present in both selection branches, but more sequeneces disappeared in the 5m branch than the 3h branch. Even at the end of selection, hundreds of unique sequence families were present. **C)** Previously identified clones were distributed among HTS-identified families in a manner consistent with family abundances measured by sequence reads (i.e., clones were mostly derived from higher-abundance families). Blue bar chart (and left y-axis) shows the distribution of clones obtained in the previous study in round 8 (5m); red (and right bar) shows the distribution obtained from HTS in that round.

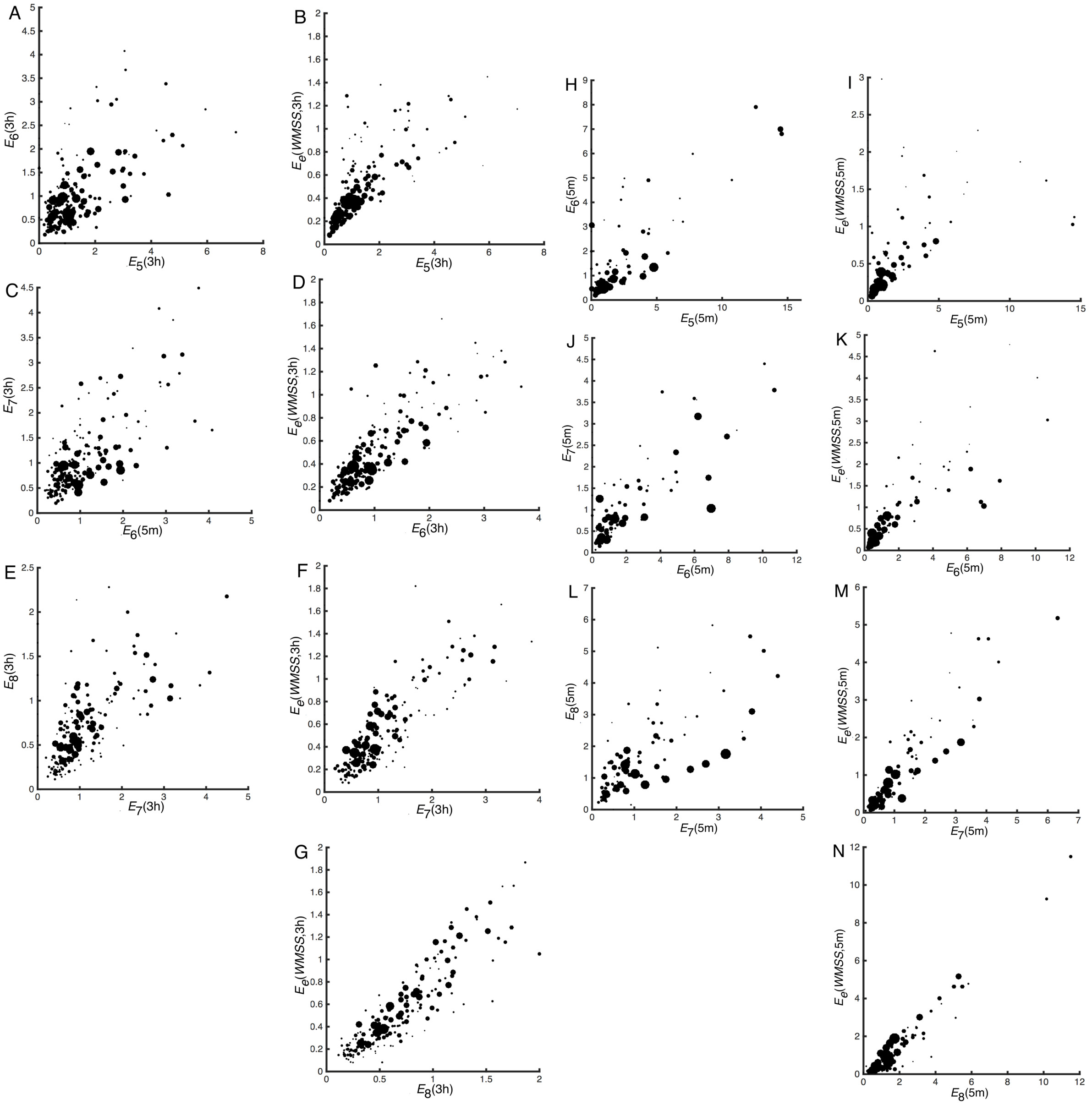


**Figure 3.2. Enrichment correlation over clusters, across multiple rounds**

**(A-G)** Cluster-based enrichment tracked across the 3-hour-incubation selection branch for Rounds 5-8, with dot area corresponding to cluster abundance (For $E_R$ and $E_{R+1}$, cluster area corresponds to round R abundance $A_R$). **(H-N)** Cluster-based enrichment similarly tracked for the 5-minute-incubation selection branch. "WMSS" cases (B,D,F,G,I,K,M,N) indicate fitness estimate from Method 4, showing somewhat higher correlation values across all rounds.

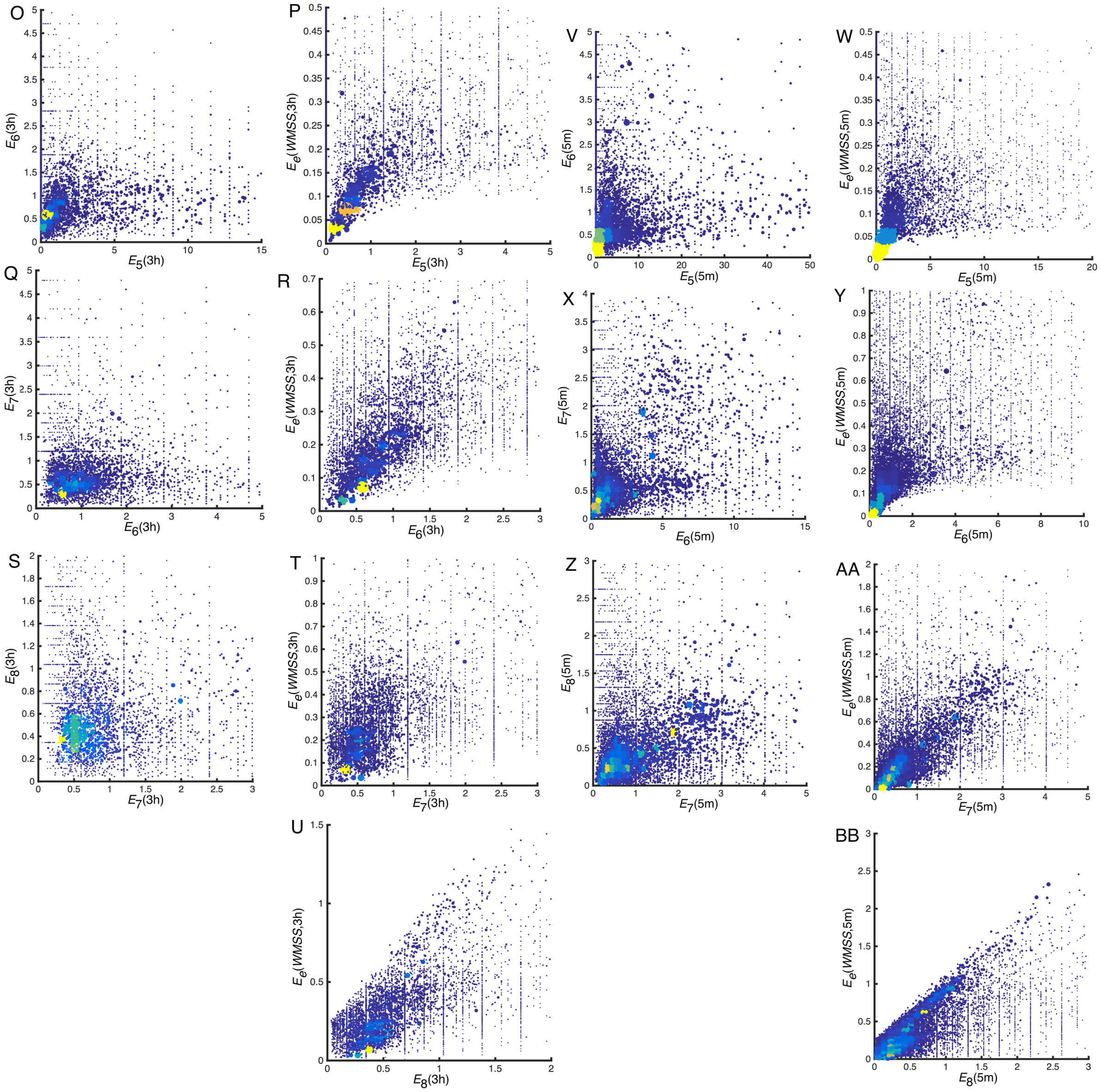


**Figure 3.2. continued: Enrichment correlation over sequences, across multiple rounds**
**(O-V)** Individual sequence enrichment in the 3-hour branch for the same rounds, with dot area corresponding to sequence abundance and dot color corresponding to local sequence density (based on the total sequence count present in each bin of a 50x50 grid, with blue corresponding to lowest sequence density and orange to highest density). (**W-BB)** Individual sequence enrichment in the 5-minute branch. “WMSS” cases (P,R,T,U,W,Y,AA,BB) indicate fitness estimate from Method 4, showing somewhat higher correlation values across all rounds.

### 3.3. Theory: Heuristic evaluation of methods for estimating activity

Prior to this analysis, while existing work had attempted to estimate fitness of aptamers and ribozymes through high-throughput selection, no work had sought to categorize the accuracy or consistency of more than a few of the estimates. As discussed in section 1.4, simple metrics such as abundance can be a poor estimate of chemical activity. We instead sought a heuristic method to evaluate the "goodness of fit" for various fitness estimation methods, as well as a specific way to calculate expected error of estimation from multiple-round selection data. From these, we sought to categorically determine the best method (of both described methods and new ones) to estimate the activity of a large number of ribozyme sequences from the provided data, which consisted of multiple rounds' sequence abundances in two separate selection branches.

With the ability to evaluate the best fitting method over a given selection, we believe such evaluative criteria could be applied to a large range of selection/molecule types in order to determine what is an appropriate fitness estimation, and what range of error to expect, in other experimental systems.

We tested six methods of calculating "estimated fitness" $F_e$, an estimate each sequence's "true" (but hidden) fitness value $F_i$. Although the ideal metric for evaluating $F_i$ would be the correlation between $F_i$ and $F_e$, it is not possible to independently determine $F_i$ without testing individual ribozymes, which is infeasible for the large number of unique sequences involved in the selection. Therefore, because $F_i$ should be proportional to $E_{R,i}$ for each round (equation (3.3)), the methods for calculating $F_e$ were evaluated by correlating the round-scaled estimated overall enrichment ($E_{R,e}$) to ($E_{R,i}$) for the same sequences at different rounds of selection, where $E_{R,e} = \frac{F_e}{\langle F_R \rangle}$. For a correlation metric evaluating the constancy of the estimations across rounds, we used the weighted coefficient of determination ($r^2$) constant. As abundance in the first of two consecutive rounds represents the number of individual times the sequence is observed undergoing selection, we used sequence/cluster abundance of each molecular species as weights  in calculating $r^2$. While each $E_{R,e}$ estimation differs from $F_e$ by a scaling factor of $1/\langle F_R \rangle$, the $r^2$ correlation between two data sets is unaffected by linear scaling of either set, allowing comparison of $E_e$ and $E_{R,e}$ without regard to normalization. Thus rescaling was only performed for the method chosen for further analysis (Method 4), with which we can calculate $F_e$ by comparison of $E_{R,e}$ at multiple incubation times, as described later in this section.

Single enrichment ratios have been used as a rough predictor of sequence fitness previously.[95-97] Therefore, Method 1 of fitness estimation used the previous round (PR) enrichment as an estimate of overall fitness (thus picking one round and comparing it to the others to give our heuristically-chosen correlation metric $r^2$). Here, enrichment $E_{R\text{-}1,e}$ (normalized by average enrichment for round $R\text{-}1$) is a predictor of $E_R$ (normalized by average enrichment for round $R$) as a baseline against which the correlation of other fitness predictors could be measured, such that $E_e(\text{PR}) = E_{R\text{-}1}$ at round $R$.

Information from multiple rounds of selection can be integrated in several possible ways, and we expected that the additional information would lead to increased accuracy of estimation. In other words, $E_5$, $E_6$, $E_7$, and $E_8$ could be used together to estimate an underlying $E_e$ that does not correspond to any one specific round (but is then scaled by an unknown constant). The simplest multi-round method tested, Method 2, uses the geometric mean (GM) of $E_{R,e}$ over the selection rounds, an approach previously used in population genetic studies of artificial selection.[87] The geometric mean is a natural choice when populations are sampled only every several generations, effectively allowing only a geometric mean of enrichment to be observed; the result, $E_e$(GM), would be $F_e$(GM) scaled by the geometric mean of $\frac{1}{\langle F_5\rangle \ldots \langle F_8\rangle}$ . $E_e$(GM) is calculated, for a sequence *i*, with $E_R$ measured from Round *I* to Round *J*, as:

$$E_e(GM) = \left(\prod_{R=I}^{J} E_{R,i}\right)^{J-I+1} \qquad (3.7)$$

Methods 3-5 were based on multivariable least squares estimation. In each method, estimated overall enrichment $E_e$ of a sequence or cluster was calculated as a linear combination of $E_5$, $E_6$, $E_7$, and $E_8$. Coefficients $C_R$ were chosen to minimize squared residual values of $E_e$ vs $C_R E_R$ for each sequence, summed across *R* from 5 to 8, and weighted by sequence observation, as in equation 3. Weighting parameters were chosen to maximize weighted $r^2$ values summed across all rounds of comparison, with Methods 3-5 varying by different approaches to the importance of each round's contribution.

In Method 3, which we term Weighted Multiple Sum of Squares (WMSS), $E_e$ of sequence *i* was calculated as

$$E_e(WMSS) = \left(\sum_{R=I}^{J} C_R E_{R,i}\right) \Big/ (J - I + 1) \qquad (3.8)$$

essentially providing an average of observed enrichment ratios at rounds I=5 through J=8. Constants $C_R$ were chosen to minimize the weighted sum of squared residuals (with each sequence's contribution weighted by its abundance), as follows:

$$C_J = 1;\ C_I \ldots C_{J-1} = \operatorname{argmin} \sum_{\text{all } i} \left( \sum_{R=I}^{J} \alpha_{R-1,i} \left[ C_R E_{R,i} - E_e(WMSS) \right]^2 \right) \qquad (3.9)$$

That is, the contribution to the summation of squared residuals from $E_e$(WMSS) for sequence (or cluster) i was weighted by the normalized abundance $\alpha_{R-1,i}$ of that sequence (or cluster) during each round. The rationale for this weighting is that the more abundant sequences represent a greater number of observations, and therefore offer greater accuracy in estimation of $E_e$. As each round's contribution to this summation function is an independent function of $C_R$ with a single inflection point, gradient descent was used to find minimizing $C_R$ values for each data set, using code written in MATLAB.

In Method 4, which we refer to as Population-Weighted Multivariable Sum of Squares (PWMSS), for sequence $i$, each round's contribution to $E_e(PWMSS)$ was weighted by its abundance in that round (and thus observational certainty) instead of using an identical set of round-specified factors for each sequence as in WMSS. As we expected greater stochasticity in selection than sequencing, observational certainty and total real information was expected to correlate with abundance rather than raw sequence counts. The coefficients in the linear combination estimate varied from sequence to sequence, following:

$$E_e(PWMSS) = \left(\sum_{R=I}^{J} C_R \alpha_{R-1,i} E_{R,i}\right) \Bigg/ \sum_{R=I}^{J} \alpha_{R-1,i} \tag{3.10}$$

with $C_R$ again chosen by minimizing weighted square residuals as in WMSS (equation (3.9), replacing WMSS with PWMSS). Method 4 resembles Method 3, but also takes into account the difference in round-dependent appearance of each individual sequence or cluster. Here we also note that we no longer expect equal contributions from every round, as many sequences appeared at lower count in earlier rounds.

In Method 5, termed Scedasticity-Consistent Multivariable Sum of Squares (SCMSS), $E_e(SCMSS)$ was calculated as in Method 4 (Equation (3.10)) but with information present in observations weighted by the expected variance of each observation. In a scedasticity-consistent modified sum of squares, squared residual values are divided by expected variance of each data point. After a round of selection, we expect sequence variance equal to mean next-round abundance (that is, abundance times viability), as in equation (3.2), but note that the variance of each individual round contribution's impact should scale with its mean of $C_R E_e$. Thus, we divide weighting terms in Equation 3 by $C_R E_e$, giving Method 5 following equation (3.10) but with $C_R$ chosen as:

$$C_J = 1;\ C_I \ldots C_{J-1} = \operatorname{argmin} \sum_{\text{all } i} \left( \sum_{R=I}^{J} \frac{\left[C_R E_{R,i} - E_e(SCMSS)\right]^2}{C_R E_e(SCMSS)} \right) \tag{3.11}$$

The last method of fitness prediction (Method 6) evaluated here used a Maximum-Likelihood Estimator (MLE) for each sequence's fitness. As described previously, we expect NF to be high enough for most sequences persisting through the end of selection to be approximated with a normal distribution. As this incarnation of the MLE method requires an initial estimate of the average sequence fitness for each round, the PWMSS estimate was used, with a conditional probability distribution $\mathrm{P}\left(\mathrm{E_e(MLE)} = \mathrm{E_{R,i}} | \mathrm{E_e(PWMSS)}, \alpha_{\mathrm{R-1,i}}\right)$, as a normal distribution whose variance scaled with abundance

$$E_e(MLE) = \operatorname{argmax} \left( \prod_{n=I}^{J} P\left(E_e(MLE) = E_{R,i} | E_e(PWMSS), \alpha_{R-1,i}\ \right) \right) \tag{3.12}$$

This probability assumption P(..) used normal distributions for expected population after selection and PCR, based on the distribution of scale-dependent noise observed with Method 4 (as described in noise analysis below). For each i, a range of 1000 $\mathrm{E_e}$ values were chosen ranging from the highest to lowest suggested by each round, scaled by the round-weight parameters generated by Method 3 (as these were assumed to be the best easily-obtainable approximations of the relative scaling between enrichment in each round).

Overall, Methods 4 and 5 showed the highest correlation to most rounds, for both individual sequences and clusters of similar sequences (which in the clustered approach were assumed to share a single activity). Figure 3.3 shows how the weighted $r^2$ metric stacks up for rounds 5-8, while Figure 3.2 shows how method 4 provided a more consistent estimate of $E_R$ than sing a single round's enrichment. For the round 5-8 triphosphorylation enrichment data, Method 2 had a tendency to over-fit to round 5 enrichment, while Method 6 over-fit round 8 enrichment, with both fitting poorly to all other round. Method 4 was chosen as the "best" method for this particular analysis, as it was mathematically simpler than method 5 while showing similar goodness of fit across the population.

As the triphosphorylation selection occurred in two separate selection branches with different time, the relationship between these two was used to normalized $E_e$(5m) and $E_e$(3h). We note that $F_e = A_i\left(1 - e^{-k_i[S]t}\right)$ and thus $E_e/C_e = A_i\left(1 - e^{-k_i[S]t}\right)$, where $C_e$ is a fundamentally unknown scaling constant resulting from any of our fitting methods, though in this case specifically Method 4. Then we note that, for each individual sequence,

$$\frac{E_e(t_1)}{A_i} = S_e(t_1)\left(1 - \left[1 - \frac{E_e(t_2)/A_i}{C_e(t_2)}\right]^{\frac{t_1}{t_2}}\right) \tag{3.13}$$

a relation derived from setting $k_i[S]$ equal at two different incubation times, with $E_e$ being a linear rescaling of $F_e$ for all sequences.

The relationship of $E_e(t_1)/A_i$ to $E_e(t_2)/A_i$ over many sequences was used to determine the constants $C_e(t_1)$ and $C_e(t_2)$, for $t_1$ = 5 min and $t_2$ = 180 min. Curve fitting analysis between $E_e$ for the two pools was performed with the Matlab curve fitting toolbox, using standard settings for a nonlinear weighted fit to Equation 7 and weighting each sequence by a geometric average of total counts in 5m and 3h pools (summed over rounds 4-7).

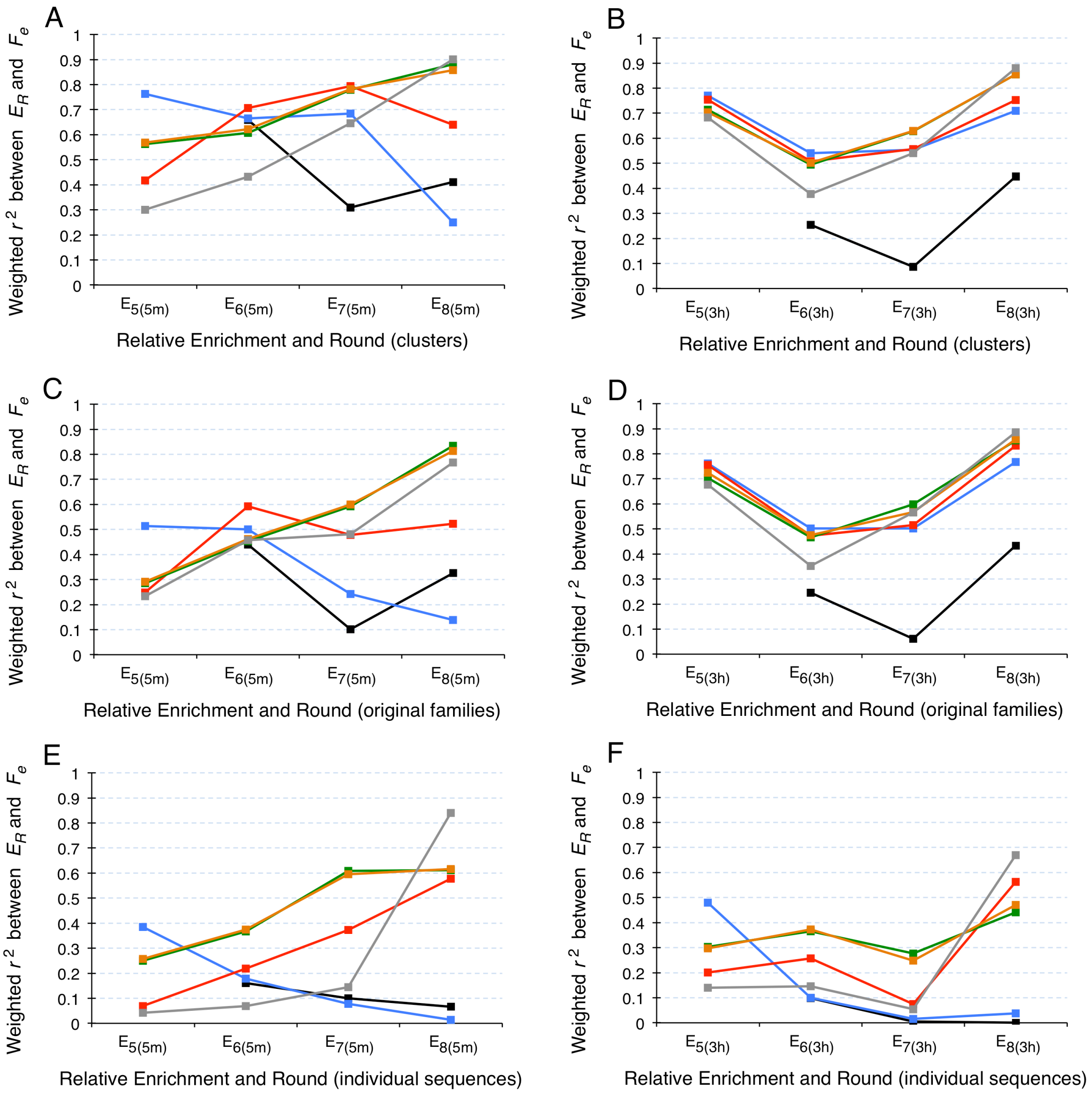


**Figure 3.3. Fit correlation of multi-round fitness estimates**
Correlations were greater between $E_e$ and $E_R$ when measuring the propagation of clusters and families **(A-D)** than when measuring individual sequences **(E-F)**. Additionally, splitting into clusters based on the presence or absence of notable mutations (A and B) gave higher correlations than families grouped solely by similarity (C and D). Nevertheless, trends in the effectiveness of fitness estimation methods were similar for both sequence and cluster analysis, for both selection branches. $r^2$ correlation does not depend on scaling or normalization. Black: Method 1; Blue: Method 2; Red: Method 3; Green: Method 4; Orange: Method 5; Gray: Method 6.

### 3.4. High-throughput activity estimation finds better ribozymes

We sought to determine whether the HTS fitness analysis could identify higher activity ribozymes than those previously identified through the arbitrary sampling and Sanger sequencing of 36 clones from Rounds 5 and 8. We chose eight sequences with high $F_e$ and high prediction confidence (see Appendix Table A.1 for details), and tested their activities by reaction with TMP and ligation of the triphosphorylated ribozyme by a ligase ribozyme, thus mimicking the conditions of the selection procedure [99]. Importantly, conditions were chosen to precisely represent those of the *in vitro* selection, thereby measuring fitness as experienced during the evolution procedure. Each experimentally-measured reaction was done in triplicate, with multiple calculated values used to obtain average and standard deviation, as fit to equation (4.2). Several sequences from the previous publication were chosen and tested alongside these for comparison.

New sequences reached considerably higher experimental activity than the ribozymes previously identified from the same selection (Figure 3.4, Appendix Table A.1); by contrast, sequence activity showed no correlation to abundance. Overall, this approach identified ribozymes with substantially greater activity while testing fewer individual sequences; six of the eight high-$F_e$ sequences showed activity greater than or equal to the best of 36 previously-tested ribozymes, by a factor of up to 10-20-fold.

Most of the clusters of highest estimated fitness carried notable mutations, in that they contained sequences that outcompeted the original highest-count sequence in the family between Rounds 4 and 8. In the more stringent 5m selection branch, 34 out of the 59 highest-abundance peaks at Round 8 displayed at least one notable mutation from the central sequence of Round 4, such that a large portion of the pool consisted of sequences similar to these mutants. In such cases, the notable mutation appeared to demonstrate significantly increased survival fitness, with the new cluster rapidly enriching to outpace the old sequence center, with one such sweep shown in Figure 3.5. For clusters with high abundance at Round 4, notable mutations (that would dominate in later rounds) typically each accounted for less than 1% of the cluster population at Round 4; thus, out-competing the original center over the next four rounds of selection would require a mutation with at least $100^{1/4}$ times (~3x) the fitness of the original center. To determine the effect of notable mutations on ribozyme activity, four sequences (1-S, 2-S, 6-S, 11-S) that clustered with previously tested clones, but also possessed notable mutations, were among those assayed experimentally. Three of the four mutants exhibited higher activity compared to the best previously identified clone from the same cluster (up to a five-fold increase), indicating that the notable mutations were usually beneficial. One sequence (2-S) showed a five-fold increase in activity despite a difference of only a single nucleotide (Appendix Table A.2).

Overall, understanding the effects and prevalence of mid-selection mutations falls outside the main scope of this project, although chapter 6 describes epistatic analyses that may give insight into directed evolution during a selection. That said, our predictive methods here appear to estimate the activity of both high-count and rarer, mid-selection-mutation-derived sequences with similar accuracy. This suggests that similar high-throughput approaches may be useful for understanding or optimizing mutational steps and parameters in addition to the selection-driven portion of controlled evolution.

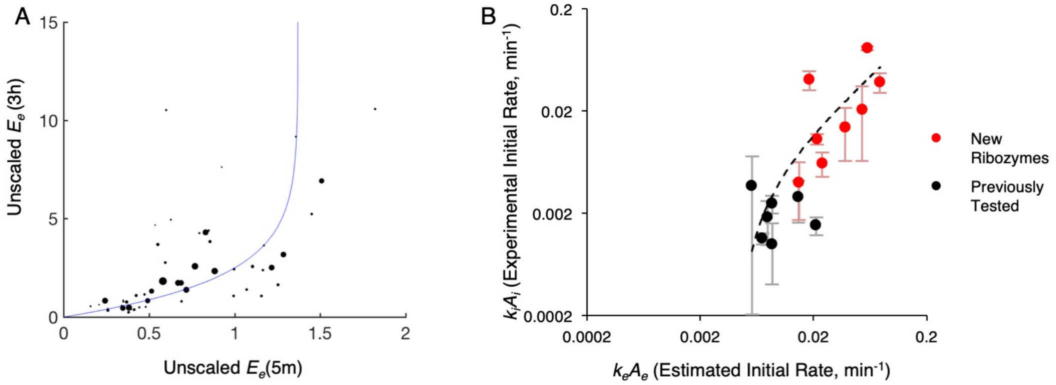


**Figure 3.4. Estimating catalytic rates for triphosphorylation ribozymes**
**(A)** Comparison of fitness estimated from the 5m and 3h branches (weighted $r^2 = 0.87$), fit by setting k equal at two time points according to equation (3.13); the fitted parameters provide scaling constants $C_e$(5m) and $C_e$(3h). The area of each dot is proportional to relative abundance of each cluster. **(B)** Comparison of $k_eA_e$ estimated from HTS fitness (Method 4, split clusters along 5m branch), with $k_iA_i$ observed as experimental chemical activity of individual sequences, for sequences described in Appendix Table A.1. Black points correspond to previously identified and tested sequences; red points correspond to eight new sequences expected to have high fitness and tested biochemically in the present study. Observed values of $k_iA_i$ were obtained in triplicate, with error bars corresponding to standard deviation. The error ranges for $k_eA_e$ for individual sequences are expected to be on the order of ±50% (Figure 3.6). Overall, these points (with the linear trend line shown as a dotted black line) show an $r^2$ correlation of 0.52.

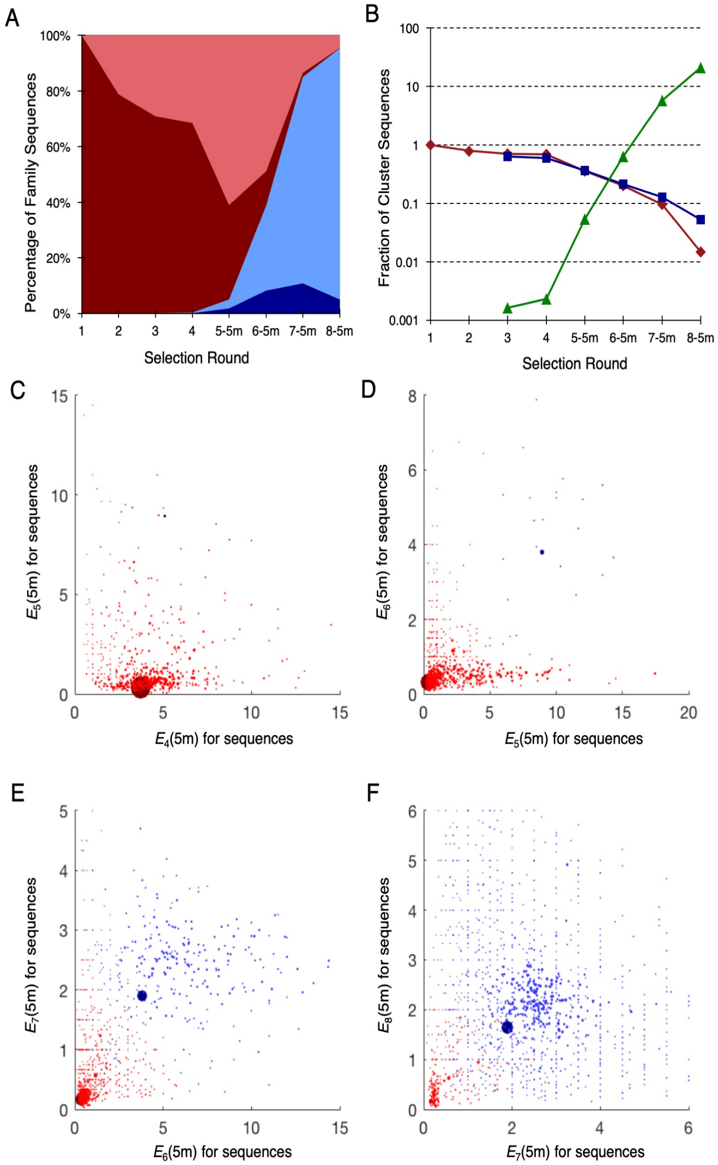


**Figure 3.5. Emergence of a notable mutation within a ribozyme family**

**(A)** Relative abundance of sequences within a split family (Family 1, with clusters 1-O and 1-S), the highest-abundance cluster at Round 8(5m). The cluster originated from in Round 1, consisting of a single center sequence (dark red), with similar mutants (light red) appearing in subsequent rounds. At Round 4, a notable mutation arose and swept through the family (center sequence in dark blue), accompanied by its mutants (light blue, consisting of all sequences closer in edit distance to the new cluster) as the original cluster center was outcompeted. Note that Rounds 4-8 experienced mutagenic PCR, resulting in sequence mutants appearing far more frequently. **(B)** The proportion of sequences clustered with the original center or shifted center over rounds. Red points/line show the fraction of the non-mutant cluster (cluster 1-O) that is composed of the original center sequence; blue shows the fraction of the new split cluster (cluster 1-S) composed of the new center sequence. The green line shows the ratio of new cluster sequences divided by old cluster sequences in the population. **(C-F)** Observed enrichment for Rounds 4-8 for all identified sequences within this family. On each graph, the x-axis denotes $E_R$, and the y-axis denotes $E_{R+1}$, with dot area corresponding to abundance at Round R. Blue dots denote sequences with an identified notable mutation (with the new center sequence in a darker blue); red dots lack that notable mutation (with the original center sequence in a darker red). This family underwent a clear shift toward the mutant cluster.

### 3.5. Consistency and noise in multiple-round selection enrichment

To analyze the noise present in observations of fitness during the selection experiments, $E_e$ values were calculated for sequences and split clusters using Method 3. As Method 3 weighs contributions from all rounds of selection equally, its weighting constants $C_R$ (calculated as part of the linear combination estimate) were used as estimates of average round-by-round total enrichment and used to normalize $E_R$ for each round to the same scale as $E_e$. An approximate distribution of noise was generated from the absolute difference between $E_e$ and $E_R/C_R$ in each round, using proportional error $\frac{|E_e - E_{R,i}/C_R|}{E_e}$ for each sequence. Plotting abundance against this proportional error, we calculated a sliding standard deviation across the distribution, using a sliding window of width ±10 clusters or ±200 sequences. This analysis was also intended to judge the impact of abundance on fitness estimation error. In a selection behaving as ideally as possible, genetic drift would still be expected to contribute abundance-driven noise, as a normal distribution with mean and variance both equal to the ratio of sequenced pool size to real pool size.[100]

Analysis showed that lower-abundance sequences enriched with greater noise, but enrichment noise did not drop below a certain proportional threshold for high-abundance sequences. This suggests that abundance-dependent noise (e.g., genetic drift) dominated the early rounds of selection and abundance-independent noise or error (e.g., experimental variations) had a greater effect on the enrichment of high-abundance sequences in later rounds of selection. Scale-dependent noise impacted individual sequences more than clusters, suggesting that in this case of isolated, narrow sequence clusters with long conserved motifs, enrichment of sequence clusters might be a better predictor of the fitness of individual sequences (Figure 3.6).

Notably, the proportional magnitude of abundance-independent noise was similar to the variation observed in experimental activity measurements. Such variation in sequence enrichment from round to round may be the result of experimental variation or sensitivity to minute changes in reaction conditions; over many rounds of selection, it may also have been responsible for some of the lack of agreement between sequence abundance and kinetic activity.

To evaluate whether the evolution experiment would be suitable for retrospective analysis (described below), we measured the extent to which the selection as a whole followed ideal behavior as predicted by a basic Fisher's Theorem (FT) analysis (see section 2.1). In the case of discrete selection rounds, FT states that, assuming each allele's fitness does not change, the change in average fitness of a population should equal the variance of sequence fitness in the population (normalized by mean fitness). While FT can predict general changes to a population's fitness distribution, we used it here as an accuracy test. Specifically, FT was used to gauge the self-consistency of fitness estimation with actual changes in population composition, made possible by the multiple rounds of selection data. Obedience to FT by an evolving population would imply that the evolutionary dynamics are well-behaved and governed by rules of natural selection, and this was the case for evolutionary dynamics with fitness estimates for both clusters and individual sequences (Figure 3.7 A,B), suggesting that selection was the primary factor driving changes in the estimated fitness distribution of the triphosphorylase ribozyme pool. This concurrence indicates that that the selection behaved predictably and confirms that fitness mean and variance were accurately estimated. Sequence clusters followed FT more closely than individual sequences, consistent with our earlier observation that a cluster-based analysis is subject to less noise.

As expected, mean expected fitness $\langle F_R \rangle$ increased during the selection, with the higher stringency 5m selection branch resulting in approximately 2-fold greater $\langle F_R \rangle$ than the 3h branch by Rounds 7-8. Interestingly, the variance of fitness also increased over time in both selection branches. Intuitively, this increase is expected during a selection from random sequence space, as the distribution of fitness is initially sharply centered near zero, and then spreads to include higher fitness values. To quantify this effect, we used the generalized discrete form of Fisher's Theorem at the level of third moments, which can be reconfigured to state that that the change in fitness variance ${\sigma^2}_R$ between rounds is expected to equal the mean-scaled skewness $\frac{E[F_R - \langle F_R \rangle]^3}{\langle F_R \rangle}$ of the fitness distribution minus the change in average fitness squared: $({\sigma^2}_{R+1} - {\sigma^2}_R) = \frac{E[F_R - \langle F_R \rangle]^3}{\langle F_R \rangle} - (\langle F_{R+1} \rangle - \langle F_R \rangle)^2$, derived from equation (2.14). The fit of the data to this equation reflects whether the shape of the fitness distribution, as captured by the first through third moments, obeys expected dynamics. As skewness is a higher-order shape factor than mean or variance, this relation is expected to be more sensitive to noise or inaccuracies in the estimated shape of the fitness distribution. Clusters followed this corollary well, although individual sequences did not (Figure 3.7 C,D). Therefore, $F_e$ based on sequence clusters gave a reasonably accurate estimate of skewness over the course of selection, and the shape of the fitness distribution based on clusters behaved in a predictable manner. Overall, these results suggest that Fisher's theorem analysis may be a useful second tool for evaluating the consistency and usefulness of fitness estimation, in cases where multiple rounds of good selection data are present.

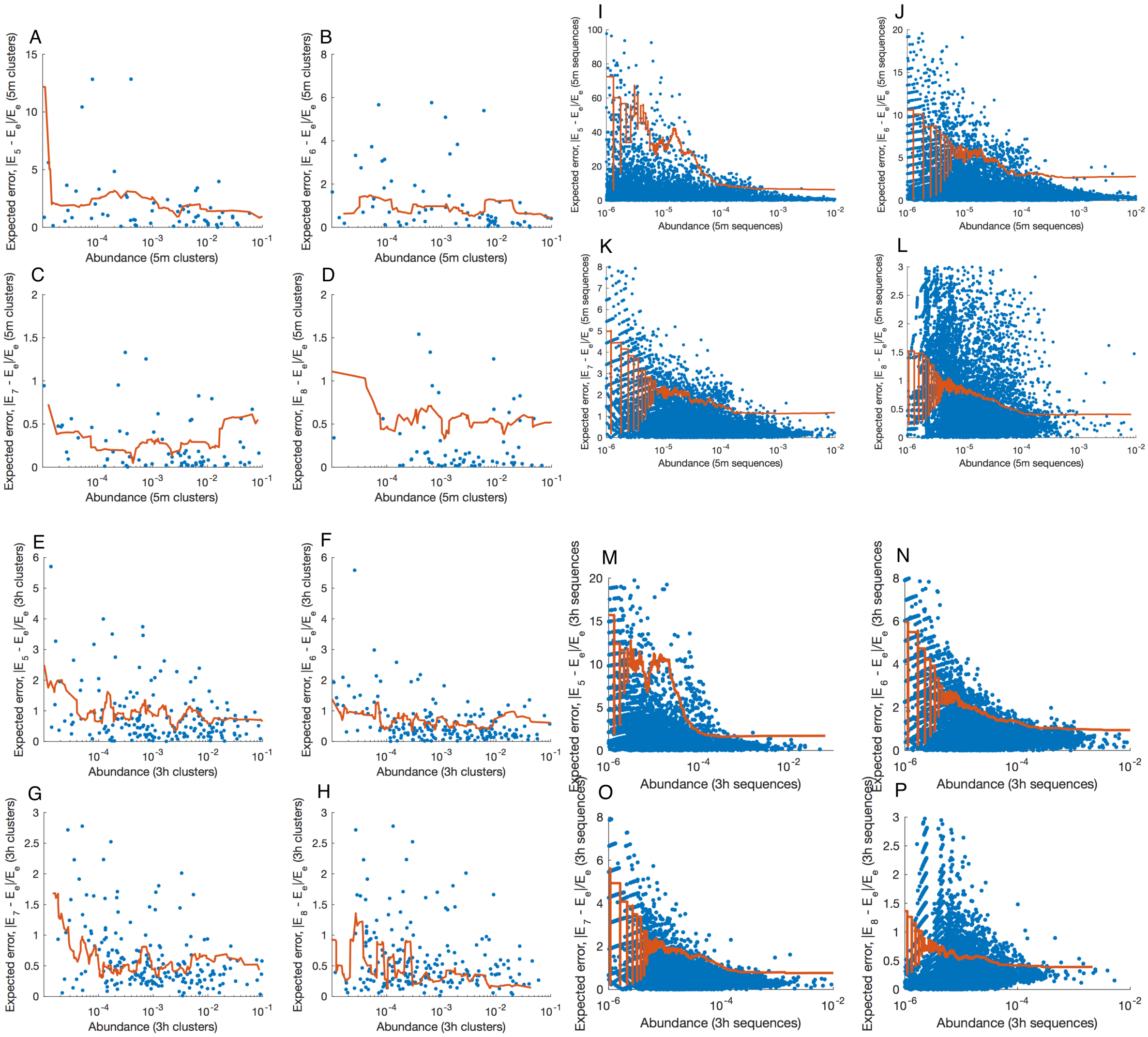


**Figure 3.6. Distribution of enrichment noise across TMPase selection**

The distribution of noise present in this evolution experiment. All panels calculate a moving standard deviation for proportional error $\frac{E_e - E_R/C_R}{E_e}$, (red line) shown above the scatter plot of the absolute value of proportional error, $\frac{|E_e - E_R/C_R|}{E_e}$, for all sequences (blue dots), plotted against abundance. **(A-D)** show error/noise for all split clusters in the 5m selection branch; **(E-H)** show 3h branch clusters. **(I-L)** show all sequences in the 5m selection branch, and **(M-P)** all sequences in the 3h branch. For the moving standard deviation, a sliding window of ±10 points was used for cluster distributions, and ±200 points for individual sequence distributions (this may potentially underfit the long distribution tail).

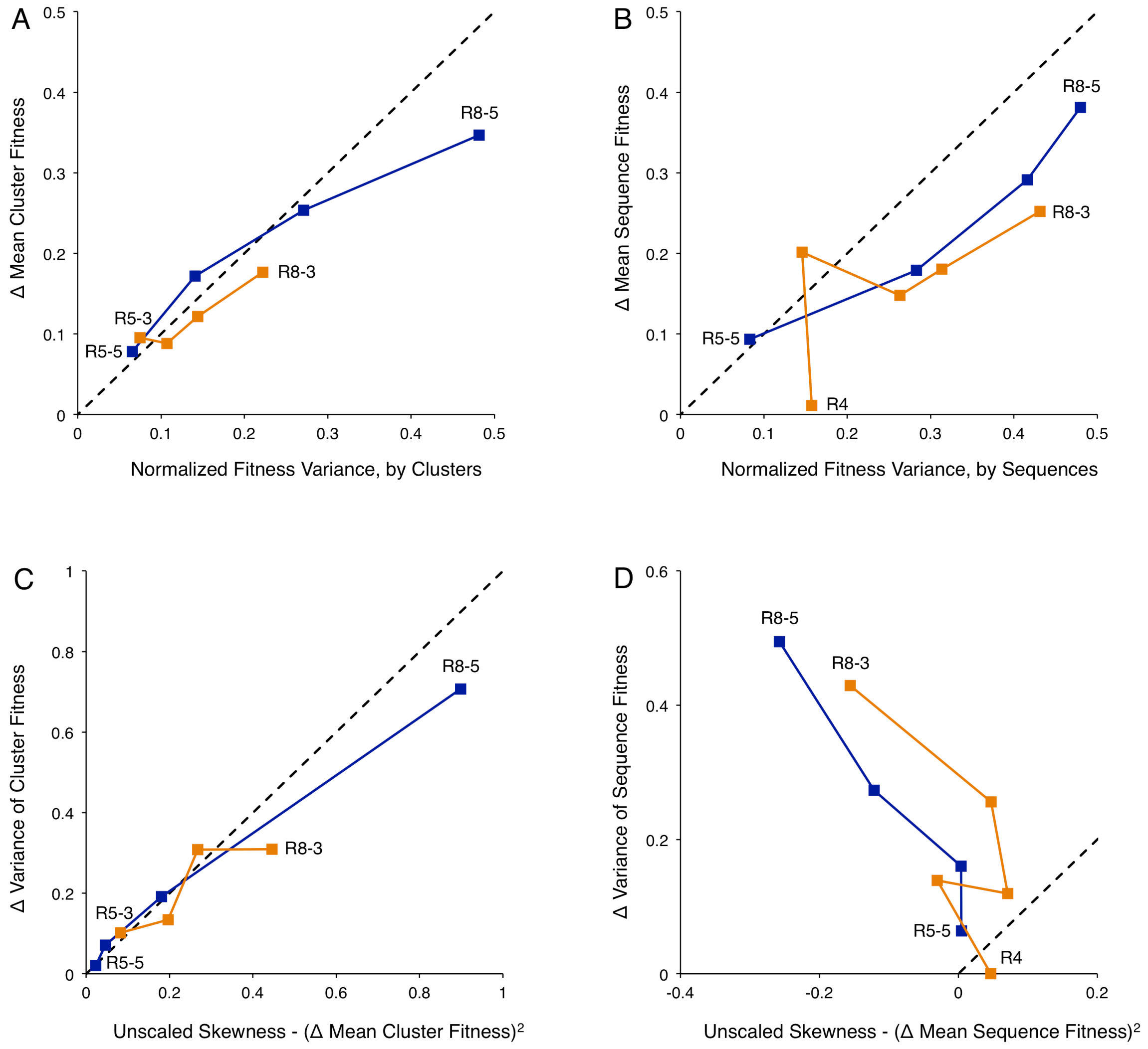


**Figure 3.7. Fisher's Theorem analysis of TMP selection**

**(A)** Cluster-based analysis shows good conformity to FT, suggesting changes in estimated fitness distribution agree with ideal selection dynamics. **(B)** Sequence-based analysis shows some conformity to FFTNS, though less than cluster-based analysis. **(C)** GFT at the level of skewness appears to be mostly followed in sequence clusters, whose skewness and change in variance follow the expected higher-order relation. **(D)**. Sequence-based analysis does not appear to follow the same GFT pattern, with skewness and change in variance of fitness distributions behaving unexpectedly. Dotted lines correspond to the expected 1:1 relationships. Blue: 5m branch; Orange: 3h branch.

### 3.6. Retrospective evolutionary analysis predicts an initial distribution of triphosphorylation activity

Following analysis, cluster-based Method 4 fits were chosen as the most accurate estimate of sequence fitness for selection populations. Since the 3h selection branch underwent constant selection conditions (3 hour incubation time) from Round 1 through 8, $F_e(3h)$ was used to characterize the overall distribution of $F_e$. $F_e$ values were binned into a 40-box histogram to approximate the probability distribution function (pdf) of estimated fitness at round 4, the earliest round of selection in which most sequences had an estimated activity.

Fitness distributions for Rounds 1-3 were calculated by a retrospective approach. In principle, the process of selection translates mathematically into the multiplication of the existing pdf by a selection function, analogous to equation (2.13). Dividing the estimated fitness pdf of one round by the pdf of the preceding round should yield this selection function, which was calculated from observed fitness pdf for each of Rounds 5-8 (Figure 3.8), and which roughly fit to an expected linear shape. As described in the previous section, we consider Fisher's theorem analysis of this evolution of fitness distribution consistent enough to approximate with the ideal case, which follows equation (3.2)—that is, each sequence's abundance is multiplied by its fitness divided by a round's average fitness. Thus, the pdf of fitness for Rounds 1-3 and for the initial pool of random RNA (Round 0) was estimated by back-calculation using $F_e$ (figure 3.9); instead of dividing by average fitness (and potentially incurring errors from any inaccuracy in average fitness), we normalized each round's pdf to integrate to 1, which was expected to have the same effect, such that $p_R(F) = \frac{p_{R-1}(F)/F}{\int_0^1 (P_{R-1}(F)/F)dF}$. (As described in section 3.1, $F$ behaves as fitness value $x$ in Chapter 2, but as a reacted fraction, is bounded to a maximum value of 1).

However, most sequences are expected to fit into the lowest-fitness histogram bin, which corresponds to a sequence space where most random RNA sequences have no catalytic activity, instead triphosphorylating at the base uncatalyzed rate. These exceedingly abundant, low-activity sequences are unlikely to survive past the first round except through stochastic processes, making the lower end of the distribution the hardest to accurately measure. As this is also the overwhelmingly most abundant part of the initial fitness distributions, small variations in the chance survival of a few low-activity sequences could have a large effect on the overall normalization of the initial distribution. This, fitting to a predicted round 0 probability distribution included multiplying proposed distribution functions by an arbitrary scaling factor, to account for this variability.

This initial distribution of reacted fractions/selection fitness was then converted into a distribution of catalytic constants $k_i$ following equation (3.5), which required assuming a single $A$ value, for which we used a rough average of all sequences whose kinetics were had been determined. No previous work had measured or approximated the distribution of a catalytic activity over random molecular space of this scale; here, analysis of HTS data allowed for the conversion of fitness information into kinetic parameters. The affinity of aptamers has been posited to be log-normally distributed, based on a model and experimental data for dsDNA-protein interactions in which individual base pairs contribute independently to overall binding energy [92,93,101]. However, energetic contributions that are correlated along the sequence, such as from DNA bending, could alter this distribution [102]. In the case of RNA folding, a theoretical model suggests that the activation energies of melting follow a Gaussian distribution,[94] while folding simulations suggest that the distribution of minimum free energies for random RNAs is non-Gaussian.[103] In microbial populations, the distributions of fitness effects from new mutations have been fit to a variety of distributions, including normal and exponential.[104] In addition, extreme value theory indicates that a high-value tail can be approximated as a Pareto (scale-free) distribution if the value is comprised of independent random variables, a trend observed in previous comparisons of pool size and activity.[91] We attempted to fit the empirically derived ribozyme rate constant distribution to log-normal, exponential, and scale-free distributions, as a log-log fit, with a log-normal distribution showing the best fit.

To determine whether the three fitness distributions tested (log-normal, exponential, scale-free) could be differentiated in the presence of stochastic noise in abundances, simulations were also conducted. Each fit for the retrospectively inferred Round 0 $F_e$ distribution was assumed to be an initial distribution and binned into a 30-bin histogram. Over four rounds of simulated "selection," the abundance of each bin was increased by the *F* of that bin divided by that round's average *F*, as $F_{bin}/\langle F_R \rangle$. To simulate random noise, the abundance of each bin was also multiplied by a normally-distributed random variable with mean 1 and standard deviation 0.5 (with a minimum bound of 0.05), expected to be equal to or greater than the actual noise in most real sequence's enrichment. Following four rounds of simulated forward selection, each bin's abundance was then *divided* by $F_{bin}/\langle F_R \rangle$, simulating a retrospective inference to recover the original distribution. Finally, the inferred $R_0$ distribution from each simulation was fit on a log-transformed scale to the candidate distributions (log-normal, scale-free, and exponential) and the $r^2$ was calculated, averaged over ten simulations (Figure 3.10).

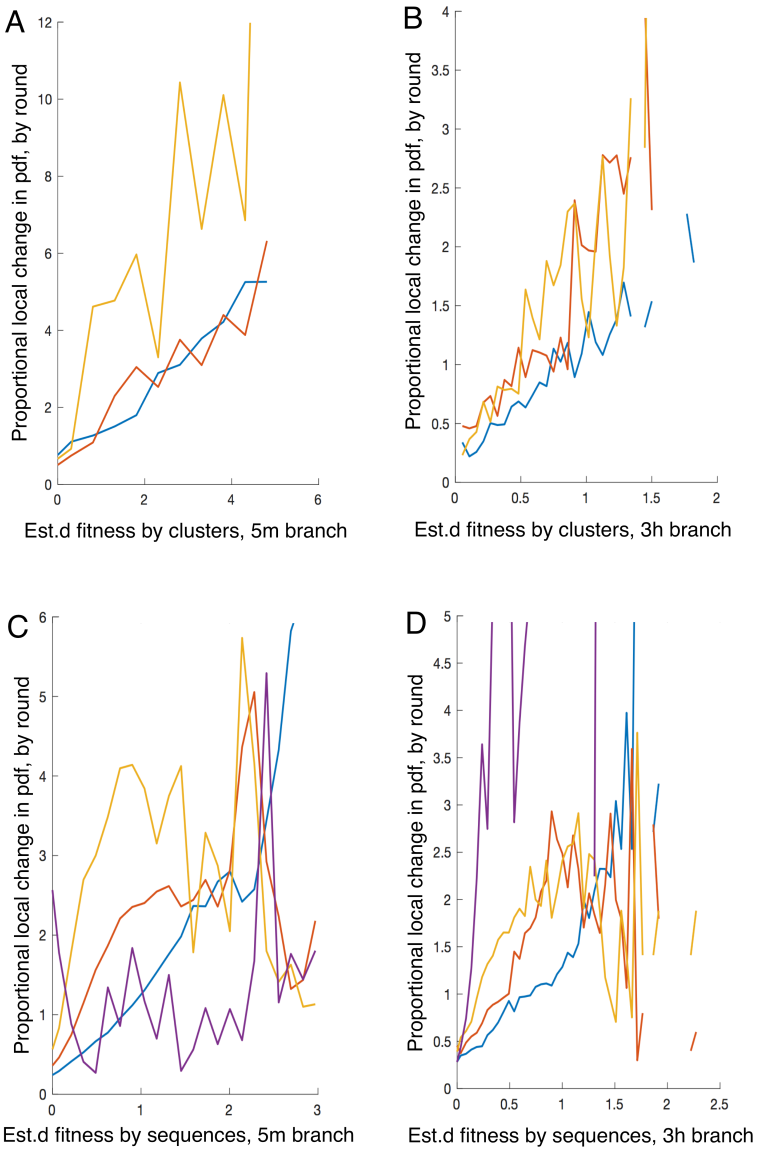


**Figure 3.8. Measured selection function for fitness distribution change in TMPase selection**
The function $p_R(E_e)/p_{R-1}(E_e)$, plotted on the y-axis, is expected to be proportional to $E_e$ (calculated from Round 4-Round 8 analysis), plotted on the x-axis (see Methods: Retrospective inference of underlying fitness and activity distributions; gaps in lines denote fitness histogram bins where $P_R(E_e) = 0$). **(A-B)** Cluster-based analysis (left, 5m branch; right, 3h branch). Approximate linear relationships are indeed observed in cluster-based analysis for ratios of Round 8/Round 7 (Blue), Round 7/Round 6 (Red), and Round 6/Round 5 (Yellow). Since $E_e$ is proportional to $F_e$, these relationships indicate that a similar linear relationship holds for $F_e$. **(C-D)** Sequence-based analysis (left, 5m branch; right, 3h branch) was too noisy for retrospective inference.

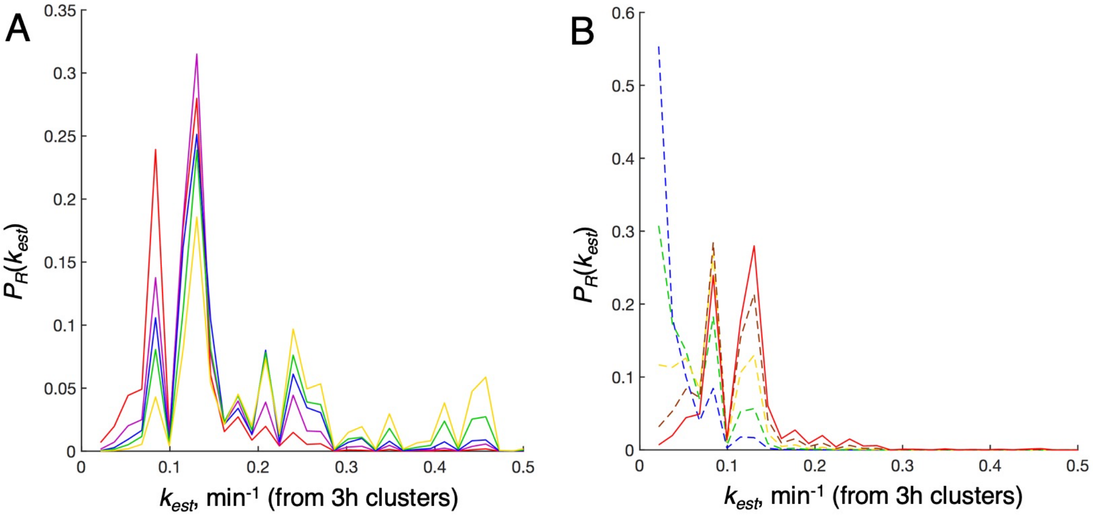
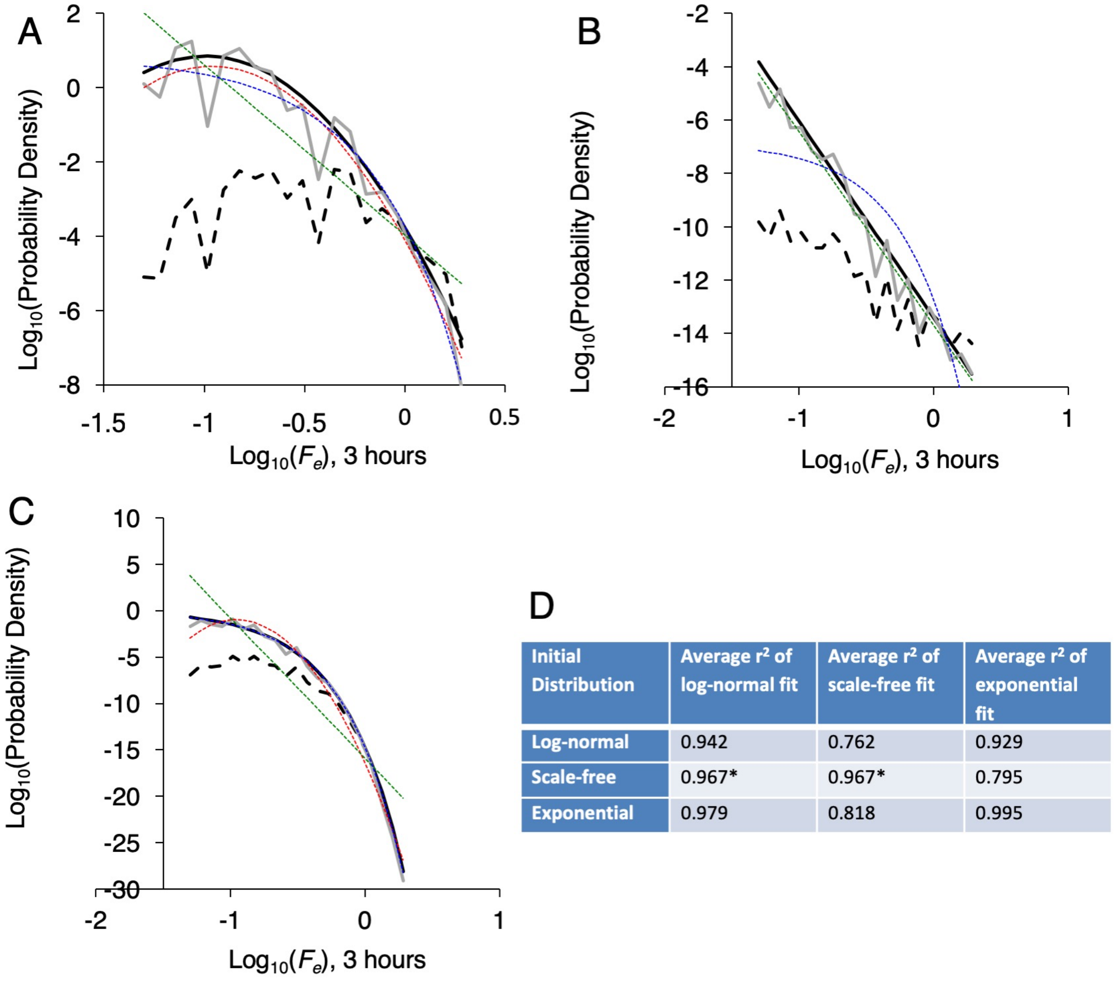


**Figure 3.9. TMPase fitness distribution and predicted distribution evolving across rounds**
**(A)** Distribution of estimated rate constants ($k_{est}$) for sequence clusters over multiple rounds of selection; lines correspond to distributions measured from HTS data of Rounds 4-8 in the 3h selection branch (Yellow: Round 8; Green: Round 7; Blue: Round 6; Purple: Round 5; Red: Round 4). **(B)** Inferred distribution of $k_{est}$ for sequence clusters in earlier rounds of selection; solid red line corresponds to Round 4 distribution, while dashed lines represent the retrospectively inferred distributions for Rounds 0-3 (Brown: Round 3; Yellow: Round 2; Green: Round 1; Blue: Round 0, i.e., initial pool).



| Initial Distribution | Average r² of log-normal fit | Average r² of scale-free fit | Average r² of exponential fit |
|---|---|---|---|
| Log-normal | 0.942 | 0.762 | 0.929 |
| Scale-free | 0.967* | 0.967* | 0.795 |
| Exponential | 0.979 | 0.818 | 0.995 |

**Figure 3.10. Testing back-prediction with different simulated fitness distributions**
Test of fitness distribution inference in the presence of stochastic noise. Simulations began with the candidate initial distributions fit to the HTS data. Candidate initial distributions were log-normal (**A**), scale-free (**B**), and exponential (**C**) (in scale-free distribution, the scale-free and log-normal fits are overlapping). Black line: original distribution (as fit to estimated R0 data). Dashed line: distribution (not renormalized) after four rounds of enrichment + random noise. Gray line: R0 distribution retrospectively inferred from simulation of R4 + noise. Dotted lines are curve fits of the inferred R0 distribution (red line: log-normal; green line: scale-free; blue line: exponential). **(D)** Goodness of fit for different distributions fitting the inferred R0 distributions from simulations with added noise. Each $r^2$ value is the average of 10 trials. Overall, retrospective inference recovers the original distribution of $F_e$. Since high $r^2$ values can be found for different distributions, the pattern of residuals should be taken into account to determine the best fitting distribution to approximate the underlying curve.

### 3.7. Conclusion: A log-normal distribution of catalytic activity over random RNA space

Despite experimental and theoretical interest, there is little consensus in the literature on the nature and shape of any such distribution for any ribozyme function, or for that matter any chemical function evolved de novo from a biopolymer sequence space. Here, in one particular activity and sequence space, we found that the inferred distribution of ribozyme rates in a random pool fit well to a log-normal distribution (Figure 3.11A). A log-normal distribution of catalytic constants k for ribozymes across sequence space could indicate a normal distribution of the corresponding activation energies. Although the observed distribution cannot be quantitatively translated into activation energies without knowledge of the Arrhenius pre-exponential factor (which we here call *Arr* to avoid confusion), we note that the probability density drops precipitously as rate increases, such that high-fitness ribozymes occur in the population as extremely rare events. A normal distribution of log(k) would then imply a normal distribution of $E_a/RT + \ln(Arr)$. If we posit that *Arr* is likely to be similar for most ribozymes of similar chemical function, the standard deviation of ln(k) should equal the standard deviation of $E_a/RT$, such that our calculated standard deviation $\sigma_{\log(k)} = 0.665$ corresponds to activation energy deviation $\sigma_{Ea} = 1.6$ kJ/mol.

This distribution of activation energies is surprisingly steep; under such a distribution, a ribozyme cluster with 100-fold higher activity than the mean would occur only once in a pool of $10^{11}$ sequences. The steep drop-off of the distribution can be put in terms of the expected activity of the best ribozyme in a pool of a given size (Figure 3.11B), which shows that even very large increases in pool complexity result in relatively small gains in activity. Interestingly, increasing the size of the initial selection pool under such a relation would provide diminishing returns; knowing the size of all of sequence space, we can hypothesize an upper limit for the most active possible ribozyme. While various possible initial distributions of ribozyme activity have been proposed, the fit of rate constants to a single log-normal distribution (and the fit of activation energies to a normal distribution) suggest that the ribozymes, despite the large heterogeneity of sequence, share an underlying pattern for the emergence of function, a pattern that may only become obvious when considering hundreds of unrelated catalytic motifs. One possible interpretation is that a normal distribution of activation energies reflects the energetic contributions of many independent interactions with finite variance, with the ribozymes each using a similar number of interactions. The apparent independence of small contributions could be tested by combining mutations.[105] HTS could be used to expand and systematize such an approach.

A few caveats should be mentioned regarding the use of HTS data to infer the distribution of rate constants. First, while sequence clusters exhibited behavior that was consistent with evolutionary theory, individual sequences were less well-behaved. The effect of this limitation can be seen in the somewhat imperfect correlation between kinetics estimated from HTS and kinetics determined biochemically for individual sequences. Additionally, any conversion of fitness to rate constants relies on assumptions about the kinetic model. In this case, the assumption of constant $A$ necessary for estimating an initial activity distribution is known to be a simplification. $A$ is presumably influenced by RNA folding and activity as a ligase substrate, in addition to the triphosphorylation reaction. While it may be justified as discussed above, based on the ribozymes observable in later rounds, it is possible that the statistical properties of $A$ differ for lower activity ribozymes. In that case, it would not be possible to disentangle the effect of $k$ and $A$ on fitness, although $F$ would still have biochemical meaning as the fraction reacted.

With respect to the distribution itself, it should be noted that there is presumably an upper limit to activity that truncates any probability distribution function in reality. That is, increasingly precise arrangements of nucleotides at the active site, and correspondingly higher catalytic rates, are presumably limited by the RNA's steric mobility around a catalytic site. Structural and/or chemical limits in RNA would provide an upper limit for the possible catalytic rate enhancement, affecting the distribution at very high activity; these limits may or may not be close to the estimated upper activity limit suggested by the overall shape of the initial distribution.

Finally, the surprisingly steep drop-off in the frequency of high-activity ribozymes, and the accompanying flatness of the expected maximum *k* vs. complexity curve (Figure 3.11B), suggests that the ribozyme activity level in this case is largely determined by the nature of the function, not the complexity or diversity of the library. For some functions, a relatively low-complexity pool of RNA may thus possess ribozymes of biochemically significant activity, suggesting that the emergence of some catalytic functions may be substantially easier than of others, making high-throughput estimation of activity distributions a useful tool for identifying the most evolvable chemical functions in both RNA world studies and directed evolution for other *de novo* functions.

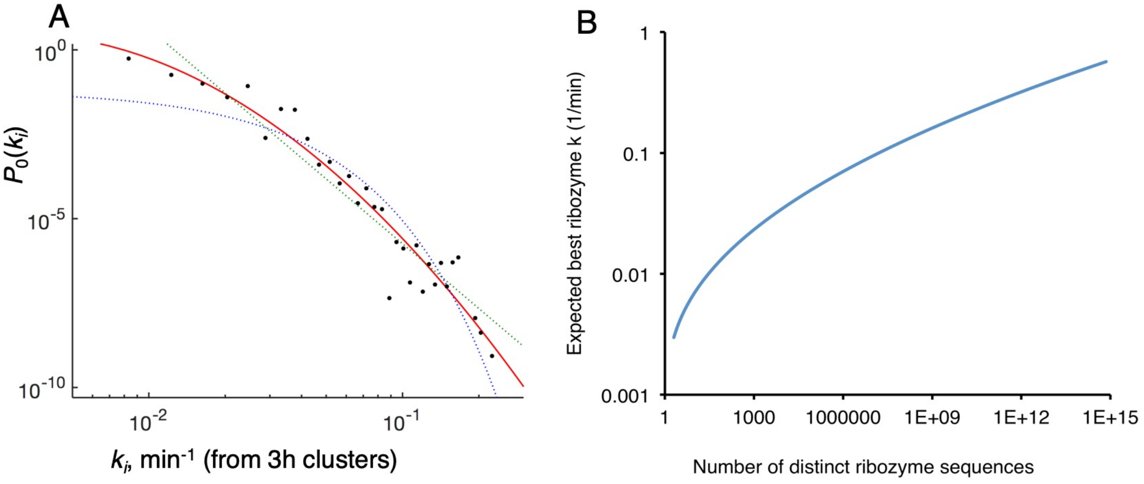


**Figure 3.11. An initial fitness distribution for TMPase ribozyme activity**
(**A**) The high-activity (right-sided) tail of the initial distribution of rate constants (k) is fit by a log-normal distribution (Red line: log-normal distribution $\mathrm{P_0} = \frac{1}{\sigma k\sqrt{2\pi}}\exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right]$ with $\sigma = 0.665$, $\mu = -5.375$, pdf scaling factor = 54.3, and nonlinear $r^2$ = 0.933; dotted green line corresponds to fit scale-free distribution $\mathrm{P_0} = \frac{6.38*[27.9]^{6.38}}{k^{7.38}}$, with pdf scaling factor = 0.00807 and nonlinear $r^2$ = 0.910; dotted blue line corresponds to fit exponential distribution $\mathrm{P_0} = 89.08e^{-89.08k}$, with pdf scaling factor = 1390 and nonlinear $r^2$ = 0.859. **(B)** Best ribozyme activity predicted by pool size, according to the inferred log-normal distribution of sequence activities. This curve is quite flat and suggests that large increases in complexity are needed for relatively modest gains in the activity of the best ribozyme.

# 4. Measuring ribozyme activity over an entire evolutionary space

Parts of this section were adapted from Pressman, A., Liu, Z., Janzen, E., Blanco, C., Muller, U.F., Joyce, G.F., Pascal, R, Chen, I.A., *J Am Chem Soc.* (not yet published at the time this was written). Reprinted with permission from ACS.

## 4.1. Background: Prior limitations of fitness landscape approaches

Section 3.1 describes the challenges and recent advances in estimating evolutionary fitness or direct kinetic activity for a large number of sequences within an evolving population. But if our goal is mapping out potential pathways of evolution for a de novo function or mechanistic change, it is not sufficient to measure the frequency with which catalysts of various strength are distributed across molecular space. We also want a picture of the individual peaks in fitness space, the valleys between them and paths by which they may be crossed, as well as the local topography surrounding evolutionary-optima and how that might affect evolvability. As described in section 1.4, fully understanding the evolutionary possibilities of a molecular space requires measuring the fitness of every possible molecule, creating a topographical "fitness landscape." And while various parameters can capture key features of a fitness landscape, an ideal approach would be able to simultaneously measure the chemical activity of every possible molecule capable of performing a given function.

To this end, existing research has taken one of two approaches. Molecular screening methods, described in section 1.4, directly measure the kinetics of every molecule in a population. While originally based in microarray techniques,[24,52-54] newer screening approach have used High-Throughput Sequencing (HTS) itself as a screening tool. In studies of self-cleaving RNA,[59,60] sequencing has proven an accurate measure of the fraction of each sequence that survives a selection step; studies in aptamers have gone further, using HTS to characterize the fraction of each unique sequence that survives a pulldown.[106] Such non-array screening methods can accurately characterize each sequence in a population, as long as each sequence is present at multiple count in the sequenced pools. However, with the limits of current DNA sequencing technology, HTS screening methods can only measure the activity of ~$10^6$-$10^7$ unique sequences, while array-based methods are even more limited in their depth. One solution to this has been choosing a sufficiently small scope: the work described above focuses on either extremely small sequence spaces of 10 varying nucleotides or less, as well as single-peak fitness landscapes covering only small mutations around a wild-type core sequence.

In contrast, one other approach to generating complete fitness landscapes, based on work in the Chen lab, has focused on the use of selection to cover an extremely large landscape. Jimenez et al.[43] demonstrated that a typical starting pool of ~$10^{14}$ sequences is sufficient to cover a sequence space with 20 random nucleotides at 100-fold (that is, an average of 100 copies of every possible sequence in the starting pool). Under this selection scheme, nearly every aptamer of sufficiently high affinity was expected to survive selection, and nearly every aptamer surviving selection was expected to have high affinity. 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. More generally, 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.[42,58]

Such "selection-heavy" coverage approaches have been able to identify numerous unique peaks corresponding to every likely active motifs across a sequence space, as well as many of the sequences in those peaks. But they have also been unable to accurately measure selection fitness or chemical activity of those peak sequences. As the goals of a fitness landscape approach include viewing the evolutionary tradeoffs along pathways between peaks, as well as the slope, roughness, and other topology of high-activity regions, a fitness landscape that only outlines peaks and potential pathways is of limited value. Instead of vague patterns of abundance or enrichment, we ideally seek a landscape in which the chemical activity of every possible sequence is directly measured. Thus, as described in the next section, the work in this chapter sought to combine selection and screening approaches, in order to map a fitness landscape both A) large enough to contain multiple peaks of unrelated catalytic mechanism, and B) with the same accurate activity measurement of a screening approach.

## 4.2. Theory: SCAPE, *k*-Seq and "virtual arrays"

As described in the previous section, there are fundamentally two main strategies to measure fitness landscapes: selection-based approaches, which can reduce a large and diverse pool of sequences down to a small and manageable number, and screening approaches, which can accurately measure many molecular activities in parallel but are limited by experimental design. To combine the two, we have developed a methodology that combines full-coverage selection with screening-based activity measurement to observe the kinetics of every high-activity ribozyme (or other functional molecule) capable of surviving selection. In the current work, we use 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 oxazolone-aminoacylation, an activity that would be foundational to protein translation, whose patterns of de novo emergence are of general interest to the study of how life evolved (Section 1.5). Biotinylated methyl tyrosine oxazolone (BTO) was chosen as a reaction substrate, due to its relative ease of manufacture and storage (vs. other activated amino acids) and synthesized by our collaborators, Robert Pascal and Ziwei Liu.

The SCAPE strategy begins with a population of molecules that covers nearly all possible sequences (here, $N$=21 nucleotides, for a starting library that includes ~80 copies of every possible RNA polymer sequence). 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 ribozyme activities, a pool from the selection that includes many different active sequences (~$10^8$, with ~$10^5$ appearing multiple times) is reacted at multiple substrate concentrations and products are isolated and sequenced on the Illumina platform. The sequencing output functions as a "virtual array"—a certain percentage of each sequence is self-modified and survives a purification step, thus giving a signal analogous to the fluorescent density of a microarray spot with ribozyme activity that reacts to a tagged substrate. We refer to this second step as kinetic sequencing (*k*-Seq), and in the aminoacylation selection it was used to measure the percentage of each ribozyme activated at four different substrate concentrations, in triplicate. While the *k*-Seq assay can only measure the activity of a small number of sequences (relative to all possible monomer patternings), we expect most high-activity sequences to survive selection, and most selection-end sequences to be of high activity. Thus, by running *in vitro* selection until the population is tractably small (5 rounds of selection here enough to reduce diversity from ~$10^{12}$ to ~$10^{8}$), the remaining sequences can be measured, allowing *k*-Seq to quantify reaction products and rates of potentially hundreds of thousands of sequences in parallel.

The *k*-Seq step is experimentally similar to previous rounds of selection; a representative round with manageable but sufficient diversity ($10^{5}$-$10^{6}$ expected unique sequences) is used for one final selection round. In this case, the selection is carried out many times in parallel, with different reactions varying the concentration of substrate (or the reaction time, or another selection parameter). The total number of recovered sequences is measured, by use of a spiked-in nonreactive sequence or other method; after sequencing each individual reacted pool, the reacted fraction can be calculated for every sequence present under every tested reaction concentration. In the SCAPE methodology, it is assumed that most high-activity sequences survive selection until the *k*-Seq round, and that most surviving sequences are of high activity. Thus, while SCAPE cannot test every single sequence present in a large sequence space, it can measure high-activity sequence (or nearly every one, accounting for stochastic losses) present in the starting population. If the sequence space is small enough to synthesize with high coverage, it then becomes possible to effectively measure biochemical activity across an entire fitness landscape of possible molecules.

In principle, the mathematics involved in *k*-Seq are fairly simple. For first-order ribozyme catalysis, data points were fit to equation (3.5), where each catalyst has the individual constants $A_i$ and $k_i$ representing overall stability and kinetic rate. For an *in vitro* enzyme selection, we would use the same math, with $A_i$ representing the rate of proper protein expression. In the case of an aptamer or *in vitro* antibody or peptide aptamer selection, we would instead use equation (3.6), etc., as described in section 3.1.

In applying the SCAPE method, our oxazolone aminoacylation selection was carried out on a library of 71-nucleotide RNA sequences, with a 21-nt randomized central region; this gave a total sequence space of 4x10$^{12}$ possible sequences, allowing the first round of selection to be carried out at 85-fold coverage. Six rounds of *in vitro* selection for aminoacylation activity were conducted, using a biotinylated tyrosine analog, biotinyl-Tyr(Me)-oxazolone (BTO), with aminoacylated RNA sequences recovered through the use of Streptavidin beads and replicated through reverse-transcriptase polymerase chain reaction (RT-PCR). 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 to confirm reproducibility of the selection. After the selection was completed, a set of "*k*-Seq rounds" were selected from the round 5 RNA pool, using four different substrate concentrations, each in triplicate, at similar sequencing depth. These *k*-Seq sequence abundances, normalized by a nonreactive spike-in sequence of known concentration, were used to generate activity curves for approximately 9x10$^6$ sequences, of which about 3x10$^5$ displayed increased activity over the baseline uncatalyzed aminoacylation rate.

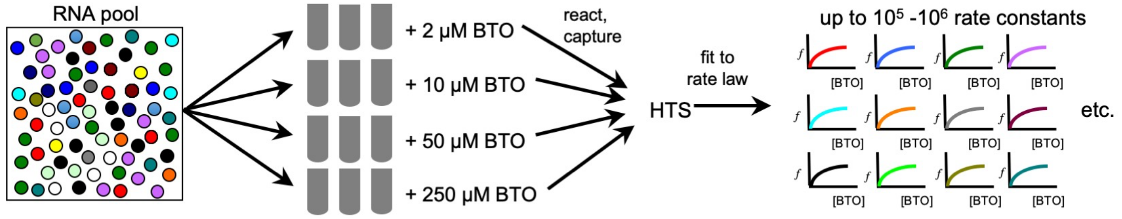


**Figure 4.1.** ***k*-Seq schematic**

In k-Seq, an RNA pool enriched for active ribozymes is reacted at multiple BTO concentrations, in triplicate. Captured RNA is then reverse-transcribed and sequenced. Activity curves are constructed for sequences detected in the enriched pool.

## 4.3. Oxazolone aminoacylation selection results

Overall, numerous related families of active ribozymes were observed in the aminoacylation selection, with distinct repeated motifs appearing in Round 4 (Figure 4.2A,B). Three distinct motifs were identified from the top 20 sequence families, which comprised 80% of sequence reads by Round 6. Notably, each family showed one of these three motifs present at a slightly different location within the 21-nucleotide random region, suggesting conserved sets of active nucleotides that survived selection despite being located at different positions within sequence space. Every one of the top 20 families present in the first selection, RS1, were also present in a duplicate selection, RS2, with similar rates of enrichment (Figure 4.2C). Of these three motifs, Motif 1 contained the shortest conserved region (Figure 4.2D) and encompassed the greatest number of unique sequences. Motif 1 could be further categorized into three sub-motifs (1A, 1B, 1C) by differences in the conserved region, with 14 of the top 20 families belonging to Motifs 1A and 1B. Motif 2 contained fewer unique sequences than Motif 1, but more than Motifs 1A, 1B, or 1C. Motif 2 also included Family 2.1, the most abundant family of the pool. Motif 3 comprised the smallest fraction of the pool and encompassed the fewest unique sequences.

*k*-Seq estimates for activity could be obtained for $8.9\text{x}10^6$ sequences, but the majority of sequences were present at low count and correspond to low activity (Figure 4.3). $\sim10^5$ unique sequences were found to have activity >10-fold above the non-catalytic background rate (i.e., catalytic ratio $r_i > 10$, where $r_i = k_iA_i/k_0A_0$, with $k_0$ and $A_0$ being reaction parameters for uncatalyzed aminoacylation of random RNA). To determine how well *k*-Seq results corresponded with results of the standard assay, we chose ten sequences that are close to the consensus sequences of the high- or medium-activity families (Appendix Table A.3, with all five motifs and sub-motifs represented) and measured aminoacylation activity by a standard Streptavidin gel-shift assay after reaction with the biotinylated substrate (explained further in Appendix Figure A.5). In the *k*-Seq assay, we reacted a heterogeneous pool from *in vitro* selection, containing many different RNA sequences, with BTO and pulled down the aminoacylated RNAs with streptavidin beads. These sequences were expected to span approximately one order of magnitude in their overall catalytic rates. Rate constants determined from *k*-Seq matched well with gel-shift measurements (Figure 4.4A,B, Appendix Table A.3). All *k*-Seq and gel-shift measurements were performed in triplicate and the standard error was similar between *k*-Seq and gel-shift measurements (Figure 4.5A). Measurement error during *k*-Seq decreased as sequence read abundance increased, becoming substantial for sequence read abundances $<10^{-6}$, as expected for stochastic noise (Figure 4.5B).

Notably, the noise in sequence estimation was lower for most sequences than in the case described in section 3.5, suggesting *k*-Seq as a more robust activity estimation method than our best estimates from multiple-round selection data. However, the relative scarcity of high-activity, measured sequences in all but the final round of selection prevented the use of a detailed Fisher's Theorem analysis as described in the same section.

The best ribozyme found here has a rate constant comparable to that of ribozymes using a biologically derived aminoacyl adenylate,[40,107] indicating that these reactions could proceed efficiently even with only prebiotic substrates.  High-activity sequences (e.g., the center of Family 2.1, with $r_{S\text{-}2.1\text{-}a} = 1010$ and $k_{S\text{-}2.1\text{-}a} = 779 \pm 21$ min$^{-1}$M$^{-1}$) exhibit saturating kinetics from *k*-Seq, providing both the rate constant ($k_i$) and the maximum amplitude of reaction ($A_i$). However, the reaction for lower activity sequences (approximately $k_i < 20$ min$^{-1}$M$^{-1}$) appears linear under the conditions tested, so that $k_i$ and $A_i$ are difficult to estimate separately using these data; instead the combined parameter $k_iA_i$ can be estimated (Figure 4.5D).

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 Methods for sequence nomenclature) for characterization of the reactive site. Identification of the site was performed by another member of the Chen group, and is described further in Appendix A.2. 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 mechanisms.

The log-normal distribution shape for catalytic activity $k_iA_i$, is consistent with prior findings.[58,108] Since the rate constant scales exponentially with the activation energy, it was of interest to determine the distribution of $k_i$ alone. For the highest activity family (2.1), many ribozymes could be characterized by $k_i$ and $A_i$ separately. $k_i$ was observed to fit a log-normal distribution, indicating that activation energies are normally distributed for a ribozyme family (Figure 4.5E,F). The distribution of $A_i$, which represents the maximum extent of reaction and may indicate the fraction of RNA that is well-folded, also fit a log-normal distribution well, suggesting that folding energies may also be normally distributed for a ribozyme family. For the regime in which $k_i$ and $A_i$ could be determined separately, these parameters are not well-correlated with each other, suggesting no relationship between the catalytic rate and fraction folded.

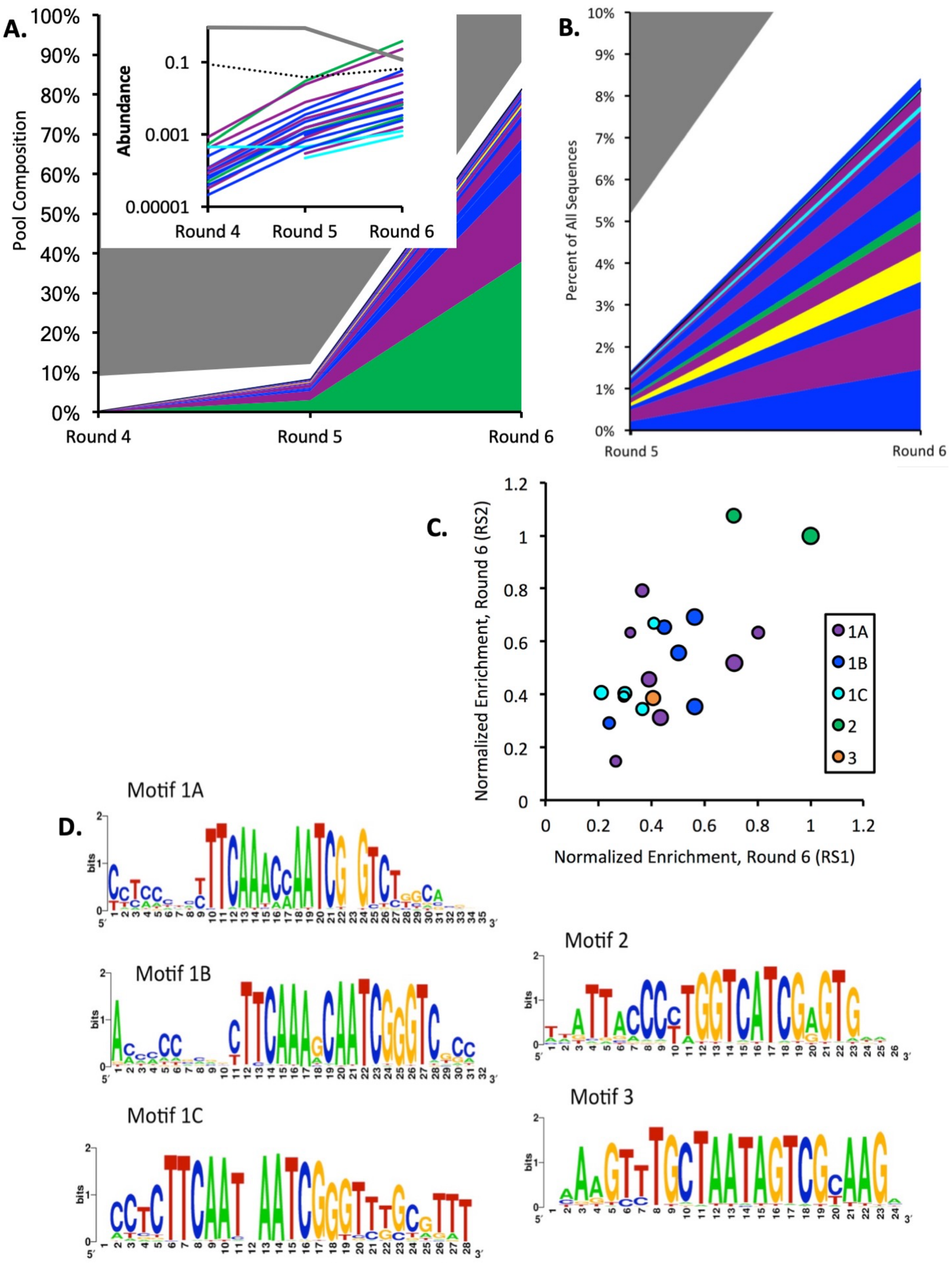


**Figure 4.2. Results of aminoacylase selection**
**(A)** Pool composition over Rounds 4-6 after clustering. The top 20 families are indicated in non-neutral colors; gray corresponds to un-clustered sequences; white corresponds to families with rank by abundance >20.

Families from sub-motif 1A (purple), 1B (dark blue), and 1C (cyan), and 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). **(B)** Magnified version showing families ranked 6-20. **(C)** Plot demonstrating concordance of top 20 families discovered in two selection branches. Dot size corresponds to family rank in RS1 (see Methods). Enrichment ratio across rounds shows some correlation. **(D)** SeqLogo representations of motifs.

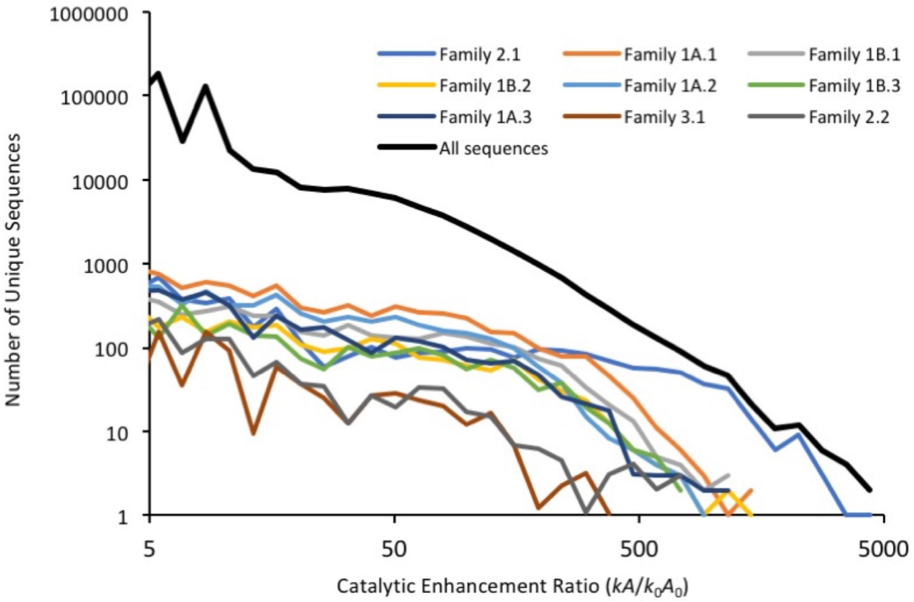


**Figure 4.3. Initial activity distribution for aminoacylation ribozyme space and families**
Distributions for the 9 highest activity families are plotted individually. The overall curve, including all families (black), is driven at higher activities by contributions from the families shown, which make up a significant portion of high-activity sequences.
of unity.

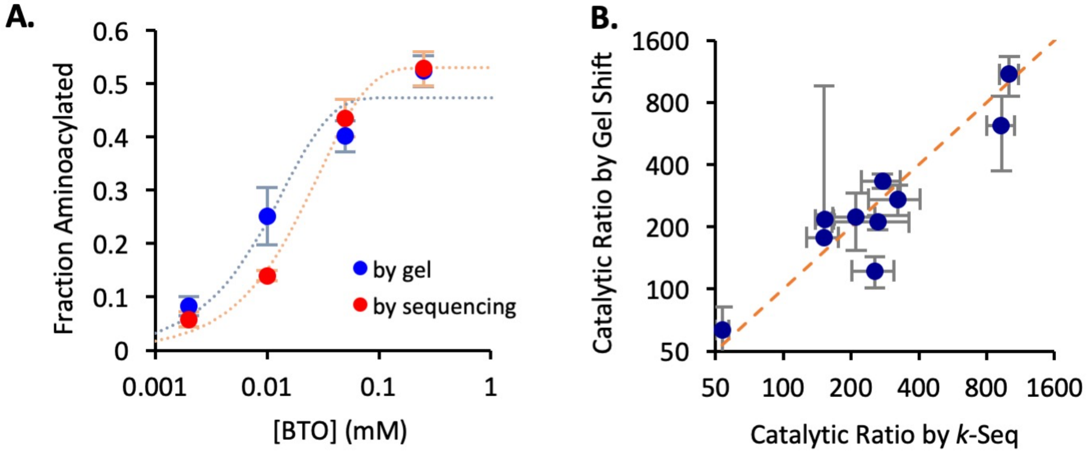


**Figure 4.4. Comparing *k*-Seq to experimental activity measurements**
**(A)** Aminoacylation at various [BTO] for ribozyme S-2.1-a observed by both gel shift (kA = 706 ± 117 $min^{-1}M^{-1}$) and k-Seq (kA = 652 ± 3 $min^{-1}$ $M^{-1}$). Data for all other measured ribozymes are shown in Supporting Figure 3.

Error bars correspond to standard deviation among triplicates. **(B)** Correlation between catalytic ratio of ten ribozymes, measured by gel shift assay and k-Seq. Error bars correspond to standard deviation among triplicates. ($R^2 = 0.87$; catalytic ratio calculated from values given in Appendix Table A.3). Dotted orange line indicates line of unity.

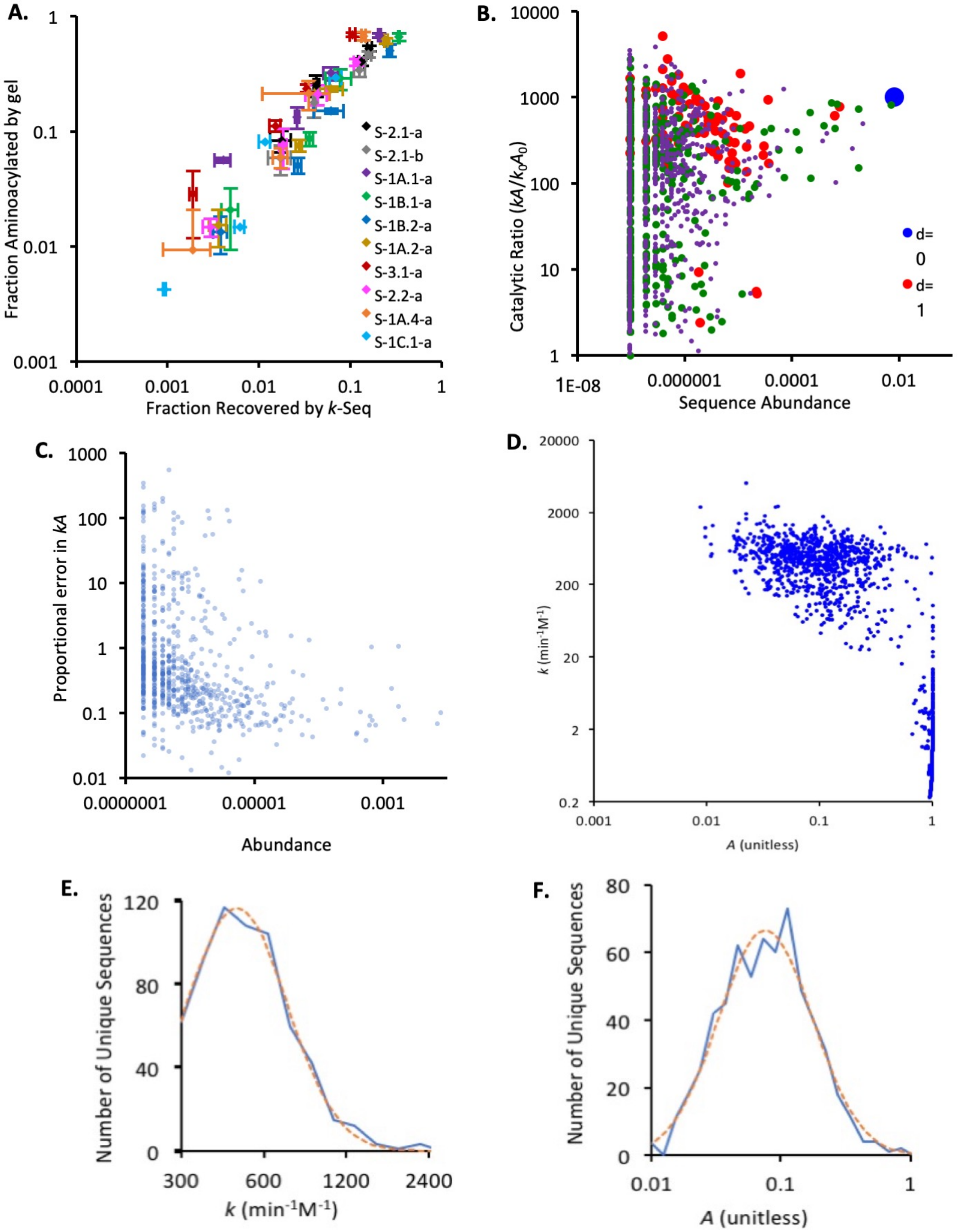


**Figure 4.5. The distribution of noise in *k*-Seq measurements**

**(A)** All data points from gel-shift assays (10 sequences, 4 concentrations of BTO, 3 replicates), compared to k-Seq measurements of sequence recovery for the same sequences. Error bars are standard deviation of triplicates. **(B)** Catalytic rate enhancement (*k*-Seq) vs Round 6 abundance for all sequences in family 2.1. Sequences sorted by distance from peak center (d = 0,1,2,3, colored as blue, red, green, and purple, respectively). No correlation observed between abundance and catalytic ratio. **(C)** For all sequences in family 2.1, proportional standard error is higher at very low abundance than at higher abundance. Most sequences having abundance $>10^{-6}$, and nearly all sequences having abundance $>10^{-5}$ have proportional error < 1. These values may be consistent with stochastic noise due to stochastic sampling of sequencing reads. **(D)** k vs A for all sequences in Family 2.1. For sequences of low k, a fit with A=1 is found; k A is still expected to be accurately estimated, but parameters cannot be estimated separately with accuracy as the curve appears linear in this range. For sequences of higher $k_i$, there appears to be little correlation between $k_i$ and $A_i$. The distributions of k **(E)** and A **(F)** fit well to log-normal distributions (dotted red line, nonlinear $R^2$=0.99 for k, and $R^2$=0.97 for A).

## 4.4. A large, quantitative fitness landscape and its evolutionary pathways

Overall, the results described in the previous section suggest several key features of the aminoacylation activity landscape constructed through SCAPE: 1) Our landscape likely identified all major activity peaks in the 21-varying-nucleotide sequence space, measuring at least the highest-activity catalysts with decent accuracy; and 2) Conserved regions appear small and varied enough, and the random region large enough, that our landscape contains several different catalytic mechanisms for achieving the same chemical function. This is significant, as prior measurements of fitness landscapes achieve only one of these points (section 4.1); we are thus able to observe multiple different evolutionary paths to the same biochemical outcome, with substantial resolution and kinetic data around each unique reactive mechanism.

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 (mentioned also in section 1.5) 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 4.6, Appendix Table A.4). A broad network of pathways existed 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 two top 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 five consecutive intermediates expected to have only baseline activity (i.e., $r \sim 10^3$-fold less than $r_{S\text{-}2.1\text{-}a}$ and with aminoacylation likely to occur at random sites or the 3'-OH end). The global optimum (Family 2.1) is particularly 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.

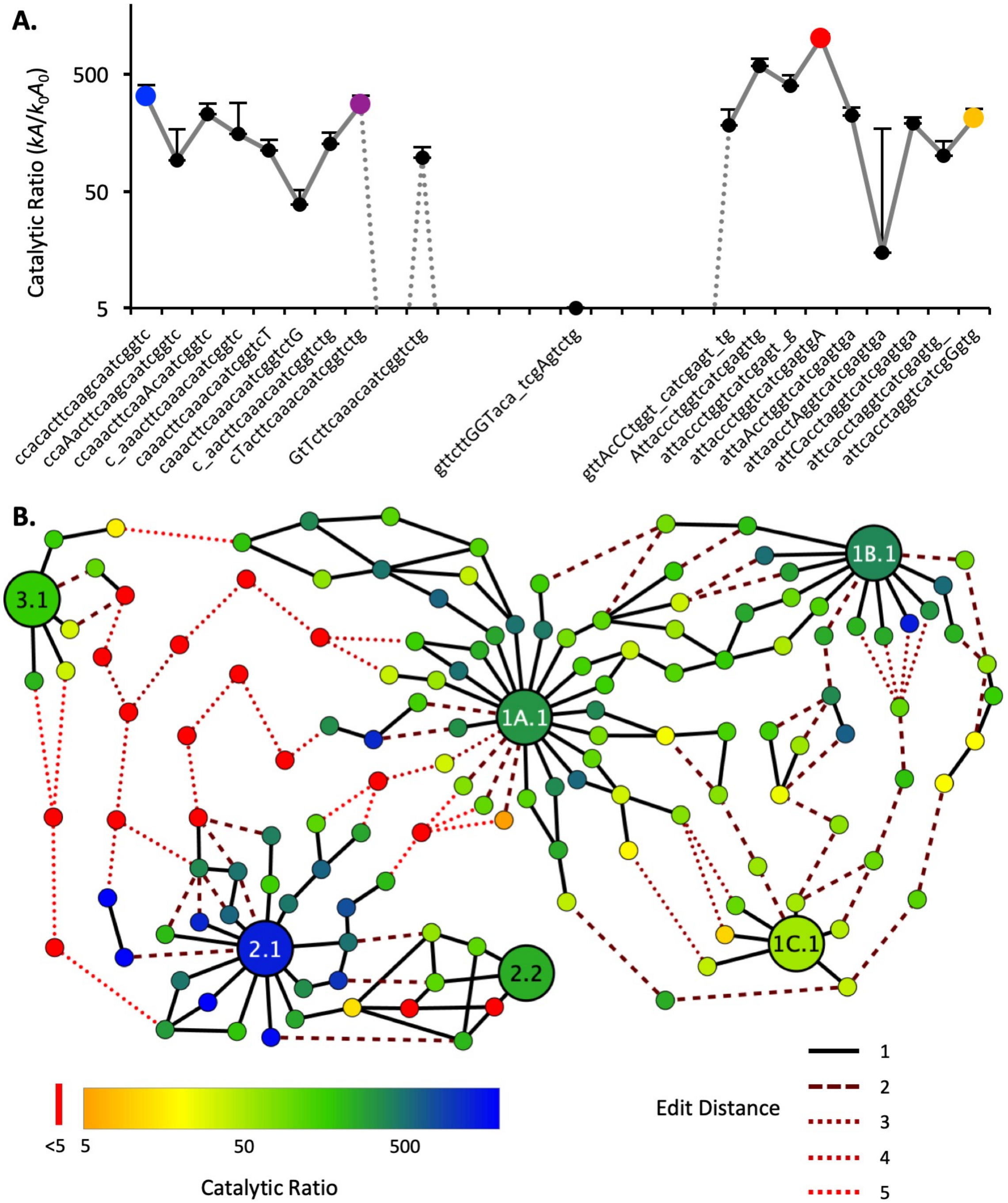


**Figure 4.6. Evolutionary pathways for oxazolone aminoacylase ribozymes**

**(A)** Catalytic ratio along the best pathway discovered from the center of Family 1B.1 (blue, S-1B.1-a), to 1A.1 (purple, S-1A.1-a), to 2.1 (red, S-2.1-a), to 2.2 (orange, S-2.2-a). Capital letters denote sequence positions changing at each step; underscore indicates a deletion. Note the large drop in activity required to cross between Motif 1 and Motif 2. Error bars are standard deviation from triplicate measurements. **(B)** Evolutionary network displaying the best 10 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 sub-smotif, as well as the two most active centers from motif 2. Each node is an individual sequence with activity measured by *k*-Seq as indicated by color (see legend; red indicates activity equal to or below the baseline uncatalyzed rate). The strength of the lines indicates mutational distance between sequences; intermediate sequences along the dotted lines have baseline activity. The majority (67%) of the edits along these pathways are substitutions; the remainder are insertions/deletions.

## 4.5. Conclusion: SCAPE reveals a highly frustrated evolutionary network

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 ~$10^5$ unique sequences (a number that in general depends on pool diversity, activity distribution, and sequencing depth). Using SCAPE, we mapped a first comprehensive fitness landscape for catalytic activity. We discovered ribozymes that self-aminoacylate using a 5(4H)-oxazolone, a key step toward the genetic code. The best ribozyme found here has rate constant comparable to that of ribozymes using a biologically derived aminoacyl adenylate,[40,107] indicating that these reactions could proceed efficiently even with only prebiotic substrates.

Fundamentally, the motivation for SCAPE analysis is to learn about molecular evolution by measuring all viable evolutionary pathways and networks through sequence space, providing a complete map for how a given biochemical function can evolve (or be evolved) into or between different catalytic peaks of high activity. We found that, 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. In general, valleys appear to occur when mutations that are separately beneficial or neutral become detrimental together.[109-112] The likely reason is that the three major motifs differ in mechanism and structure, as indicated by their different aminoacylation sites. It appears that the mechanism cannot be changed without 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 described for functional RNA. Specifically, it addresses the hypothesis of neutral networks (a proposal discussed in section 1.5). While in general theoretical studies have suggested the possibility of comprehensive, evolvable networks between diverse peaks in fitness space for early life-analog ribozymes,[94,113,114] such networks have been difficult to find in practice.[115] One possible intuitive interpretation of these aggregated results is that evolutionary pathways for ribozymes may be viable when structural and/or mechanistic features can be conserved, but that pathways toward a new fold or motif are usually not viable. A second interpretation is that the appearance of "neutral" networks may to some extent be an artifact of selection processes; while we were indeed able to find "pathways" of sequences connecting all major peaks in our landscape, real catalytic measurement of these intermediate sequences revealed them to have little to no activity, likely simply the result of a few random sequences surviving the initial rounds of selection. An important caveat is that the landscape has been described under constant selection for a particular catalytic activity; changing environments or selection pressures may significantly alter this picture, but this is something that may be testable by future fitness landscape studies varying additional environmental conditions under *k*-Seq.

The evolutionary frustration inherent to the ribozyme activity landscape is analogous to frustration found in other physical systems. A classic illustration of frustration is the anti-ferromagnetic spin glass, in which energy would be minimized by antiparallel placement of neighboring electronic spins.[41,116,117] In certain configurations (e.g., a triangle), no placement of spins can satisfy all desired constraints, leading to rugged energy landscapes. Walks on such energy or evolutionary landscapes are characterized by sensitivity to initial conditions, frustrated optimization, and multiple possible outcomes. Thus, in the absence of recombination or other mechanisms to increase diversification,[112] the emergence of a globally optimal sequence is likely to result from chance events rather than natural selection, as shallower local optima become effective fitness traps. This is a significant result for the emergence of one particular ribozyme function in the origin of life (or in *de novo* selection for an engineered aminoacylation system), but the extent to which these results can be expanded to other examples of evolution relies on application of the SCAPE methodology to additional evolutionary systems.

# 5. Limitations of the *k*-Seq assay

## 5.1. The curious *k*-Seq conundrum for a TMP phosphorylase

As a follow-up to the successful generation of an oxazolone aminoacylation landscape, we decided to attempt a similar analysis on the TMP triphosphorylation ribozyme selection discussed in Chapter 3. This experiment did not fall under the typical SCAPE methodology, as the pool of ribozymes had been selected from an infinitesimally small partial coverage of sequence space, with mutagenic PCR generating sequence-variant fitness peaks around each high-activity sequence present in the initial pool. However, the presence of > 300 apparently-unrelated active catalytic motifs was determined to be a suitable subject of interest.

Here, the actual *k*-Seq step was performed by the Müller lab, which had previously carried out the triphosphorylation selection; we carried out sequencing and analysis of the data similarly to the aminoacylation case, but the *k*-Seq rounds were carried out under a different protocol. Specifically, our collaborators suggested varying reaction time instead of substrate concentration so that all reactions could be taken out as quenched aliquots from a single reaction. The presence in the selection of a ligation step whose efficiency varied by sequence also reduced the ability of a spike-in sequence to accurately quantitate recovery through normalization, such that kinetic fitting could be assumed to accurately determine $k_i$ but not $A_i$ for each sequence. Additional complicating steps in the selection meant that only a single replicate was carried out, instead of the parallel triplicates of each sample used in the aminoacylation experiment. Overall, while the aminoacylation selection had been designed from the start to fit the criterion for an effective SCAPE experiment, the triphosphorylation selection contained additional complicating steps, and was chosen as a test of *k*-Seq in a non-ideal, more chaotic setting.

Unfortunately, the results of this *k*-Seq selection proved inconclusive. For a number of sequences synthesized and tested, chemically-measured and *k*-Seq-predicted catalytic activity showed almost no correlation (Figure 5.2A). After some consideration and further analysis, we settled on several hypotheses for why this approach was not nearly as accurate as our previous use of *k*-Seq:

1) Fundamentally, the triphosphorylation selection had been an extremely sensitive one; larger variations were observed in round-to-round sequence enrichment in this selection than in the aminoacylation selection (compare Figures 3.6, 4.5C). In fitting kinetic curves to *k*-Seq data points, the data themselves were slightly less “curve-shaped” in the triphosphorylation selection than in the aminoacylation selection, resulting in the latter giving kinetic fits with lower correlation (Figure 5.2C). This may be a relatively isolated problem, as the triphosphorylation was especially finicky and sensitive, with its activity varying substantially more between measurements. It may also be the result of an inability to accurately fit only $k_i$ and not $A_ik_i$, as we saw in the aminoacylation case that *k*-Seq can fit $A_ik_i$ substantially more accurately than either single parameter for lower-activity sequences.

2) The data inconsistencies could be an artifact of sequence choice. While aminoacylation ribozymes were chosen as the centers of large sequence families, the triphosphorylation selection contained far more diversity of active motifs. Thus, sequences were chosen based on their activity (as an expected 100%, 90%, 80% etc. of the highest-activity sequence) rather than the specifics of their sequence. If we expect a mean random variance in sequence activity of ±30%, with a large population of sequences, we could conceivably see at least a handful with observed activity over 2x greater than the highest actual activity; thus, the highest-activity sequence (and all sequences with at least half its

activity) might very well be all or mostly casualties of stochastic effects or measurement noise, as described in sections 3.5 and 4.3. Over the range of activities observed and tested, this might be capable of providing enough chaos to obfuscate any correlation between *k*-Seq and chemically-tested activity measurements.

3) The addition of many rounds of mutagenized PCR to the triphosphorylase selection left most sequences at extremely low copy number in the selected pool, and most of those chosen were of low abundance. Ongoing theoretical work by another member of the Chen group has suggested that *k*-Seq becomes inherently unreliable once sequence abundance falls below a certain threshold in an idealized case. While the triphosphorylase ribozymes chosen had abundance above this threshold, the actuality might require a slightly larger abundance cap for reliable prediction than the theoretical simulations suggest.

Unfortunately, lack of a *k*-Seq replicate makes it impossible to determine which combination of these factors may have confounded the triphosphorylation attempt. As at least one possibility is inherent randomness in the reaction itself, it was decided that rather than spend additional time and resources on redoing the experiments, future focus should be on specific methods to improve the accuracy of existing *k*-Seq data, as described in the next section.

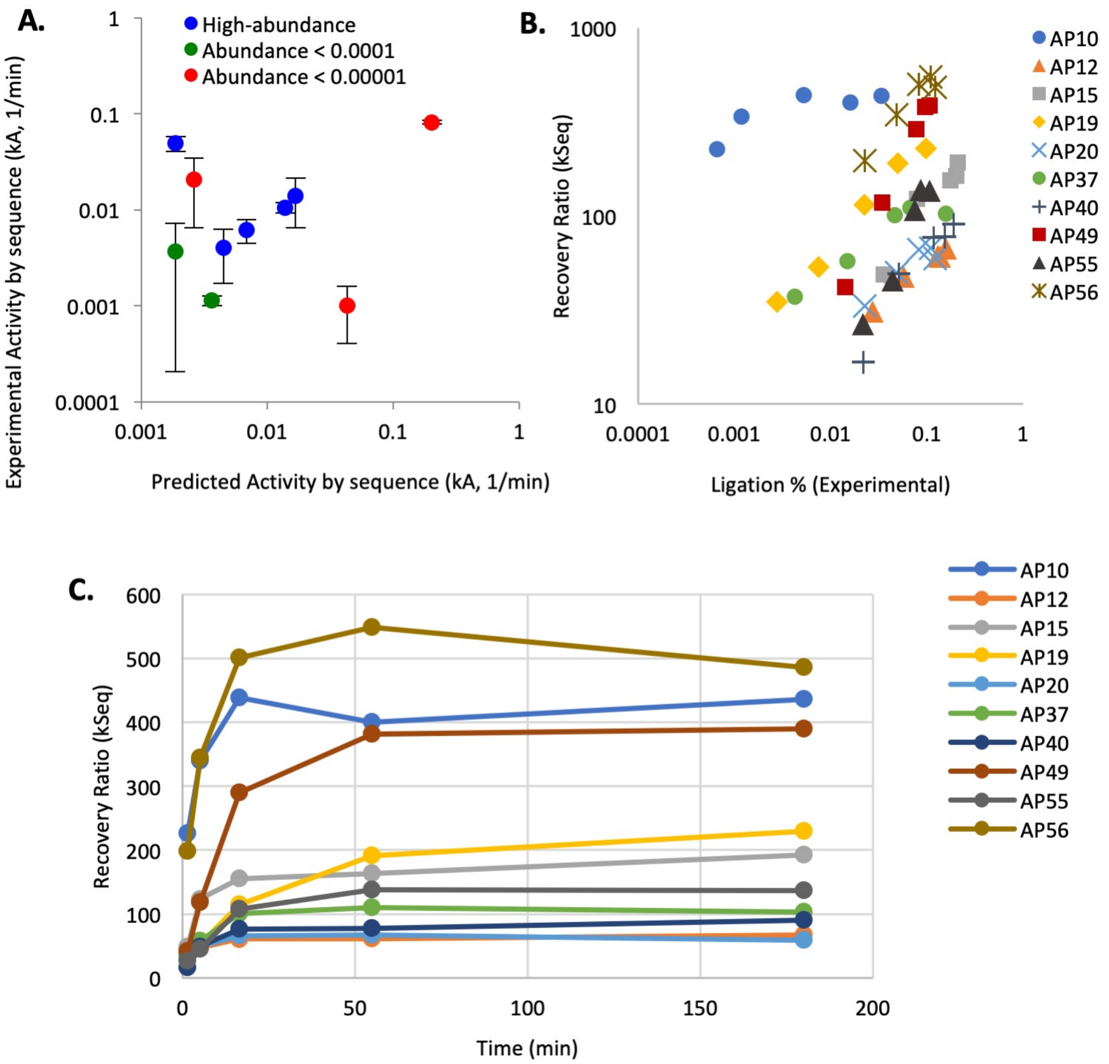


**Figure 5.1. Kimitations of *k*-Seq analysis applied to TMP ribozyme selection**
**(A)** *k*-Seq estimated vs. experimentally-determined activity for 10 new sequences chosen specifically for *k*-Seq analysis of the TMP triphosphorylase ribozyme system. Sequences were chosen to correspond to a range of expected activity. Higher-abundance sequences showed some correlation HTS-predicted and experimentally-measured catalytic ratio; taken as a whole set, very little correlation is observed. **(B)** The individual ligation rates, measured both experimentally and by *k*-Seq, for each incubation time point for all 10 new sequences. Compared to figure 4.5A, we observe significantly less correlation between the two values under individual measurement conditions. **(C)** Reaction vs. time curves for investigated sequences show some irregularity.

## 5.2. Improving *k*-Seq accuracy through library manipulation

The presence of a known non-reacting RNA spike-in sequence in both k-Seq tests led, among other things, to an interesting dilemma: constructing a “peak” around this dummy sequence led to a number of single-mutant variants, making up about 3% as many sequences as the correct version in the aminoacylation case and about 20% as many sequences in the oxazolone case. Since each spike-in sequence had been ordered as a single DNA sequence, this implied the introduction of mutational variation either through transcription, reverse transcription, PCR, or sequencing itself (with the longer triphosphorylation spike-in acquiring more errors). A close reading of enzyme manufacturers' protocols suggests that T7 polymerase had the highest error rate of any of these steps, and was thus likely introducing unintentional slight mutation rates at every round of selection.

In the aminoacylation selection, miscopying 3% of every sequence's random region gives about 0.5 copies of each *specific* mutant for any copied sequence. While relatively small, this still functions as a limiter on the accuracy of certain points in our SCAPE-generated landscape. Due likely to stochastic effects in the early rounds of selection, the highest-count sequences in each of the top 3 aminoacylation peaks appear at greater than 2000x the abundance of the lowest-count sequences, which gives T7 polymerase errors the potential to slightly bias measurement of low-count sequence activity. The effects of such bias are likely to be relatively small, and likely to impact mostly the sequences closest to family centers, leaving most of the shape of our predicted fitness landscape (and the pathways between peaks) intact. It does, however, limit our confidence in measurements of fine peak topography such as epistasis and certain roughness parameters (discussed in the next section), whose analysis relies specifically on the variation in activity between high-abundance sequences and their closest neighbors.

A number of methods already exist to address these concerns, and it remains to be seen whether or not such remediation even changes epistasis calculations. Techniques for normalizing libraries of highly-varied sequence abundance have existed for decades,[118,119] and represent one possible step to add to the SCAPE protocol if necessary. A simpler option, and one we are currently testing, involves synthesizing new libraries comprising the local sequence space around all major peaks. While this method may sacrifice some coverage of fitness space, the initial SCAPE approach is still able to pinpoint all possible active motifs, and a more detailed follow-up may be able to determine whether the local topography of disparate fitness peaks is homo- or heterogeneous in different regions of sequence space. The next chapter describes how this approach may be able to answer additional open questions about evolutionary spaces and how to map them.

# 6. Potential high-definition landscape topography

The work in this chapter primarily corresponds to my contribution to an in-progress publication, which is awaiting one final set of experimental data (*k*-Seq over a library corresponding only to local peak sequences) to confirm or contradict the findings briefly proposed here.

## 6.1. Background: Epistasis in real fitness landscapes

While evolutionary pathways determine global evolvability of functional molecules across a fitness landscape, local evolvability is generally governed by subtler local topography. The ease with which a sequence can reach a local maximum in activity depends on its epistasis (as introduced in Section 1.3), which is the dependence of fitness effects on multiple sites. In an evolutionary system with little epistasis, swapping out one monomer at a time, and choosing the replacements that increase activity, would be sufficient to evolve the best possible catalyst or reagent sequence for a given task. In a system with significant epistatic effects, evolving the highest-activity sequences would require a significant mutation rate and many rounds of selection.

Theoretically, a system with complete or near coverage of sequence space—such as our SCAPE landscape—benefits less from epistatic analysis, as no mutation or additional evolution is required to identify the most active sequences. But epistasis distributions and roughness measurements provide a simple means to generalize evolvability of a landscape, which helps us to understand how a given function *could* evolve in a history-of-live setting, or how a similar but unstudied function could be engineered to evolve. And epistasis is already widely studied in lower-coverage fitness landscapes, especially those of natural proteins, where only a small local band of fitness space can be studied. Analysis of epistasis on *in vivo* protein landscapes is more common than on *in vitro* systems, but generally limited to a small number of peptide sites, a limited library of amino acid substitutions, or one specific set of evolutionary paths.[27] 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,[120] 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.[57] 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. We thus chose our aminoacylation landscape as a subject for epistatic analysis to both A) investigate SCAPE's ability to provide high-quality epistasis data that might be useful in future comparison studies and B) observe whether patterns of epistasis in a synthetic ribozyme mirror or differ from those in naturally-occurring protein-based catalysts.

One may also ask whether real fitness landscapes can be generally fit to rough empirical fitness landscapes models. One 2013 meta-analysis found general trends in ruggedness and epistasis across a number of such studies, with many showing reasonable agreement[41] with patterns expected from the Rough Mt. Fuji model, in which a perfectly-smooth fitness landscape is overlaid with an uncorrelated pattern of random roughness.[121,122] Efforts to connect empirical data to these models are important for gaining an intuitive grasp of the topography of fitness landscapes. It currently remains an open question whether any empirical fitness landscape models can also describe effects fitness landscapes over a larger scale, or ones that contain multiple unrelated high-activity peaks.

### 6.2. Theory: Topography and ruggedness of fitness landscapes

Overall, the dominant types of epistasis give a sense of the sort of “curvature” of a fitness peak in evolutionary space. Here, the terminology is somewhat confusing, and can depend on the sequence chosen as a baseline or “wild-type” (wt) for investigating epistatic effects (Table 5.1). While multiple definitions of epistasis exist, for purposes of studying epistasis on our aminoacylation landscape we chose log-fitness, such that the epistasis of combined mutations a and b,

$$\varepsilon_{a,b} = \ln\left(\frac{r_{ab}}{r_{wt}}\right) - \ln\left(\frac{r_a}{r_{wt}}\right) - \ln\left(\frac{r_b}{r_{wt}}\right) \qquad (6.1)$$

where $r_{wt}$ is the catalytic enhancement ratio of a chosen peak center “wild-type,” and $r_a$ the enhancement of the baseline sequence modified by mutation a.

For simplicity, we choose the highest-activity sequence in a fitness peak as our baseline (such that all single monomer changes are deleterious). Then an abundance of positive magnitude epistasis, where the effects of two individual mutations are greater in combination, leads to a “dome-shaped” peak, where fitness drops off first gradually and then quickly with an increased number of sequence substitutions—in such a case, the peak may function as more of a small plateau of high local evolvability. Negative magnitude epistasis, where two individual mutations have less of an effect in combination, leads to a “needleshaped" peak, where fitness drops off sharply with any changes to the ideal sequence and then gradually with additional changes—such a peak might be harder to find through random evolutionary walks but easier to maintain through additional rounds of evolution once found. Sign epistasis, where two negative mutations in combination have a less negative effect than one individual mutation, leads to a peak surrounded by local smaller peaks or "saddle points"—in this case, changes to the ideal sequence might be offset by compensating mutations at other monomer sites.

Beyond basic, two-mutation combinations, other types of epistasis can be investigated at well. Higher-order epistasis looks at either the effects of more than two changes, or of the effects of two larger changes (i.e. the epistatic effects of two multiple-monomer substitutions). To evaluate interaction between a double mutation and a third mutation ($c$), we calculated

$$\varepsilon_{a,bc} = \ln\left(\frac{r_{abc}}{r_{wt}}\right) - \ln\left(\frac{r_a}{r_{wt}}\right) - \ln\left(\frac{r_{bc}}{r_{wt}}\right) \qquad (6.2)$$

with $\varepsilon_{b,ac}$ and $\varepsilon_{c,ab}$ defined and calculated in analogy, for all combinations of three mutants observed; this is referred to as "a-bc epistasis", in analogy to the "a-b epistasis" of double mutants. We also calculated a triple "a-b-c epistasis", defined as

$$\varepsilon_{a,b,c} = \ln\left(\frac{r_{abc}}{r_{wt}}\right) - \ln\left(\frac{r_a}{r_{wt}}\right) - \ln\left(\frac{r_b}{r_{wt}}\right) - \ln\left(\frac{r_c}{r_{wt}}\right) \qquad (6.3)$$

although this could not be classified in direct analogy to the other epistatic observations.

Overall, the ratio of different epistatic effects can be appraised through various ruggedness parameters, which give general numerical descriptors of how "smooth" or "bumpy" a landscape is. For our purposes, the most useful is probably epistatic correlation

$\gamma_d$, further discussed in Appendix A.3, which is defined for each possible pair of mutations over a varying background of edit distance *d*. [123] $\gamma_1 = 1$ indicates a perfectly smooth landscape with no epistasis, $\gamma_1= 0$ indicates a completely uncorrelated landscape, and $\gamma_1 = -1$ indicates a maximally rugged, anticorrelated landscape dominated by reciprocal sign epistasis. The changes in $\gamma_d$ as *d* varies indicate how many key positions are involved in determining a sequence's sensitivity to variation. If $\gamma_d$ remains high, we expect the same positions in a key motif to be conserved even as the other positions vary, but if it drops off quickly as *d* increases we expect the key active monomers to change more fluidly across sequence space.

| Single mutants | Double mutant | Category |
|---|---|---|
|  | $r_{ab} = r_a r_b$ | no epistasis |
| $r_a > r_{wt}$ and $r_b > r_{wt}$ | $r_{ab} > r_a$ and $r_{ab} > r_b$ | magnitude epistasis |
| $r_a < r_{wt}$ and $r_b < r_{wt}$ | $r_{ab} < r_a$ and $r_{ab} < r_b$ | magnitude epistasis |
| $r_a > r_{wt} > r_b$ | $r_a > r_{ab} > r_b$ | magnitude epistasis |
| $r_a < r_{wt} < r_b$ | $r_a < r_{ab} < r_b$ | magnitude epistasis |
| $r_a > r_{wt}$ and $r_b > r_{wt}$ | $r_{ab} < r_a$ and $r_{ab} < r_b$ | reciprocal sign epistasis |
| $r_a < r_{wt}$ and $r_b < r_{wt}$ | $r_{ab} > r_a$ and $r_{ab} > r_b$ | reciprocal sign epistasis |
| $r_a > r_{wt}$ and $r_b > r_{wt}$ | $r_a < r_{ab} < r_b$ or<br>$r_a > r_{ab} > r_b$ | partial sign epistasis |
| $r_a < r_{wt}$ and $r_b < r_{wt}$ | $r_a < r_{ab} < r_b$ or<br>$r_a > r_{ab} > r_b$ | partial sign epistasis |
| $r_a > r_{wt} > r_b$ | $r_a < r_{ab}$ or $r_b > r_{ab}$ | partial sign epistasis |
| $r_a < r_{wt} < r_b$ | $r_a > r_{ab}$ or $r_b < r_{ab}$ | partial sign epistasis |

**Table 6.1. Explanation of epistasis terms used in this chapter**
As the language of epistasis can often be confusing or inconsistent, we provide here the definitions of various discussed categories of epistatic interactions.

### 6.3. Possible epistatic analysis of a multi-peak ribozyme activity landscape

As discussed in section 5.2, the realities of polymerase-induced errors throw question onto the use of our aminoacylation fitness landscape for epistatic analysis. Thus, current experiments on this system are focused on measuring a new k-Seq data set from a more-normalized library, generated by heavily-mutagenizing the top sequence peaks from every active motif. As no previous study has ever compared the shape and topography of multiple unrelated peaks in a fitness landscape, such analysis has merit, regardless of what specific patterns it shows. Specifically, it may be able to answer whether different evolved mechanisms for the same chemical function, generated from the same biopolymer system, show similar or different patterns of evolvability. To that end, one result is suggested by our existing SCAPE data; once the new *k*-Seq data has been sequenced, new numbers will be run through the same scripts and workflow, with a potential future publication presenting the resultant epistasis analysis.

In analysis using existing SCAPE data, major aminoacylation ribozyme families displayed similar patterns of epistasis. Epistasis was substantial, as the measured activity of double mutants did not correlate with expectation based on single mutants. The distribution of epistasis values was roughly symmetric around 0 (i.e., interactions were roughly equally likely to result in higher vs. lower activity than expected), with the typical epistatic effect having a magnitude $|\varepsilon| \approx 1.5$, indicating that the typical double mutant showed activity ~5-fold different from expectation (Figure 6.3). Most combinations of beneficial mutations show negative epistasis (i.e., a concave-down curvature immediately near the fitness peak), consistent with expectations from diminishing returns epistasis [124]. Measures of higher-order epistasis (triple mutants; see previous section) show a distribution of epistasis values and categories similar to that of double mutants (Figure 6.3), demonstrating that addition of a third mutation has quantitatively similar effects as addition of the second mutation.

The majority of interactions exhibited magnitude epistasis (i.e., the activity of a double mutant was aligned in direction with the expected linear combination of single mutants but differed in magnitude). However, roughly one third of interactions exhibited sign or reciprocal sign epistasis (i.e., the activity of the double mutant was affected in direction opposite to expectation). Sign epistasis represents a drop in activity along either the path wt→*a*→*ab* or the path wt→*b*→*ab*, while reciprocal sign epistasis represents a drop in activity along both such paths, suggesting that 20–40% of evolutionary steps within an edit distance of 2 from the center sequence would be blocked in the strong-selection/weak-mutation regime for each family (Table 6.1).

Analysis of $\gamma_d$ for individual ribozyme families, carried out by another member of the Chen group (Appendix Figure A.6) indicates that the overall topography of each peak can be described as a combination of two components: a 'smooth' component (~40%) in which mutations have additive effects on catalytic activity, and a 'rough' component (~60%) that represents deviations from additivity (i.e., epistasis). This combination resembles the so-called Rough Mt. Fuji model, which consists of a perfectly smooth peak overlaid by uncorrelated ruggedness. However, sequences that did not survive selection were included and assigned baseline activity, then families showed a gradual decrease of $\gamma_d$ as $d$ increased, matching the exponentially decreasing pattern expected for a rugged NK model (Appendix Figure A.6).[125] Here the decreasing pattern of the $\gamma_d$ curve is consistent with an important role for reciprocal sign epistasis associated with highly deleterious mutations and implies that ruggedness increases over larger length scales, consistent with separate peaks in a landscape being uncorrelated. Overall, the different peaks show similar ruggedness, with high-activity regions resembling a RMF model and the overall region resembling an NK landscape.

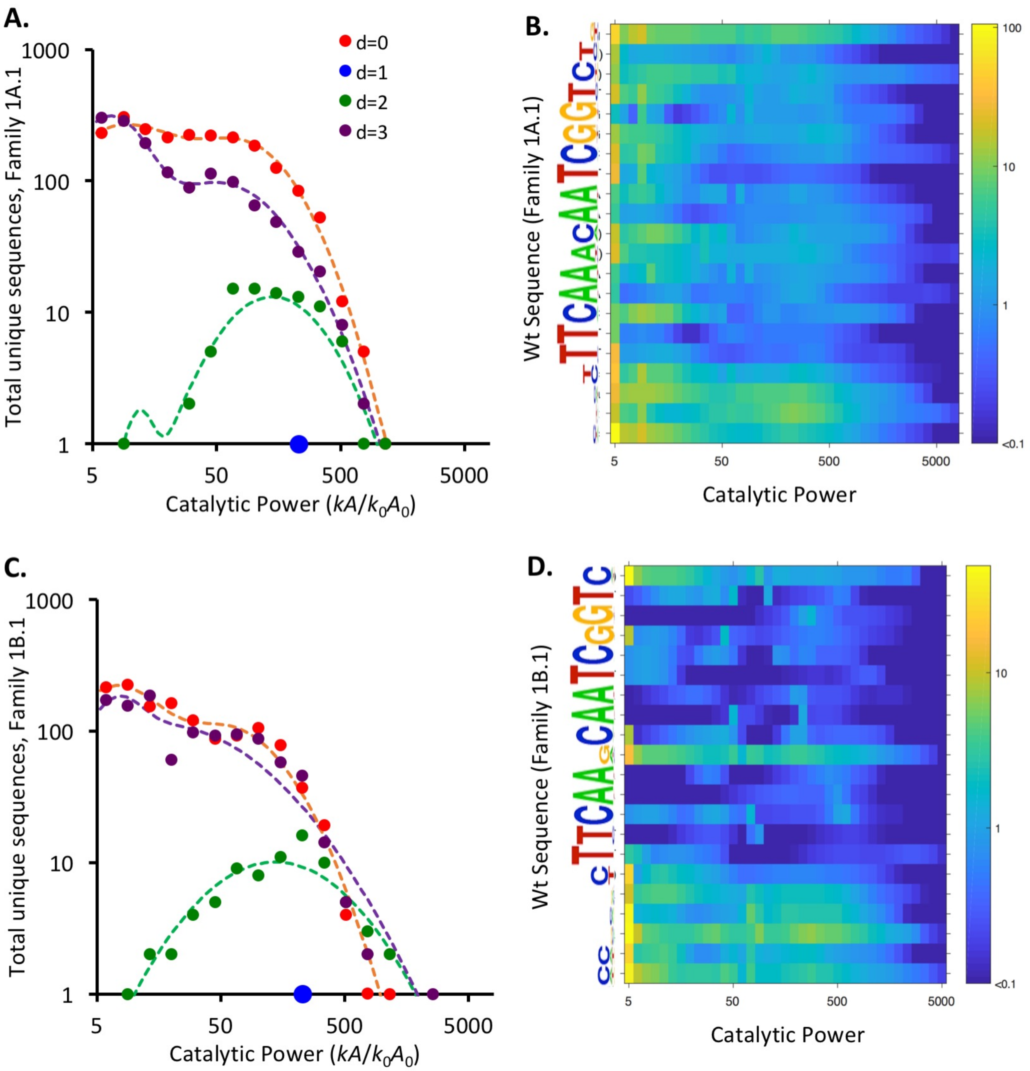


**Figure 6.2. Distribution of fitness effects across individual ribozyme families**
**(A)** Number distribution of catalytic power for sequences of Family 2.1. The blue filled circle indicates the center sequence (edit distance d=0). Single (d=1), double (d=2), and triple (d=3) mutants are shown in red, green, and purple filled circles, respectively. Dotted lines show log bimodal normal fits to each mutant distribution ($R^2$ = 0.98, 0.96, and 0.99, respectively). The overall distribution of fitness effects from single, double, and triple mutants may be conserved across different key mechanistic structures performing the same ribozyme function; further data will help to confirm or refute this. **(B)** Heat map representation of number distribution for single and double mutants of Family 2.1, by nucleotide position (y-axis; SeqLogo given along y-axis). Distributions were smoothed as a sum of normal distributions based on the average and standard error of each sequence's activity (kA) measured by *k*-Seq (weighted based on expected Round 1 loss). Sequences not observed were assumed to have baseline activity. Such approaches may be able to identify sites with a greater portion of deleterious/neutral mutations within large sequence peaks.

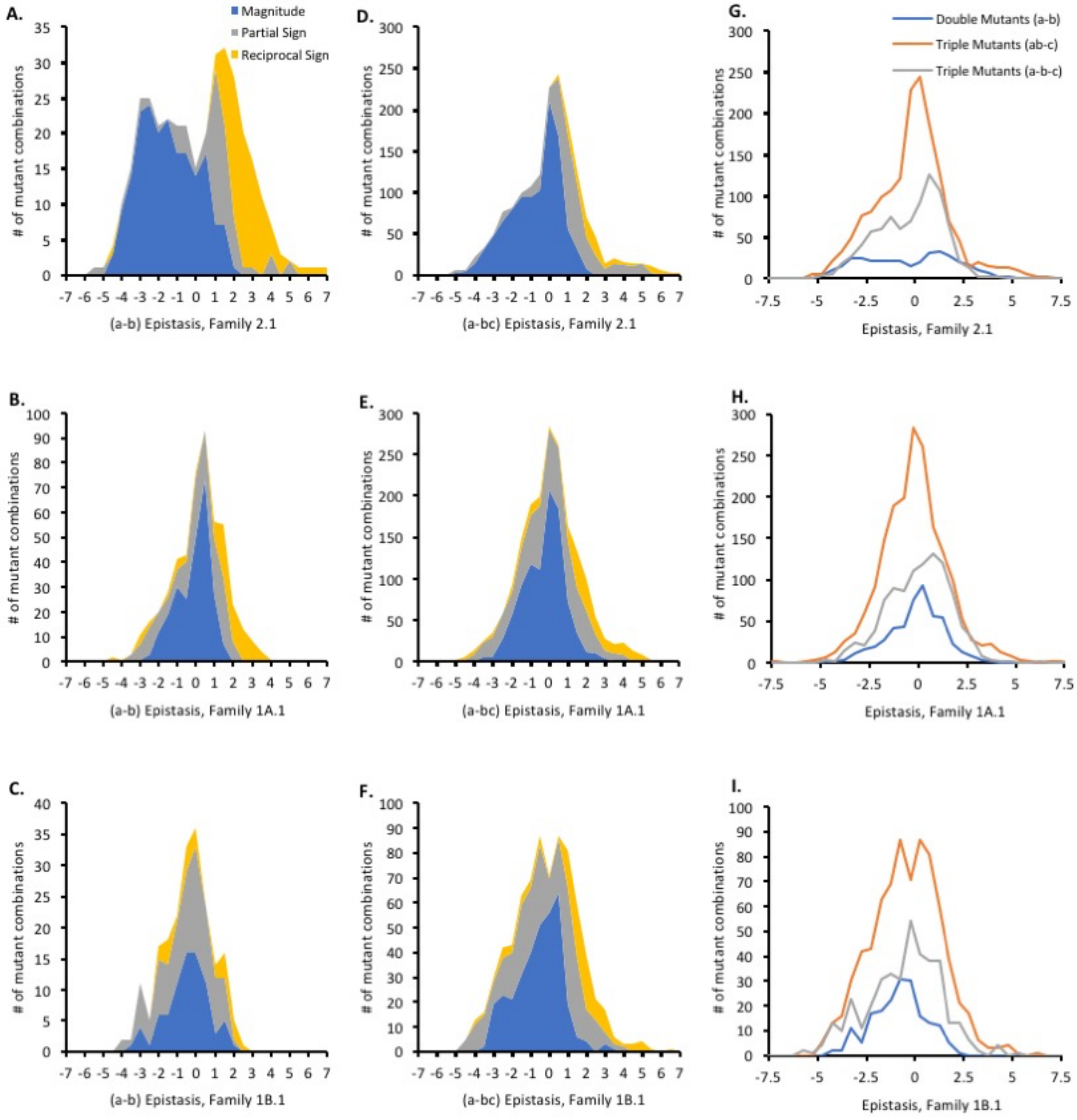


**Figure 6.3. Distribution of epistatic effects across individual ribozyme families**
(**A-C**) show categorization of double mutant (a-b) epistasis and (**D-F**) show categorization of triple mutant (a-bc) epistasis across three different ribozyme families. Overall, the general shape and proportional composition of epistasis appears remarkably similar for each different active ribozyme motif. Further data may be helpful in confirming or refuting this first observation of epistasis compared across different fitness landscape peaks. If true, the extent of homogeneity is significant. (**G-I**) show the overall distribution of epistasis values for a-b (orange), a-bc (blue), and a-b-c epistasis (gray). Results suggest that epistasis may be similar in its magnitude for these peaks at both higher and lower numbers of interacting mutation sites.

### 6.4. Conclusion: Aminoacylation landscape may show surprising homogeneity

The prevalence of epistasis, and the decrease of $\gamma_d$ to 0 across low-activity sequences, suggest our aminoacylation landscape may be consistent with the "thermodynamic threshold model" in which biomolecular activity falls abruptly when mutations decrease stability below a sustainable level.[126,127] This may help to explain the relatively small number of sequences ($>10^5$ out of $>10^{12}$) displaying activity significantly higher than the baseline; it may also be a feature of intermediate-to-low-activity sequences simply failing to survive the initial selection rounds of SCAPE. Unlike previous studies of epistasis across a substantial local fitness landscape,[126,128] we observed roughly equal frequency of positive and negative epistasis values. That is, an additional mutation was equally likely to interact synergistically or antagonistically with the genetic background. Epistasis and ruggedness metrics are surprisingly consistent across several unrelated ribozyme families, suggesting that epistatic findings may be generalizable either across the entire fitness space of RNA-catalyzed aminoacylation, or possibly across a larger category of ribozymes as a whole. Overall, we expect further, higher-quality epistasis data to soon answer these questions, and in the future provide an additional analytic step downstream of SCAPE analysis.

# 7. Future prospects and conclusions

*"A mathematician may say anything he pleases, but a physicist must be at least partially sane."*

– J.W. Gibbs, *On the Relation of Mathematics and Physics, Dec. 1944*

## 7.1. Landscapes for molecular specificity and environmental conditions

A current open question in the RNA world is whether changing pH, salt concentration, or molecular crowding in an early earth ocean might make certain evolutionary pathways more or less accessible; but tolerance to environmental conditions is also a general concern in the optimization of many functional proteins.[129-134] Studies of local fitness landscapes in proteins and functioning cellular RNA have suggested that landscapes can vary significantly under different reactive or growth conditions, including changes to the overall evolvability and epistasis of these systems. [135,136]

The *k*-Seq methods described in this work should be able to easily investigate such questions as they apply to environmental effects on the evolvability of an aminoacylation ribozyme. The actual environmental conditions of an early earth environment are fairly unknown; it remains to be seen whether varying reaction buffers or other conditions might increase the appearance of neutral pathways between currently-isolated fitness peaks in our aminoacylation landscape. The answers to such questions require only repeating certain experimental steps under different reaction conditions—and the ability to build fitness landscapes under multiple different conditions may be of special interest to bioengineered systems designed to function both *in vitro* and *in vivo*.

Beyond selecting sequences with high and predictable fitness, another important consideration in *in vitro* selections is specificity. In aptamers and antibodies, a highly specific binder forms a very stable with its target molecule and not other, similar molecules; in ribozyme and enzyme selections, a high-specificity catalyst incorporates one substrate into a reaction far more readily than other similar alternatives. Aptamers and antibodies, when selected naively, sometimes display too much nonspecific binding to be useful in a diagnostic or biological setting. Often, a selection can generate several high-fitness sequences whose specificities vary dramatically, and it can be hard to predict whether a selection will produce aptamers or ribozymes of useful specificity, requiring characterization and testing of each final sequence to measure this.[137,138] Counter-selection to remove sequences from a pool with an "off-target" substrate can help to control specificity,[139] but counter-selection also increases the runtime and complexity of selections, with no current work examining ideal conditions for such steps.

One of the challenges in dealing with specificity of selection products is that no research exists comprehensively investigating the distribution of specificities in any *in vitro* selection. Competing theories in aptamer selection suggest either selecting high-affinity binders and then optimizing with mutagenesis for specificity of binding, or selecting high-specificity binders and then optimizing those to improve binding fitness.[23,137] The effectiveness of either approach depends on the combined distributions of fitness and specificity over sequence space; a pool whose peaks show mostly constant affinity but locally-varying specificity would suggest the first approach, while a pool whose peaks show mostly constant affinity with locally-varying affinity would suggest the second. Measuring the specificity space of ribozymes relevant to early life could also potentially answer questions about an RNA world, demonstrating which ribozyme functions may have been more likely to descend from a single progenitor with mutation-varied substrate specificity than from parallel evolution for the same.

Conceptually, mapping and building specificity landscapes for a given selection or family of selections should function similarly to calculating fitness landscapes. The next step of the aminoacylation landscape study, currently being carried out by new members of Chen lab, involves using SCAPE to build parallel fitness landscapes for aminoacylase ribozymes selected with different amino acid substrates. We expect that this will allow construction of the first large, complete landscape of catalytic *specificity*, answering questions of whether one basal aminoacylase could evolve into many specific variants, or whether different aminoacylases more likely evolved independently. Demonstrating the use of this method for building specificity landscapes should also show the extent of its feasibility for answering questions of specificity in general, which we expect, combined with questions of environmental generation, to form the next generation of fitness landscape approaches to understanding engineered evolution.

## 7.2. SCAPE and *k*-Seq tools

Beyond the experimental side, other work being carried out on this project involves making existing tools for SCAPE and k-Seq easier to use. To that end, the collection of scripts currently used for this analysis is being refined into a set of easier-to-use tools for distribution on the Chen lab github. (github.com/ichen-lab-ucsb). Concurrent work by another member of the Chen group has identified minor improvements to the mathematics of the SCAPE method, which are described in Appendix A.4.

### 7.3. Final Thoughts

The overall goal of this project was not simply to study several specific questions, but to build a better analytical framework for studying evolutionary spaces—and, to some extent, to shift the level at which fitness landscapes are discussed. The practical uses of artificial selection are, to some extent, not as pronounced today as they were five years ago. Advances in molecular screening and DNA synthesis allow sizeable landscapes of molecules to be constructed and immediately tested from scratch, in workflows increasingly paired with *in silico* predictions of putative catalysts or aptamers. Directed evolution still remains a powerful and widely-used tool for maturation of pharmaceutical antibodies, as well as our best method for investigating possible reactions in the origin of life, but rational design methods are beginning to answer some questions that once only selections could touch. At the conferences I have personally attended, selection just feels like a less sexy research area than it used to; synthetic biology, as a field, seems to be moving away from new ways to find functional molecules and towards new things to do with those molecules and methods.

But artificial selection, the oldest tool in the bioengineer's handbook (by around 10,000 years), isn't going away just yet. And the work described in this report demonstrates several ways that analysis of selections can answer questions it previously could not. Very few basic fitness distributions have been mapped out at any notable resolution, for any biological function. With a new theory-based approach to approximate these "starting curves" for directed evolution, we have the ability to greatly expand our repertoire of knowledge regarding the distribution of chemical activity over random biopolymer space. Such distributions may one day be helpful in providing realistic bounds on *in silico* simulations, and the ability to estimate these distributions as they evolve may be of direct aid now to the automation of genetic engineering, which remains a rapidly-growing field of interest.

Plus, no matter how much progress is made in biopolymer synthesis and molecular simulation, it will be a long time before any method other than selection can investigate sequence spaces on the order of trillions of functional molecules. With fitness estimation and *k*-Seq methods, we are able to bridge some of the gap between screening and selection methods, allowing any simple selection to be transformed into a direct chemical activity assay. By combining the search space of selection with the quantitative nature of such screens, our SCAPE methodology has successfully built the largest activity landscape ever characterized, covering a sequence space six orders of magnitude larger than anything similarly measured. We have observed the first comprehensive fitness landscape with multiple unrelated peaks, quantifying the paths between unrelated ribozyme mechanisms as evolutionarily unfeasible in an aminoacylation system. Questions of homogeneity across a large and diverse fitness landscape may also be answerable, though final confirming data is still needed. Perhaps most significantly, we have effectively mapped all possible routes by which a specific catalytic activity can evolve from random RNA space, hopefully creating a new standard that can be used in many future studies on the evolutionary possibilities of novel chemical functions.

Evolution is a powerful engineering tool, and its applications are far more diverse than a single reaction, or even a single type of functional molecule. It is my sincerest hope that this work is able to inspire other researchers, across different areas of biochemistry and engineering, to approach their research on evolution and artificial selection with a greater eye for mathematical analysis, and a better set of tools to perform that analysis.

## Academic Acknowledgements

This work benefitted significantly—more than I ever could have expected—from discussions and collaborations with a great number of other scientists.

Some of the ideas for this project began before I even joined the Chen lab, primarily through conversations with Greg Campbell, as well as with Scott Ferguson and Andrew Csordas. Andrew, Greg, and JP Wang taught me a large quantity of my present lab skills as they relate to *in vitro* selection. My work on Fisher's theorem and the use of population dynamics in optimizing benefitted from conversations with Tetsuya Yomo, Angela Zhang, David Ross, and Kevin Esvelt. Uli Müller, Janina Moretti, Ziwei Liu and Robert Pascal all provided experimental work and chemical insights through various collaborations. The experimental basis for *k*-Seq was developed with significant assistance from Evan Janzen, while the theoretical basis for the SCAPE methodology was developed with insight from Evan, Uli, Yuning Shen and Jerry Joyce. This project benefitted from base code written by Greg and Ramon Xulvi-Brunet; Yuning and Celia Blanco helped sanity check my own code. Celia also assisted with understanding the minutiae of fitness landscapes and their extensive and tumultuous depictions in the literature. And of course, Irene helped with basically all of the above.

# Appendix A: Additional Protocols and Data

## A.1. Data on tested sequences

Here we present kinetic data on all sequences tested as part of the analysis in Chapters 3 and 4. Unless otherwise indicated, error bars are the result of triplicate analyses.

| Sequence Name | Initial rate estimated by HTS ($k_eA_e$) ($min^{-1}M^{-1}$) | R8(5m) Abundance | Initial rate measured experimentally ($k_iA_i$) ($min^{-1}M^{-1}$) | Previously Identified? |
|---|---|---|---|---|
| 1-S | 0.476 ± 0.238 | 0.1581 | 0.124 ± 0.033 | No |
| 2-S | 0.761 ± 0.381 | 0.0449 | 0.278 ± 0.149 | No |
| 3-S | 0.369 ± 0.185 | 0.0229 | 0.807 ± 0.171 | No |
| 4-S | 0.431 ± 0.215 | 0.0435 | 0.212 ± 0.026 | No |
| 6-S | 1.547 ± 0.774 | 0.0497 | 0.770 ± 0.166 | No |
| 8 | 0.298 ± 0.149 | 0.0759 | 0.080 ± 0.46 | No |
| 11-S | 1.192 ± 0.596 | 0.0125 | 1.638 ± 0.078 | No |
| 22-S | 1.072 ± 0.536 | 0.0060 | 0.412 ± 0.281 | No |
| R5_3C21 | 0.115 ± 0.058 | 0.0060 | 0.074 ± 0.083 | Yes |
| R8_35C18A | 0.143 ± 0.072 | 0.0097 | 0.023 ± 0.003 | Yes |
| R8_35C18B | 0.173 ± 0.087 | 0.0130 | 0.020 ± 0.001 | Yes |
| R8_55C10 | 0.423 ± 0.212 | 0.0102 | 0.030 ± 0.001 | Yes |
| R8_35C10 | 0.173 ± 0.087 | 0.0130 | 0.050 ± 0.001 | Yes |
| R8_55C18 | 0.158 ± 0.079 | 0.0030 | 0.036 ± 0.001 | Yes |
| R8_35C16 | 0.295 ± 0.148 | 0.0457 | 0.025 ± 0.001 | Yes |

**Table A.1. Activity of tested TMP ribozymes**
**Table 2.** Ribozyme activity assayed experimentally, comparing newly identified ribozymes with the best previously identified [99]. Estimated and measured values were calculated as described in Supplementary Text. Bold indicates newly-identified sequences. Errors given are ± 1 standard deviation, as described in Methods and Supplementary Text, with the standard deviation of $k_e$ calculated from cluster abundance as a scale-variant error of ±50% for cluster enrichment.

| New Sequence | Measured Ligation % (3h) | Similar Previously-Tested Sequence | Measured Ligation % (3h) | Edit distance between sequences |
|---|---|---|---|---|
| 1-S | 3 | R8_55C12 | 4 | 13 |
| 11-S | 52 | R8_35C16 | 24 | 6 |
| 6-S | 39 | R8_55C4 | 7 | 5 |
| 2-S | 21 | R8_55C6 | 4 | 1 |

**Table A.2. Triphosphorylase families undergoing a notable mutational shift**
Of sequences predicted to have high fitness and tested experimentally in isolation, four were similar to previously tested sequences[99] but differed by possessing notable mutations. In three of these four cases, the minor sequence variations caused a substantial increase in ribozyme activity; one of these differed by only a single nucleotide (the center shift of a notable mutation).

| Name | Sequence (random region) | A (by gel) | A (by k-Seq) | k (by gel) ($min^{-1}M^{-1}$) | k (by k-Seq) ($min^{-1}M^{-1}$) | Reason chosen |
|---|---|---|---|---|---|---|
| S-2.1-a | ATTACCCTGGTCATC GAGTGA | 0.450 ± 0.012 | 0.161 ± 0.007 | 1570 ± 260 | 779 ± 21 | Most abundant sequence, most abundant family |
| S-2.1-b | ATTACCCTGGTCATC GAGTGT | 0.446 ± 0.072 | 0.158 ± 0.007 | 890 ± 267 | 729 ± 28 | Second-most abundant seq., most abundant family |
| S-1A.1-a | CTACTTCAAACAATC GGTCTG | 0.708 ± 0.008 | 0.283 ± 0.069 | 303 ± 29 | 121 ± 11 | Most abundant sequence, second-most abundant family |
| S-1B.1-a | CCACACTTCAAGCAA TCGGTC | 0.708 ± 0.124 | 0.865 ± 0.185 | 247 ± 49 | 46.2 ± 17.6 | Most abundant sequence, third-most abundant family |
| S-1B.2-a | CCGCTTCAAGCAATC GGTCGC | 0.704 ± 0.238 | 0.669 ± 0.275 | 112 + 23 | 47.3 ± 11.5 | Most abundant sequence, fourth-most abundant family |
| S-1B.3-a | CCGAGTTTCAAGCAA TCGGTC | 0.700 ± 0.064 | 0.458 ± 0.313 | 194 ± 19 | 71.2 ± 20.6 | Expected to have medium-high activity |
| S-3.1-a | AAGTTTGCTAATAGT CGCAAG | 0.825 ± 0.006 | 0.134 ± 0.013 | 169 ± 12 | 142 ± 3 | Most abundant sequence, most active family of motif 3 |
| S-2.2-a | ATTCACCTAGGTCAT CGGGTG | 0.404 ± 0.050 | 0.132 ± 0.019 | 355 ± 75 | 197 ± 9 | Second-most active family from most-active motif |
| S-1A.4-a | CTCTTCAAACAATCG GTCTTC | 0.719 ± 0.249 | 0.251 ± 0.145 | 127 ± 860 | 74.9 ± 5.2 | Expected to have medium-low activity |

| | | | | | | |
|---|---|---|---|---|---|---|
| S-1C.1-a | CTCTTCAATAATCGG TTGCGT | 0.516 ± 0.048 | 1.000 ± 0.000 | 81.5 ± 20.4 | 6.65 ± 0.75 | Most abundant sequence, least active motif |
| Baseline Activity | N/A | | | 0.645 ± 0.283 | 0.124 | Baseline catalytic activity of starting pool |

**Table A.3. Aminoacylation ribozymes chosen for gel-shift assay.**
Ten sequences were chosen for gel-based activity testing; nine were the highest-abundance centers of a range of different sequence families (± indicates standard deviation of triplicates). k-Seq activity estimates were not adjusted for expected loss due to column binding and recovery, as described in Methods; expected loss from a linear fit of these data is 80.7% (or 19.3% of aminoacylated sequences retained). Since the baseline activity measurement does not include these losses, we calculate the estimated k-Seq equivalent baseline $k_0A_0 = 0.193 * 0.645 = 0.124$ min$^{-1}$M$^{-1}$, for comparison to k-Seq values in calculation of catalytic ratio.

| Start Sequence | End Sequence | Name | Total Path Length | Path Steps | Largest Step | Minimum Count |
|---|---|---|---|---|---|---|
| S-1A.1-a | S-1B.1-a | 1A-1B:1 | 7 | 7 | 1 | 4 |
| | | 1A-1B:2 | 7 | 7 | 1 | 4 |
| | | 1A-1B:3 | 7 | 7 | 1 | 4 |
| | | 1A-1B:4 | 7 | 7 | 1 | 4 |
| | | 1A-1B:5 | 7 | 7 | 1 | 4 |
| | | 1A-1B:6 | 6 | 5 | 2 | 12 |
| | | 1A-1B:7 | 6 | 5 | 2 | 12 |
| | | 1A-1B:8 | 6 | 5 | 2 | 9 |
| | | 1A-1B:9 | 6 | 5 | 2 | 9 |
| | | 1A-1B:10 | 6 | 5 | 2 | 7 |
| S-1A.1-a | S-1C.1-a | 1A-1C:1 | 8 | 6 | 2 | 4 |
| | | 1A-1C:2 | 8 | 6 | 2 | 4 |
| | | 1A-1C:3 | 8 | 6 | 2 | 4 |
| | | 1A-1C:4 | 8 | 6 | 2 | 3 |
| | | 1A-1C:5 | 8 | 6 | 2 | 3 |
| | | 1A-1C:6 | 7 | 5 | 3 | 13 |
| | | 1A-1C:7 | 7 | 5 | 3 | 10 |
| | | 1A-1C:8 | 7 | 5 | 3 | 7 |
| | | 1A-1C:9 | 7 | 5 | 3 | 7 |
| | | 1A-1C:10 | 7 | 5 | 3 | 3 |
| S-1B.1-a | S-1C.1-a | 1B-1C:1 | 12 | 9 | 2 | 2 |
| | | 1B-1C:2 | 12 | 9 | 2 | 2 |
| | | 1B-1C:3 | 12 | 8 | 2 | 2 |
| | | 1B-1C:4 | 12 | 8 | 2 | 2 |
| | | 1B-1C:5 | 12 | 8 | 2 | 2 |
| | | 1B-1C:6 | 11 | 7 | 3 | 3 |
| | | 1B-1C:7 | 11 | 7 | 3 | 3 |
| | | 1B-1C:8 | 11 | 7 | 3 | 3 |
| | | 1B-1C:9 | 11 | 7 | 3 | 2 |
| | | 1B-1C:10 | 11 | 7 | 3 | 2 |
| S-1A.1-a | S-2.1-a | 1A-2:1 | 24 | 11 | 4 | 2 |
| | | 1A-2:2 | 24 | 11 | 4 | 2 |
| | | 1A-2:3 | 24 | 11 | 4 | 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 1A-2:4 | 24 | 11 | 4 | 2 |
| | | 1A-2:5 | 24 | 11 | 4 | 2 |
| | | 1A-2:6 | 15 | 7 | 5 | 2 |
| | | 1A-2:7 | 15 | 7 | 5 | 2 |
| | | 1A-2:8 | 15 | 7 | 5 | 2 |
| | | 1A-2:9 | 15 | 7 | 5 | 2 |
| | | 1A-2:10 | 15 | 7 | 5 | 2 |
| S-1A.1-a | S-3.1-a | 1A-2:1 | 27 | 11 | 4 | 2 |
| | | 1A-2:2 | 27 | 11 | 4 | 2 |
| | | 1A-2:3 | 27 | 11 | 4 | 2 |
| | | 1A-2:4 | 27 | 11 | 4 | 2 |
| | | 1A-2:5 | 27 | 11 | 4 | 2 |
| | | 1A-2:6 | 12 | 9 | 5 | 2 |
| | | 1A-2:7 | 12 | 9 | 5 | 2 |
| | | 1A-2:8 | 12 | 9 | 5 | 2 |
| | | 1A-2:9 | 12 | 9 | 5 | 2 |
| | | 1A-2:10 | 12 | 9 | 5 | 2 |
| S-2.1-a | S-3.1-a | 1A-2:1 | 21 | 9 | 4 | 2 |
| | | 1A-2:2 | 21 | 9 | 4 | 2 |
| | | 1A-2:3 | 21 | 9 | 4 | 2 |
| | | 1A-2:4 | 21 | 9 | 4 | 2 |
| | | 1A-2:5 | 21 | 9 | 4 | 2 |
| | | 1A-2:6 | 18 | 7 | 5 | 2 |
| | | 1A-2:7 | 18 | 7 | 5 | 2 |
| | | 1A-2:8 | 18 | 7 | 5 | 2 |
| | | 1A-2:9 | 18 | 7 | 5 | 2 |
| | | 1A-2:10 | 18 | 7 | 5 | 2 |
| S-2.1-a | S-2.2-a | 2-2.1:1 | 5 | 5 | 1 | 2 |
| | | 2-2.1:2 | 5 | 5 | 1 | 2 |
| | | 2-2.1:3 | 5 | 5 | 1 | 2 |
| | | 2-2.1:4 | 5 | 5 | 1 | 2 |
| | | 2-2.1:5 | 5 | 5 | 1 | 2 |
| | | 2-2.1:6 | 5 | 4 | 2 | 40 |
| | | 2-2.1:7 | 5 | 4 | 2 | 13 |
| | | 2-2.1:8 | 5 | 4 | 2 | 13 |
| | | 2-2.1:9 | 5 | 4 | 2 | 13 |
| | | 2-2.1:10 | 5 | 4 | 2 | 3 |

**Table A.4. Properties of pathways between aminoacylase ribozyme peaks**
For each pathway found between peak centers of interest, 10 pathways were found as described in Methods. For each path, "largest step" corresponds to the edit distance of the largest step present within the pathway, and "minimum count" is the lowest Round count (# of sequence reads) of any sequence that the pathway passes through.

## A.2. Identification of aminoacylation ribozyme sites for top motifs

The work in this section was primarily done by Evan Janzen as part of an upcoming publication (though I assisted with interpreting the results), but may also be helpful in understanding the overall project described in Chapter 4.

The most highly abundant sequences from each major aminoacylation ribozyme motif were chosen (S-1A.1-a, S-1B.1-a, S-2.1-a, S-3.1-a; see Methods for sequence nomenclature) for characterization of the reactive site. Identification of the 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 2' acylation sites. The putative ribozymes were ligated to a 3' adapter in order to test for stalling of reverse transcription along the entire length of the RNA. Stalling resulted in a truncated product whose length, determined by gel electrophoresis, suggested a likely site of aminoacylation (Figure A.5A). Second, the nucleophilic importance of the 2'OH at the candidate site was verified by testing the activity of a synthetic RNA sequence modified at this position by 2'-*O*-methylation. In each case, a control synthetic RNA sequence that was instead modified at an adjacent position was also tested. Blocking of the candidate site (but not the control sites) 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 (Appendix Figure A.5). 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 mechanisms.

Note that the catalytic ratio $r_i$ 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 lower than $k_0A_0$, which we 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 [140,141]. In contrast, we found that 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 at least 100-fold greater (and likely several orders of magnitude greater) than the $r_i$ as reported in this work, suggesting significant catalytic rate enhancement is possible for oxazolone-precursor aminoacylation.

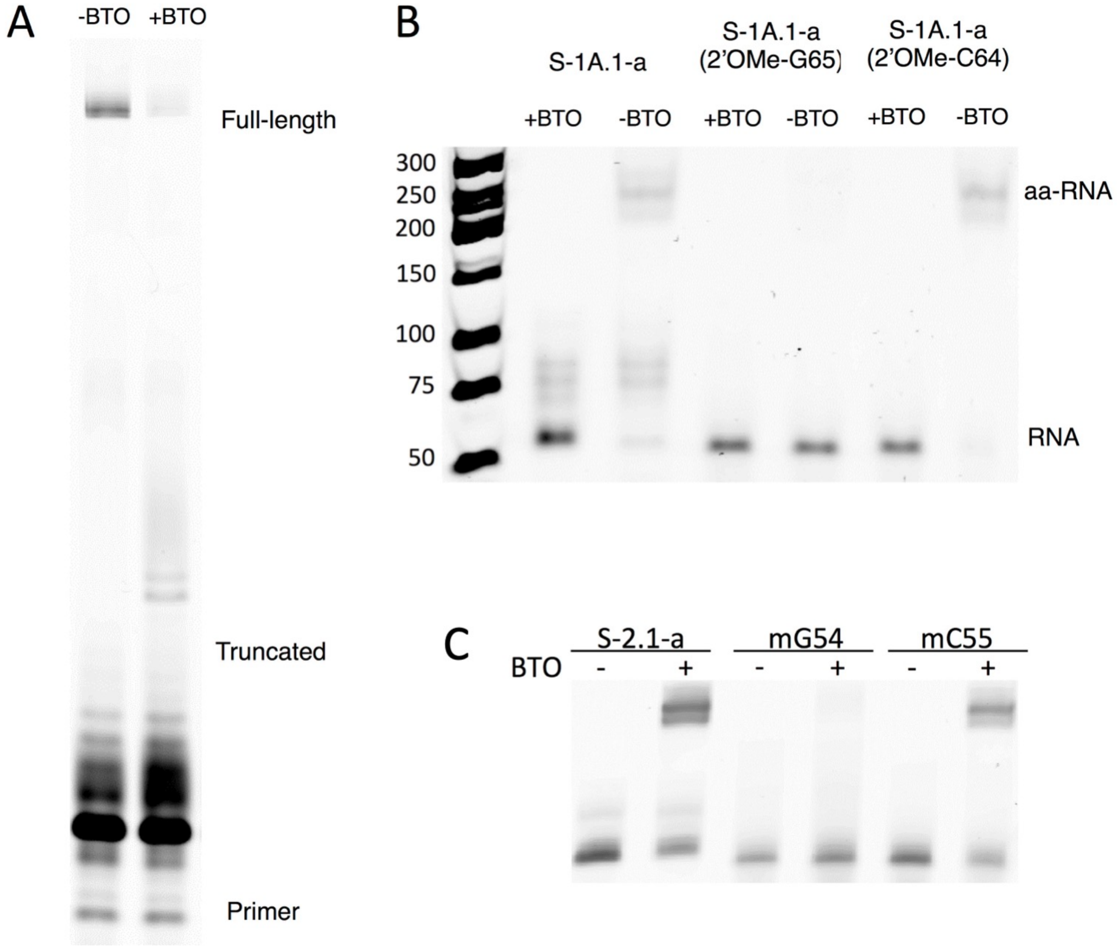


**Figure A.5. Location of ribozyme oxazolone-aminoacylation sites**
The likely site of BTO modification on ribozyme S-1A.1-a was identified by stalling of reverse transcription, resulting in a truncated product **(A)** The site, G65, was verified by loss of activity upon 2'-O-methylation, assayed by streptavidin gel shift after BTO reaction **(B)** 2'-O-methylation of an adjacent site (C64) did not show loss of activity. **(C)** For comparison, a gel shift assay similar to those used to quantify aminoacylation. Here, BTO-aminoacylated RNA can bind to streptavidin, resulting in a shifted gel band from the larger complex.

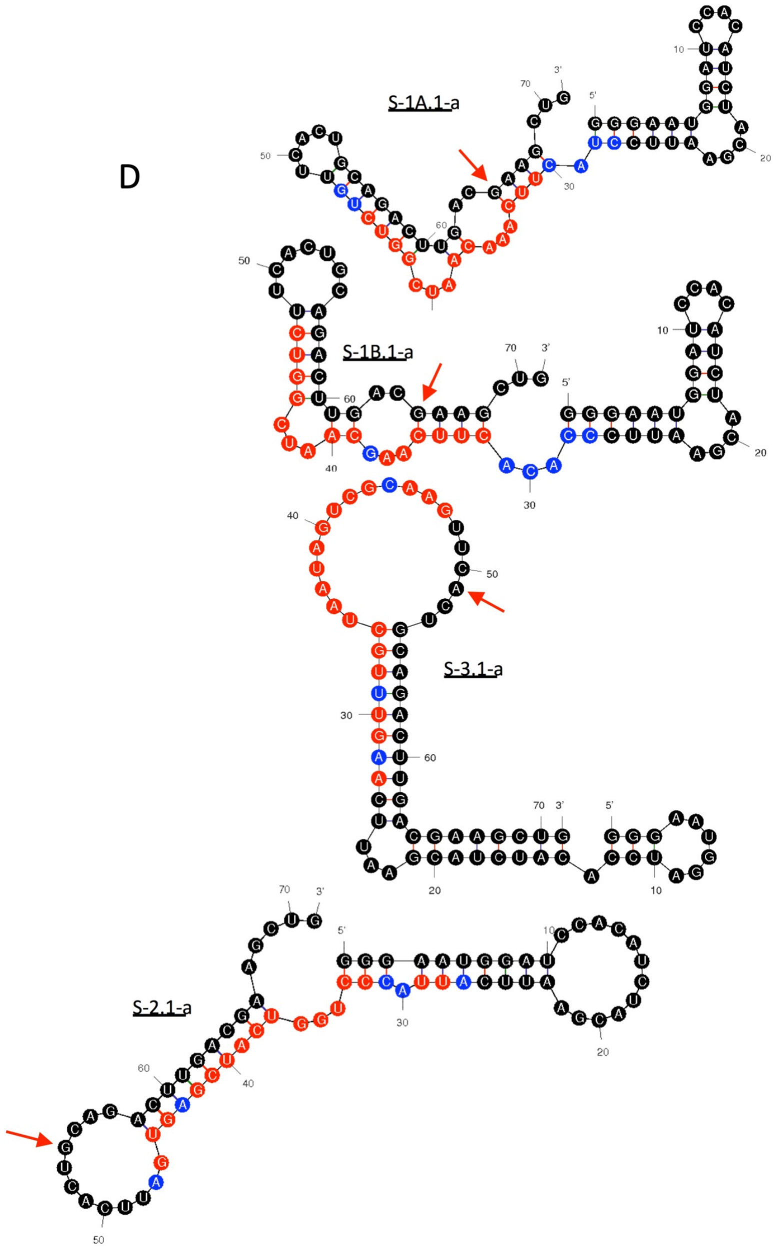


**Figure A.5 continued. Location of ribozyme oxazolone-aminoacylation sites**
**(D)** Minimum free energy secondary structures for the sequences indicated, predicted by mfold. Black denotes constant regions. Sites in the random region conserved with information entropy < 1 bit are shown in blue; sites with information entropy > 1 bit are shown in red (also see Fig. 4.2 D). Red arrows indicate the observed aminoacylation site.

### A.3. Epistatic correlation parameter for an aminoacylation landscape

The work in this section was primarily done by Dr. Celia Blanco as part of an upcoming publication (though I assisted with interpreting the results), but may also be helpful in understanding the overall project described in Chapters 4 and 6.

To understand how the overall character of the ribozyme fitness landscape compares with well-known theoretical models, we characterized the ruggedness of the ribozyme peaks. Generally, the fitness of close relatives is highly correlated to each other, but the fitness of more distant relatives is less correlated. A simple measure of ruggedness is the fitness correlation $\gamma_d$ for a ribozyme family, which is the average correlation of activity effects of single mutations between sequences at evolutionary distance $d$ of each other [123] ($d$ is the Levenshtein edit distance, i.e., the number of substitutions, insertions or deletions between two related sequences). $\gamma_d = 1$ indicates a perfectly smooth landscape, and $\gamma_d = 0$ indicates a highly rugged, completely uncorrelated landscape. $\gamma_1$ was approximately 0.4 for all families analyzed, i.e., the typical effect of a particular mutation is 40% correlated across all single-mutant backgrounds (Figure 3C, Supporting Figure 7), indicating substantial ruggedness on the fitness peaks. Interestingly, as the neighborhood size increased up to $d = 4$, $\gamma_d$ dropped only slightly, indicating that activity remained similarly correlated at longer evolutionary distances within the peaks. The relative constancy of $\gamma_d$ over a range of $d$ indicates an underlying smoothness that is felt throughout the peak.

A general interpretation of these results may be that, overall, the correlation between various sites is preserved even against a random background effect on fitness. In other words, the core conserved motif, and how its important elements bind with or are oriented towards each other, changes very little across a peak corresponding to a single reaction mechanism. As with the other epistatic analysis in this work, additional data will be helpful in confirming the precise extent of this phenomenon.

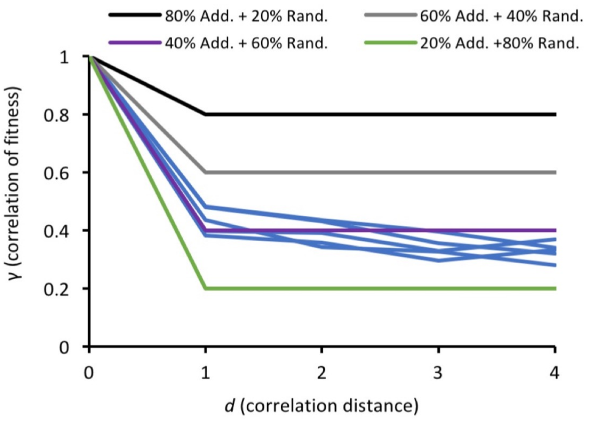


**Figure A.6. Epistatic correlation may be consistent across aminoacylation ribozyme peaks**
Average correlation of fitness effects are shown for the Rough Mt. Fuji model, calculated with different amounts of additivity vs. randomness (see legend). The observed data, corresponding to the top 5 most active peaks is shown in blue, suggesting significant agreement to this model (at least in terms of the correlatedness of fitness effects across the motif sequence).

## A.4. Refinements to the *k*-Seq calculation

Updates to the k-Seq methodology has been developed primarily by Yuning Shen, as part of an upcoming publication, developed and tested through numerous data simulations. While the extent of that evaluation is beyond the scope of this particular work, it may be helpful to know that a slight variation on the calculations will be included as a default option in future versions of the k-Seq tools. Essentially, the updated method fits an activity curve to all measured data points, rather than averaging for each set of substrate/reaction conditions. In this case, bootstrapping is used instead of random triplicate grouping in order to estimate error bars on fitness fits. This new approach seems fairly robust, and will likely be fully adopted as part of the relevant workflow.

### A.5. *k*-Seq protocols

As a potentially useful aide to future Chen lab members or other readers of this work, several write-ups of lab methods from the aminoacylation k-Seq paper are provided. The first of these is a selection protocol, which details potential minute errors in carrying out the selection and how to avoid them; the second consists of the help docs for the SCAPE analysis tools currently available on the Chen lab github. These are not meant to be of interest to the casual reader, but merely a readily-available place of documentation.

Oxazolone aminoacylase ribozyme selection protocol (also applicable to k-Seq rounds):

Start: Make some RNA
-I usually run a double transcription reaction (2 tubes) with ~200 ng of dsDNA per reaction tube, which should give 50-100 µg of RNA for that round—plenty of extra, in case the selection needs to be done a few times

Start: Reconstitute some BTO
Reminders:
-It'll take a lot of vortexing/sonication to get this stuff to dissolve
-I typically defrost a new tube every 2 weeks or so when doing the selection (it's probably fine for about 1 month in the freezer once reconstituted, though, if you're only looking for qualitative results)
-Make sure to avoid accidentally heating up any of the non-reconstituted tubes (don't take the box out of the freezer for longer than you have to)

Selection
**Step 1:** Set up the reaction
Reminders:
-RNA can stick to the sides of tubes when it freezes. Use Eppendorf or other DNA lo-bind tubes when possible; make sure to vortex RNA tubes well after they are thawed

-Make sure BTO tube has reached ~ room temperature or so (the outside of the tube isn't cold) before opening it (if you don't, water will precipitate on the sides of the tube, causing the BTO to break down faster)
Procedure:
Do the following:
-2 reactions, for 90 minutes total
-Each reaction: 100 µL buffer (the pH 6.95 HEPES selection buffer), 5 µg of RNA, 50 µM BTO
-The BTO stock should be at 25 mM. To get it to 50 µM, you're doing a 500-fold dilution. I find it's easiest to make a 1:10 dilution of this (45 µL of buffer + 5 µL of BTO stock), then add the dilution stock in a 1:50 volume ratio to each reaction

**Step 2:** Set up next few steps
I usually start at around 60 minutes after the reaction starts. This gives me 30 minutes (plenty of time) to set up the next few steps.
Reminders:
-P30 spin columns should be inverted several times; after you snap off the bottom piece of plastic, shake them once more
-If streptavidin-coated beads ever clump together, use the sonicator for a few seconds, then vortex them, to help them re-dissolve
-Magnetic beads should only be left on the magnetic rack for 1-2 minutes at a time (though try to let the liquid turn clear before you remove it). As soon as you remove the liquid from a tube containing magnetic beads, remove the tube as well (leaving the beads "dry" on the magnetic rack will cause them to roll out of the remaining liquid and into the air, where they can dry out and damage the streptavidin)
-I usually like to use an empty Falcon tube (50 mL) to quickly eject my wash liquid into (as a small "liquid waste" that can then be thrown out because there's nothing hazardous in it)
Procedure:
-Cast a denaturing PAGE gel (I use a small gel, at 6% acrylamide, here). Skip this if you're doing a qPCR instead
-Prep spin columns, 1 per 100 µL of RNA reaction (do this by spinning for 2 minutes at 1000 G, then discarding flow-through, then spinning for 2 minutes at 1000 G a second time, then placing each column into a new 1.5 mL Eppendorf tube)
-Get Streptavidin beads out of fridge (I've been using 50 µL of bead stock for every 100 µL reaction tube). Vortex the bead stock well, then aliquot the amount you're using into a new tube.
-Turn heating block on at 65° C
**Procedure** (if you have time; if not, you can put off the bead-washing until a bit later)
-Wash your new tubes of beads as follows:
1. Remove liquid with magnet, add equal volume of bead buffer (bead buffer is 1x PBS, pH ~ 7.5, with 0.01% Triton X-100 added)
2. Remove liquid with magnet, add equal volume of 20mM NaOH
3. Remove liquid with magnet, add equal volume of bead buffer
4. Remove liquid with magnet, add equal volume of bead buffer
Tips: If you're doing 200 µL of total reaction, you'll want to leave your pipet set to 100 µL here; push it all the way down (past the stop) when removing liquid, so you can get it all easily (try not to leave much in the tube).

**Step 3:** Stop the reaction (should happen ~90 minutes after reaction is started)
Reminders:
-Make sure to vortex tubes frequently (especially after eluting RNA from columns)
-Make sure to always spin tubes down before placing them on the magnetic rack, so that no liquid is up on the sides of the tube
Procedure:
-Use prepped spin columns to remove BTO from reactions (put up to 100 µL of reaction into each P30 spin column)
-Add washed beads to column elutions (At a ratio of 1 bead volume : 2 reaction volume)
-Vortex bead+RNA tubes
-Place on rotator for 10-15 minutes

Prep for next step:
-Prep additional spin columns (you probably only need 1) for formamide buffer exchange in next step
-Take out reagents for Reverse Transcription (primers, DTT, DNTPs, forward strand buffer) now so that they can thaw

**<u>Step 4:</u>** Wash + Elute RNA from beads
Procedure:
-Take bead+RNA tubes off rotator (I usually combine them into one tube at this point for ease of handling)
-Wash beads as follows:
1. Remove liquid with magnet, add equal volume of bead buffer
2. Remove liquid with magnet, add equal volume of 20mM NaOH
3. Remove liquid with magnet, add equal volume of bead buffer
4. Remove liquid with magnet, add ***half*** volume of Formamide (95% formamide, 10 mM EDTA)
(If you started with 200 μL of reaction, and 100 μL of beads, you're adding 50 μL of formamide)
-Place formamide-bead tubes on rotator for 5-10 minutes
-Place formamide-bead tubes on heat block (65°) for 5-10 minutes
-Sonicate formamide-bead tubes, then vortex and spin down. Place them on the magnetic rack. The solution will not go completely clear, but wait at least 2 minutes.
-Remove formamide solution from beads and add it to a prepped P30 spin column (don't push the pipet down past the stop; we don't need to get all the volume here, if there's a little left behind that's okay, we're trying to avoid pipetting beads)
-Use spin column to exchange formamide (it'll elute in Tris buffer)

**<u>Step 5:</u>** Reverse Transcription
Procedure:
-Prepare reverse transcription reactions
(per reaction: I usually use 25 μL eluted-RNA, +1 μL of each primer from 100μM stock, and +1 μL of dNTPs stock)
-Run reverse transcription protocol on these reactions (it should run at 65°C for 5 minutes to anneal primers, then hold at 4° C)
-Add to each tube: buffer, DTT, super-ase-in, and enzyme
(per reaction: I usually use 7 μL of 5x first-strand buffer, 1 μL DTT, 1 μL SUPERase-in, 1 μL Superscript III reverse-transcriptase)
-Put reverse transcription tubes back into thermocycler, hit next step to continue the protocol (I use the standard superscript III protocol but have extended the reaction step to 45 minutes)

Step 6: PCR
**Prep** (you can do this while the RT reaction is still running):
-Wipe down everything on bench, including pipets
-Prepare for 500 μL PCR reaction, adding to a tube as follows:
-100 μL 5x HF buffer
-5 μL of 100μM FP and RP (this is the forward & reverse primer I only use for PCR, so they're likely to be "cleaner"
-5 μL of dNTP stock
-5 μL Phusion primer
-300 μL of water
-80 μL of reverse transcription product
Procedure:
-Aliquot 50 μL of PCR into a "test PCR" tube
-Run test PCR for ~18 rounds; take a 5μL aliquot out during the annealing step of each round 7, 9, 11, 13, 15, 17, and then after round 18 (taking out DNA during the annealing step of round 7 will tell you how much DNA was produced by 6 rounds, etc.)
-Run these samples (corresponding to round 6, 8, 10, 12, 14, 16, 18) on a gel, stain, and photograph

-Determine how many rounds PCR should be run for
-Aliquot PCR mix into five PCR tubes (~100 µL/tube) and run for the right number of cycles

**Step 7:** Finishing up
-Run Qiagen kit cleanup of PCR
(Add 5x PB buffer for ~3 mL of Qiagen sample; you can run two Qiagen columns, loading each one twice with 750µL, before proceeding to the normal PE wash step—Qiagen columns can be run with multiple PB buffer aliquots before you add the second/third buffers)
(I recommend eluting each Qiagen column into 20 µL of EB buffer—instead of 15 µL twice—, then vortexing these tubes well, combining, and reading their concentration)
-Make new RNA!

Pathfinding and k-Seq scripts: Full documentation, and the scripts themselves, can be found at the Chen lab github, https://github.com/ichen-lab-ucsb/SCAPE-BYO

# References

1 Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181-186 (2015).

2 Aggarwal, R. S. What's fueling the biotech engine-2012 to 2013. *Nat Biotechnol* **32**, 32-39, doi:10.1038/nbt.2794 (2014).

3 Sundaram, P., Kurniawan, H., Byrne, M. E. & Wower, J. Therapeutic RNA aptamers in clinical trials. *Eur J Pharm Sci* **48**, 259-271, doi:10.1016/j.ejps.2012.10.014 (2013).

4 Jayasena, S. D. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. *Clin Chem* **45**, 1628-1650 (1999).

5 Balke, D., Wichert, C., Appel, B. & Muller, S. Generation and selection of ribozyme variants with potential application in protein engineering and synthetic biology. *Appl Microbiol Biotechnol* **98**, 3389-3399, doi:10.1007/s00253-014-5528-7 (2014).

6 Agresti, J. J., Kelly, B. T., Jaschke, A. & Griffiths, A. D. Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization. *Proc Natl Acad Sci U S A* **102**, 16170-16175, doi:10.1073/pnas.0503733102 (2005).

7 Illangasekare, M. & Yarus, M. Small-molecule-substrate interactions with a self-aminoacylating ribozyme. *J Mol Biol* **268**, 631-639, doi:10.1006/jmbi.1997.0988 (1997).

8 Murakami, H., Saito, H. & Suga, H. A versatile tRNA aminoacylation catalyst based on RNA. *Chem Biol* **10**, 655-662, doi:10.1016/S1074-5521(03)00145-5 (2003).

9 Schimmel, P. & Henderson, B. Possible role of aminoacyl-RNA complexes in noncoded peptide synthesis and origin of coded synthesis. *Proc Natl Acad Sci U S A* **91**, 11283-11286 (1994).

10 Kobori, S., Nomura, Y., Miu, A. & Yokobayashi, Y. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res* **43**, e85, doi:10.1093/nar/gkv265 (2015).

11 Thiel, W. H. *et al.* Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One* **7**, e43836, doi:10.1371/journal.pone.0043836 (2012).

12 Hou, H. *et al.* Aptamer-based cantilever array sensors for oxytetracycline detection. *Anal Chem* **85**, 2010-2014, doi:10.1021/ac3037574 (2013).

13 Ferguson, B. S. *et al.* Genetic analysis of H1N1 influenza virus from throat swab samples in a microfluidic system for point-of-care diagnostics. *J Am Chem Soc* **133**, 9129-9135, doi:10.1021/ja203981w (2011).

14 Swensen, J. S. *et al.* Continuous, real-time monitoring of cocaine in undiluted blood serum via a microfluidic, electrochemical aptamer-based sensor. *J Am Chem Soc* **131**, 4262-4266, doi:10.1021/ja806531z (2009).

15 Song, K. M., Lee, S. & Ban, C. Aptamers and their biological applications. *Sensors (Basel)* **12**, 612-631, doi:10.3390/s120100612 (2012).

16 White, R. R., Sullenger, B. A. & Rusconi, C. P. Developing aptamers into therapeutics. *J Clin Invest* **106**, 929-934, doi:10.1172/JCI11325 (2000).

17 Peer, D. *et al.* Nanocarriers as an emerging platform for cancer therapy. *Nat Nanotechnol* **2**, 751-760, doi:10.1038/nnano.2007.387 (2007).

18 Kamaly, N., Xiao, Z., Valencia, P. M., Radovic-Moreno, A. F. & Farokhzad, O. C. Targeted polymeric therapeutic nanoparticles: design, development and clinical translation. *Chem Soc Rev* **41**, 2971-3010, doi:10.1039/c2cs15344k (2012).

19 Li, N., Nguyen, H. H., Byrom, M. & Ellington, A. D. Inhibition of cell proliferation by an anti-EGFR aptamer. *PLoS One* **6**, e20299, doi:10.1371/journal.pone.0020299 (2011).

20 Topp, S. & Gallivan, J. P. Random walks to synthetic riboswitches--a high-throughput selection based on cell motility. *Chembiochem* **9**, 210-213, doi:10.1002/cbic.200700546 (2008).

21 Murakami, H., Ohta, A., Ashigai, H. & Suga, H. A highly flexible tRNA acylation method for non-natural polypeptide synthesis. *Nat Methods* **3**, 357-359, doi:10.1038/nmeth877 (2006).

22 Niwa, N., Yamagishi, Y., Murakami, H. & Suga, H. A flexizyme that selectively charges amino acids activated by a water-friendly leaving group. *Bioorg Med Chem Lett* **19**, 3892-3894, doi:10.1016/j.bmcl.2009.03.114 (2009).

23 Huang, Z. & Szostak, J. W. Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. *RNA* **9**, 1456-1463, doi:10.1261/rna.5990203 (2003).

24 Knight, C. G. *et al.* Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Res* **37**, e6, doi:10.1093/nar/gkn899 (2009).

25 Petrie, K. L. & Joyce, G. F. Limits of neutral drift: lessons from the in vitro evolution of two ribozymes. *J Mol Evol* **79**, 75-90, doi:10.1007/s00239-014-9642-z (2014).

26 Saito, H., Kourouklis, D. & Suga, H. An in vitro evolved precursor tRNA with aminoacylation activity. *EMBO J* **20**, 1797-1806, doi:10.1093/emboj/20.7.1797 (2001).

27 de Visser, J. A. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* **15**, 480-490, doi:10.1038/nrg3744 (2014).

28 Deris, J. B. *et al.* The innate growth bistability and fitness landscapes of antibiotic-resistant bacteria. *Science* **342**, 1237435 (2013).

29 Wedge, D. C., Rowe, W., Kell, D. B. & Knowles, J. In silico modelling of directed evolution: Implications for experimental design and stepwise evolution. *J Theor Biol* **257**, 131-141, doi:10.1016/j.jtbi.2008.11.005 (2009).

30 Savory, N. *et al.* In silico maturation of binding-specificity of DNA aptamers against Proteus mirabilis. *Biotechnol Bioeng* **110**, 2573-2580, doi:10.1002/bit.24922 (2013).

31 Breitling, F., Dübel, S., Seehaus, T., Klewinghaus, I. & Little, M. A surface expression vector for antibody screening. *Gene* **104**, 147-153 (1991).

32 Brekke, O. H. & Sandlie, I. Therapeutic antibodies for human diseases at the dawn of the twenty-first century. *Nature reviews Drug discovery* **2**, 52 (2003).

33 Hoppe-Seyler, F., Crnkovic-Mertens, I., Tomai, E. & Butz, K. Peptide aptamers: specific inhibitors of protein function. *Current molecular medicine* **4**, 529-538 (2004).

34 Arnold, F. H. & Georgiou, G. *Directed enzyme evolution: screening and selection methods*. (Springer Science & Business Media, 2003).

35 Arnold, F. H. & Volkov, A. A. Directed evolution of biocatalysts. *Current opinion in chemical biology* **3**, 54-59 (1999).

36 Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX—a (r) evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular engineering* **24**, 381-403 (2007).

37 Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).

38 Szostak, J. W., Bartel, D. P. & Luisi, P. L. Synthesizing life. *Nature* **409**, 387-390, doi:10.1038/35053176 (2001).

39 Klug, S. J. & Famulok, M. All you wanted to know about SELEX. *Molecular biology reports* **20**, 97-107 (1994).

40 Chumachenko, N. V., Novikov, Y. & Yarus, M. Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J Am Chem Soc* **131**, 5257-5263, doi:10.1021/ja809419f (2009).

41 Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & De Visser, J. A. G. Quantitative analyses of empirical fitness landscapes. *J Stat Mech-Theory E* **2013**, P01005 (2013).

42 Pitt, J. N. & Ferre-D'Amare, A. R. Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376-379, doi:10.1126/science.1192001 (2010).

43 Jimenez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R. & Chen, I. A. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc Natl Acad Sci U S A* **110**, 14984-14989, doi:10.1073/pnas.1307604110 (2013).

44 Wang, J., Rudzinski, J. F., Gong, Q., Soh, H. T. & Atzberger, P. J. Influence of target concentration and background binding on in vitro selection of affinity reagents. *PLoS One* **7**, e43940, doi:10.1371/journal.pone.0043940 (2012).

45 Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* **1**, 356--366 (1932).

46 Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563-564 (1970).

47 Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**, 11-45 (1987).

48 Schutze, T. *et al.* Probing the SELEX process with next-generation sequencing. *PLoS One* **6**, e29604, doi:10.1371/journal.pone.0029604 (2011).

49 Rowe, W., Wedge, D. C., Platt, M., Kell, D. B. & Knowles, J. Predictive models for population performance on real biological fitness landscapes. *Bioinformatics* **26**, 2145-2152, doi:10.1093/bioinformatics/btq353 (2010).

50 Curtis, E. A. & Bartel, D. P. Synthetic shuffling and in vitro selection reveal the rugged adaptive fitness landscape of a kinase ribozyme. *RNA* **19**, 1116-1128, doi:10.1261/rna.037572.112 (2013).

51 Sanchez-Luque, F. J., Stich, M., Manrubia, S., Briones, C. & Berzal-Herranz, A. Efficient HIV-1 inhibition by a 16 nt-long RNA aptamer designed by combining in vitro selection and in silico optimisation strategies. *Sci Rep* **4**, 6242, doi:10.1038/srep06242 (2014).

52 Katilius, E., Flores, C. & Woodbury, N. W. Exploring the sequence space of a DNA aptamer using microarrays. *Nucleic Acids Res* **35**, 7626-7635, doi:10.1093/nar/gkm922 (2007).

53 Fischer, N. O., Tok, J. B. H. & Tarasow, T. M. Massively Parallel Interrogation of Aptamer Sequence, Structure and Function. *PLoS ONE* **3**, e2720-e2720, doi:10.1371/journal.pone.0002720 (2008).

54 Rowe, W. *et al.* Analysis of a complete DNA-protein affinity landscape. *J R Soc Interface* **7**, 397-408, doi:10.1098/rsif.2009.0193 (2010).

55 Athavale, S. S., Spicer, B. & Chen, I. A. Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Current opinion in chemical biology* **22**, 35-39 (2014).

56 Otwinowski, J. & Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences* **111**, E2301-E2309 (2014).

57 Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, doi:10.7554/eLife.16965 (2016).

58 Pressman, A., Moretti, J. E., Campbell, G. W., Muller, U. F. & Chen, I. A. Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res* **45**, 10922, doi:10.1093/nar/gkx816 (2017).

59 Kobori, S. & Yokobayashi, Y. High-Throughput Mutational Analysis of a Twister Ribozyme. *Angewandte Chemie* (2016).

60 Dhamodharan, V., Kobori, S. & Yokobayashi, Y. Large Scale Mutational and Kinetic Analysis of a Self-Hydrolyzing Deoxyribozyme. *ACS Chem Biol* **12**, 2940-2945, doi:10.1021/acschembio.7b00621 (2017).

61 Phillips, A. M. *et al.* Host proteostasis modulates influenza evolution. *Elife* **6**, doi:10.7554/eLife.28652 (2017).

62 Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801-807, doi:10.1038/nmeth.3027 (2014).

63 Starita, L. M. & Fields, S. Deep Mutational Scanning: A Highly Parallel Method to Measure the Effects of Mutation on Protein Function. *Cold Spring Harb Protoc* **2015**, 711-714, doi:10.1101/pdb.top077503 (2015).

64 Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat Methods* **7**, 741-746, doi:10.1038/nmeth.1492 (2010).

65 Crick, F. H. The origin of the genetic code. *Journal of molecular biology* **38**, 367-379 (1968).

66 Orgel, L. E. Evolution of the genetic apparatus. *Journal of molecular biology* **38**, 381-393 (1968).

67 Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M. & Saxinger, W. C. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor symposia on quantitative biology* **31**, 723-736 (1966).

68 Athavale, S. S., Spicer, B. & Chen, I. A. Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Curr Opin Chem Biol* **22**, 35-39, doi:10.1016/j.cbpa.2014.09.008 (2014).

69 Sczepanski, J. T. & Joyce, G. F. A cross-chiral RNA polymerase ribozyme. *Nature* **515**, 440-442, doi:10.1038/nature13900 (2014).

70 Murakami, H., Kourouklis, D. & Suga, H. Using a solid-phase ribozyme aminoacylation system to reprogram the genetic code. *Chemistry & Biology* **10**, 1077-1084, doi:DOI 10.1016/j.chembiol.2003.10.010 (2003).

71 Saito, H., Watanabe, K. & Suga, H. Concurrent molecular recognition of the amino acid and tRNA by a ribozyme. *RNA* **7**, 1867-1878 (2001).

72 Bessho, Y., Hodgson, D. R. & Suga, H. A tRNA aminoacylation system for non-natural amino acids based on a programmable ribozyme. *Nat Biotechnol* **20**, 723-728, doi:10.1038/nbt0702-723 (2002).

73 Turk, R. M., Chumachenko, N. V. & Yarus, M. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci U S A* **107**, 4585-4589, doi:10.1073/pnas.0912895107 (2010).

74 Noller, H. F. The driving force for molecular evolution of translation. *RNA* **10**, 1833-1837, doi:10.1261/rna.7142404 (2004).

75 Szathmary, E. & Maynard Smith, J. From replicators to reproducers: the first major transitions leading to life. *J Theor Biol* **187**, 555-571, doi:10.1006/jtbi.1996.0389 (1997).

76 Wong, J. T. Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. *Orig Life Evol Biosph* **21**, 165-176, doi:10.1007/BF01809445 (1991).

77 Pressman, A., Blanco, C. & Chen, I. A. The RNA World as a model system to study the origin of life. *Curr Biol* **25**, R953-R963 (2015).

78 Di Giulio, M. The origin of the tRNA molecule: implications for the origin of protein synthesis. *J Theor Biol* **226**, 89-93, doi:10.1016/j.jtbi.2003.07.001 (2004).

79 Di Giulio, M. A polyphyletic model for the origin of tRNAs has more support than a monophyletic model. *J Theor Biol* **318**, 124-128, doi:10.1016/j.jtbi.2012.11.012 (2013).

80 Widmann, J., Di Giulio, M., Yarus, M. & Knight, R. tRNA creation by hairpin duplication. *J Mol Evol* **61**, 524-530, doi:10.1007/s00239-004-0315-1 (2005).

81 Illangasekare, M. & Yarus, M. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA* **5**, 1482-1489 (1999).

82 Lacey, J. C., Jr., Senaratne, N. & Mullins, D. W., Jr. Hydrolytic properties of phenylalanyl- and N-acetylphenylalanyl adenylate anhydrides. *Orig Life Evol Biosph* **15**, 45-54, doi:10.1007/BF01809392 (1984).

83 Liu, Z., Beaufils, D., Rossi, J. C. & Pascal, R. Evolutionary importance of the intramolecular pathways of hydrolysis of phosphate ester mixed anhydrides with amino acids and peptides. *Sci Rep* **4**, 7440, doi:10.1038/srep07440 (2014).

84 Darwin, C. ORIGIN OF SPECIES. *The Athenaeum*, 861-861 (1869).

85 Fisher, R. A. *The genetical theory of natural selection: a complete variorum edition*. (Oxford University Press, 1930).

86 Smerlak, M. & Youssef, A. Limiting fitness distributions in evolutionary dynamics. *Journal of theoretical biology* **416**, 68-80 (2017).

87 Ichihashi, N., Aita, T., Motooka, D., Nakamura, S. & Yomo, T. Periodic pattern of genetic and fitness diversity during evolution of an artificial cell-like system. *Molecular biology and evolution*, msv189 (2015).

88 Lessard, S. Fisher's fundamental theorem of natural selection revisited. *Theor Popul Biol* **52**, 119-136, doi:10.1006/tpbi.1997.1324 (1997).

89 Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature genetics* **38**, 484 (2006).

90 Joyce, P., Rokyta, D. R., Beisel, C. J. & Orr, H. A. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics* **180**, 1627-1643, doi:10.1534/genetics.108.088716 (2008).

91 Carothers, J. M., Oestreich, S. C., Davis, J. H. & Szostak, J. W. Informational complexity and functional activity of RNA structures. *J Am Chem Soc* **126**, 5130-5137, doi:10.1021/ja031504a (2004).

92 Vant-Hull, B., Gold, L. & Zichi, D. A. Theoretical principles of in vitro selection using combinatorial nucleic acid libraries. *Current Protocols in Nucleic acid Chemistry*, 9.1. 1-9.1. 16 (2000).

93 Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology* **193**, 723-743 (1987).

94 Tacker, M., Fontana, W., Stadler, P. F. & Schuster, P. Statistics of RNA melting kinetics. *Eur Biophys J* **23**, 29-38 (1994).

95 Schütze, T. *et al.* Probing the SELEX process with next-generation sequencing. *PLoS One* **6**, e29604 (2011).

96 Thiel, W. H. *et al.* Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One* **7**, e43836 (2012).

97 Hoinka, J. *et al.* Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res* **43**, 5699-5707, doi:10.1093/nar/gkv308 (2015).

98 Hayden, E. J., Bratulic, S., Koenig, I., Ferrada, E. & Wagner, A. The effects of stabilizing and directional selection on phenotypic and genotypic variation in a population of RNA enzymes. *Journal of molecular evolution* **78**, 101-108 (2014).

99 Moretti, J. E. & Muller, U. F. A ribozyme that triphosphorylates RNA 5'-hydroxyl groups. *Nucleic Acids Res* **42**, 4767-4778, doi:10.1093/nar/gkt1405 (2014).

100 Petrie, K. L. & Joyce, G. F. Limits of neutral drift: Lessons from the in vitro evolution of two ribozymes. *Journal of molecular evolution* **79**, 75-90 (2014).

101 Slutsky, M. & Mirny, L. A. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical journal* **87**, 4021-4035 (2004).

102 Slutsky, M., Kardar, M. & Mirny, L. A. Diffusion in correlated random potentials, with applications to DNA. *Physical Review E* **69**, 061903 (2004).

103 Wolfsheimer, S. & Hartmann, A. Minimum-free-energy distribution of RNA secondary structures: Entropic and thermodynamic properties of rare events. *Physical Review E* **82**, 021902 (2010).

104 Bataillon, T. & Bailey, S. F. Effects of new mutations on fitness: insights from models and data. *Ann NY Acad Sci* **1320**, 76-92 (2014).

105 Ekland, E. H., Szostak, J. W. & Bartel, D. P. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364-370 (1995).

106 Jalali-Yazdi, F., Lai, L. H., Takahashi, T. T. & Roberts, R. W. High-Throughput Measurement of Binding Kinetics by mRNA Display and Next-Generation

Sequencing. *Angew Chem Int Ed Engl* **55**, 4007-4010, doi:10.1002/anie.201600077 (2016).

107 Illangasekare, M., Sanchez, G., Nickles, T. & Yarus, M. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* **267**, 643-647, doi:10.1126/science.7530860 (1995).

108 Segré, D., Ben-Eli, D., Deamer, D. W. & Lancet, D. The lipid world. *Origins Life Evol B* **31**, 119-145 (2001).

109 Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111-114, doi:10.1126/science.1123539 (2006).

110 Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLoS Comput Biol* **13**, e1005541, doi:10.1371/journal.pcbi.1005541 (2017).

111 Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol* **7**, e1002302 (2011).

112 Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**, 1165-1174 (2005).

113 Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci* **255**, 279-284, doi:10.1098/rspb.1994.0040 (1994).

114 Gavrilets, S. *Fitness landscapes and the origin of species (MPB-41)*. Vol. 41 (Princeton University Press, 2004).

115 Xulvi-Brunet, R., Campbell, G. W., Rajamani, S., Jiménez, J. I. & Chen, I. A. Computational analysis of fitness landscapes and evolutionary networks from in vitro evolution experiments. *Methods* **106**, 86-96 (2016).

116 Anderson, P. W. Suggested model for prebiotic evolution: the use of chaos. *Proc Natl Acad Sci U S A* **80**, 3386-3390 (1983).

117 Amitrano, C., Peliti, L. & Saber, M. Population dynamics in a spin-glass model of chemical evolution. *J Mol Evol* **29**, 513-525 (1989).

118 Bogdanova, E. A. *et al.* Normalizing cDNA libraries. *Current protocols in molecular biology* **90**, 5.12. 11-15.12. 27 (2010).

119 Soares, M. B. *et al.* Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences* **91**, 9228-9232 (1994).

120 Weinreich, D. M., Lan, Y., Jaffe, J. & Heckendorn, R. B. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics*, doi:10.1007/s10955-018-1975-3 (2018).

121 Aita, T. & Husimi, Y. Adaptive Walks by the Fittest among Finite Random Mutants on a Mt. Fuji-type Fitness Landscape. *J Theor Biol* **193**, 383-405, doi:10.1006/jtbi.1998.0709 (1998).

122 Aita, T. & Husimi, Y. Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. *J Theor Biol* **182**, 469-485, doi:10.1006/jtbi.1996.0189 (1996).

123 Ferretti, L. *et al.* Measuring epistasis in fitness landscapes: the correlation of fitness effects of mutations. *J Theor Biol* **396**, 132-143 (2016).

124 Wunsche, A. *et al.* Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. *Nat Ecol Evol* **1**, 61, doi:10.1038/s41559-016-0061 (2017).

125 Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J Theor Biol* **141**, 211-245 (1989).

126 Puchta, O. *et al.* Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840-844 (2016).

127 Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932 (2006).

128 Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837-840 (2016).

129 Saha, R., Pohorille, A. & Chen, I. A. Molecular crowding and early evolution. *Orig Life Evol Biosph* **44**, 319-324, doi:10.1007/s11084-014-9392-3 (2015).

130 Rivas, G. & Minton, A. P. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* **41**, 970-981, doi:https://doi.org/10.1016/j.tibs.2016.08.013 (2016).

131 Mimi, G. *et al.* Crowders and Cosolvents—Major Contributors to the Cellular Milieu and Efficient Means to Counteract Environmental Stresses. *ChemPhysChem* **18**, 2951-2972, doi:doi:10.1002/cphc.201700762 (2017).

132 Saha, R., Verbanic, S. & Chen, I. A. Lipid vesicles chaperone an encapsulated RNA aptamer. *Nature Communications* **9**, 2313, doi:10.1038/s41467-018-04783-8 (2018).

133 Daher, M., Widom, J. R., Tay, W. & Walter, N. G. Soft interactions with model crowders and non-canonical interactions with cellular proteins stabilize RNA folding. *J. Mol. Biol.* **430**, 509-523, doi:https://doi.org/10.1016/j.jmb.2017.10.030 (2018).

134 Lee, H.-T., Kilburn, D., Behrouzi, R., Briber, R. M. & Woodson, S. A. Molecular crowding overcomes the destabilizing effects of mutations in a bacterial ribozyme. *Nucleic Acids Res.* **43**, 1170-1176, doi:10.1093/nar/gku1335 (2015).

135 Domingo, J., Diss, G. & Lehner, B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* **558**, 117-121, doi:10.1038/s41586-018-0170-7 (2018).

136 Baird, N. J., Inglese, J. & Ferré-D'Amaré, A. R. Rapid RNA–ligand interaction analysis through high-information content conformational and stability landscapes. *Nature Communications* **6**, 8898-8898, doi:10.1038/ncomms9898 (2015).

137 Carothers, J. M., Oestreich, S. C. & Szostak, J. W. Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J Am Chem Soc* **128**, 7929-7937, doi:10.1021/ja060952q (2006).

138 Smith, D., Collins, B. D., Heil, J. & Koch, T. H. Sensitivity and specificity of photoaptamer probes. *Mol Cell Proteomics* **2**, 11-18 (2003).

139 Guo, W. M. *et al.* Identification and Characterization of an eIF4e DNA Aptamer That Inhibits Proliferation With High Throughput Sequencing. *Mol Ther Nucleic Acids* **3**, e217, doi:10.1038/mtna.2014.70 (2014).

140 Liu, Z. *et al.* 5 (4H)-Oxazolones as Effective Aminoacylation Reagents for the 3′-Terminus of RNA. *Synlett* **28**, 73-77 (2017).

141 Liu, Z., Rigger, L., Rossi, J. C., Sutherland, J. D. & Pascal, R. Mixed Anhydride Intermediates in the Reaction of 5(4H)-Oxazolones with Phosphate Esters and Nucleotides. *Chemistry* **22**, 14940-14949, doi:10.1002/chem.201602697 (2016).