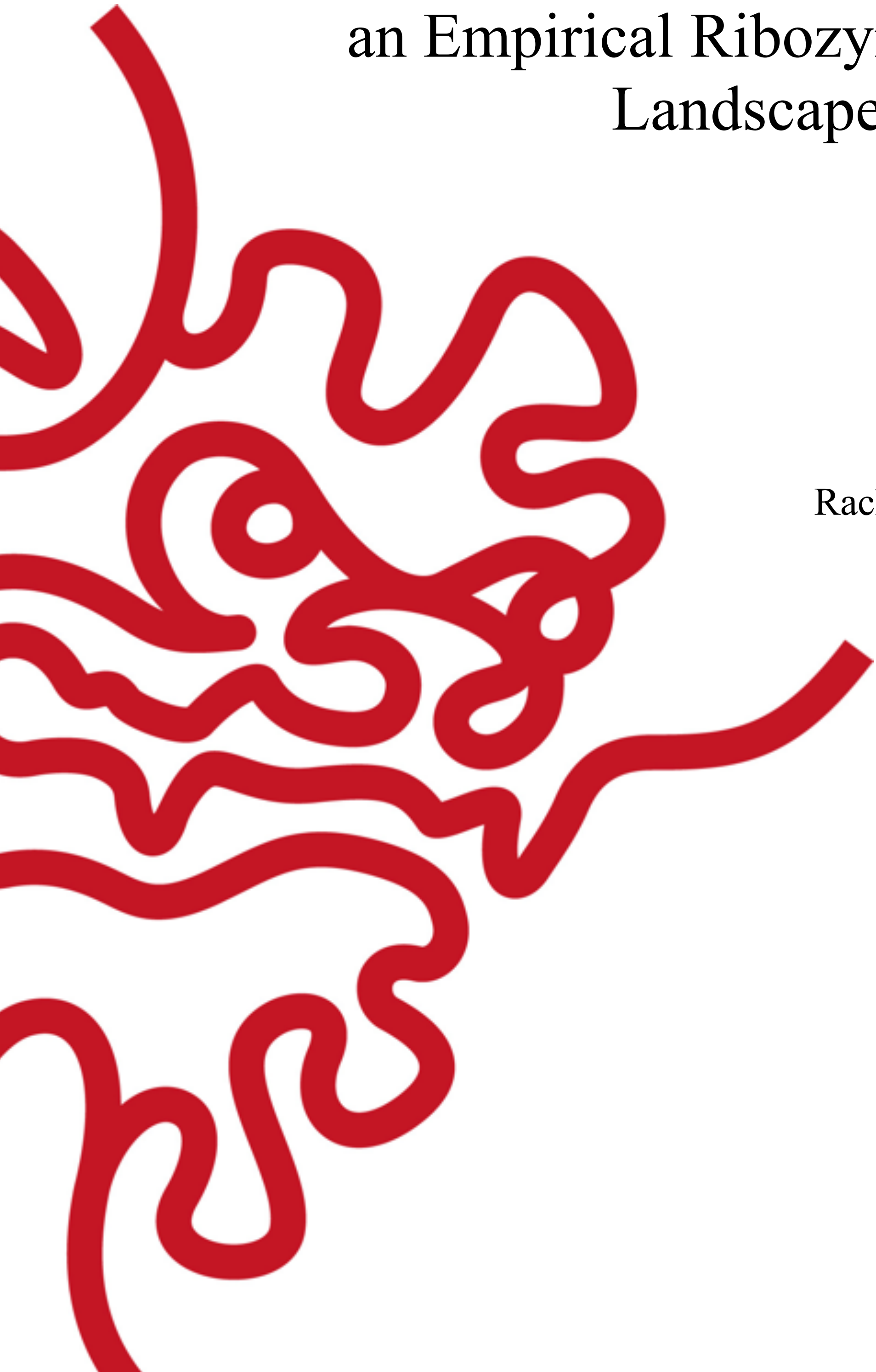
Okinawa Institute of Science and Technology
Graduate University

Thesis submitted for the degree
Doctor of Philosophy

Machine Learning Guided Exploration of an Empirical Ribozyme Fitness Landscape



by

Rachapun Rotrattanadumrong

under the supervision of

Prof. Yohei Yokobayashi

August 2023

Declaration of Original and Sole Authorship

I, Rachapun Rotrattanadumrong, declare that this thesis entitled “Machine learning guided exploration of an empirical ribozyme fitness landscape” and the data presented in it are original and my own work.

I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published and pre-printed elsewhere, with the exception of the following:

  Rotrattanadumrong, R. and Yokobayashi, Y. (2022) ‘Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning’, Nature communications, 13(1), p. 4847.
  (In this work, I designed all experiments and computational models, perform all data collection and analysis, all model training and evaluation and wrote the manuscript. Yohei Yokobayashi supervised the experimental design, evaluate the data analysis and edited the manuscript.)

Date: 18th August 2023

Signature:

Rachapun

Rachapun Rotrattanadumrong

Abstract

Fitness landscape of a biomolecule is a representation of its activity as a function of its sequence. Properties of a fitness landscape determine how evolution proceeds. Therefore, the distribution of functional variants and more importantly, the connectivity of these variants within the sequence space are important scientific questions. Exploration of these spaces, however, is impeded by the combinatorial explosion of the sequence space. High-throughput experimental methods have recently reduced this impediment but only modestly. Better computational methods are needed to fully utilize the rich information from these experimental data to better understand the properties of the fitness landscape. In this work, I seek to improve this exploration process by combining data from massively parallel experimental assay with smart library design using advanced computational techniques. I focus on an artificial RNA enzyme or ribozyme that can catalyze a ligation reaction between two RNA fragments. This chemistry is analogous to that of the modern RNA polymerase enzymes, therefore, represents an important reaction in the origin of life. In the first chapter, I discuss the background to this work in the context of evolutionary theory of fitness landscape and its implications in biotechnology. In chapter 2, I explore the use of processes borrowed from the field of evolutionary computation to solve optimization problems using real experimental sequence-activity data. In chapter 3, I investigate the use of supervised machine learning models to extract information on epistatic interactions from the dataset collected during multiple rounds of directed evolution. I investigate and experimentally validate the extent to which a deep learning model can be used to guide a completely computational evolutionary algorithm towards distant regions of the fitness landscape. In the final chapter, I perform a comprehensive experimental assay of the combinatorial region explored by the deep learning-guided evolutionary algorithm. Using this dataset, I analyze higher-order epistasis and attempt to explain the increased predictability of the region sampled by the algorithm. Finally, I provide the first experimental evidence of a large RNA 'neutral network'. Altogether, this work represents the most comprehensive experimental and computational study of the RNA ligase ribozyme fitness landscape to date, providing important insights into the evolutionary search space possibly explored during the earliest stages of life.

Acknowledgement

I would like to thank the current and past members of the Nucleic Acid Chemistry and Engineering Unit for the technical, academic and personal support over the past 5 years. I am grateful for the mentorship given by Professor Yokobayashi who has taught me a rigorous scientific approach which I hope to carry on for the rest of my research career. I would like to thank Yoko-san who, since my rotation days, has taught me all I need to know about good experimental methods. Finally, thank you to the OIST sequencing section for performing many of the high-throughput sequencing that underpin the success of this work.

Table of Abbreviations

cDNA	Complementary DNA
CNN	Convolutional neural network
DNN	Deep neural network
dsDNA	Double stranded DNA
EA	Evolutionary algorithms
EC	Evolutionary computation
EDTA	Ethylenediaminetetraacetic acid
FL	Fraction ligated
GA	Genetic algorithm
GBDT	Gradient-boosted decision trees
GP	Gaussian process
GTP	Guanosine triphosphate
HDV	Hepatitis delta virus
k-NN	k-Nearest Neighbors
LR	Logistic/Linear regression
LSTM	Long short-term memory
MFE	Minimum free energy
ML	Machine learning
MLP	Multilayer perceptron
NEB	New England Biolabs
nt	Nucleotide
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
QS	Quality score
RA	Relative activity
SVM	Support vector machine
WT	Wildtype

Dedication

For my family, to whom I owe everything.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Fitness landscape was first introduced as a way to visualize evolutionary processes (Wright, 1932) by displaying organismal reproductive success (fitness) as a function of its genotype space. This landscape is often presented with the genotype space in the horizontal plane while the vertical axis represents the fitness value (Figure 1-1). In this representation, evolution is a hill climbing process where natural selection moves the genotype towards regions of higher fitness or peaks in the landscape. This fitness landscape metaphor has been used to study many fundamental concepts of evolution including adaptation and innovation. Understanding the topology of the fitness landscape can lead to better understanding of how evolution can proceed. Mapping the topography of a fitness landscape can be done by assigning fitness value to every genotype in the landscape. However, this has proven difficult as the number of possible genotypes is astronomical (4^L for nucleotides and 20^L for proteins with L being the length of sequence). Furthermore, measurements of fitness are experimentally challenging for many systems, and many do not have high-throughput methods to do so. RNA emerged as an important system for large scale study of the fitness landscape (Pitt and Ferré-D'Amaré, 2010). The discovery of catalytic RNA or ribozyme reveals that RNA can both carry genotypic (nucleotide sequence) and phenotypic (catalytic activity) information (Kruger *et al.*, 1982). At the same time, computational algorithms were being developed that can successfully predict secondary structure from RNA sequences (Zuker and Stiegler, 1981). Under the sequence-structure-function assumption, a new paradigm emerges. The study of the fitness landscape of the RNA sequence-structure map where secondary structure was used as proxy for fitness (Figure 1-1).

Schuster and Fontana pioneered the use of RNA sequence-structure map to study evolution. Specifically they discovered that RNA sequence space contains extensive neutral network (Schuster *et al.*, 1997). A neutral network is a group of genotypes within the sequence space that share the same phenotype and are connected to each other via single step mutations that maintain the phenotype (Figure 1-1). In the evolutionary hill climbing metaphor, neutral networks represent smooth ridges or plateaus that connect multiple fitness peaks allowing evolution to travel large distances within the fitness landscape without interruption. Using an inverse folding algorithm, Schuster and Fontana discovered that many neighboring RNA sequences are predicted to fold into the same structure forming a percolating neutral network within the RNA fitness landscape. This landmark discovery led to many subsequent studies that used RNA sequence-structure maps and the concept of neutral networks to better understand how evolutionary adaptation and innovation can happen (van Nimwegen, Crutchfield and Huynen, 1999; Kun, Santos and Szathmáry, 2005; Greenbury, Louis and Ahnert, 2022; Johnston *et al.*, 2022).

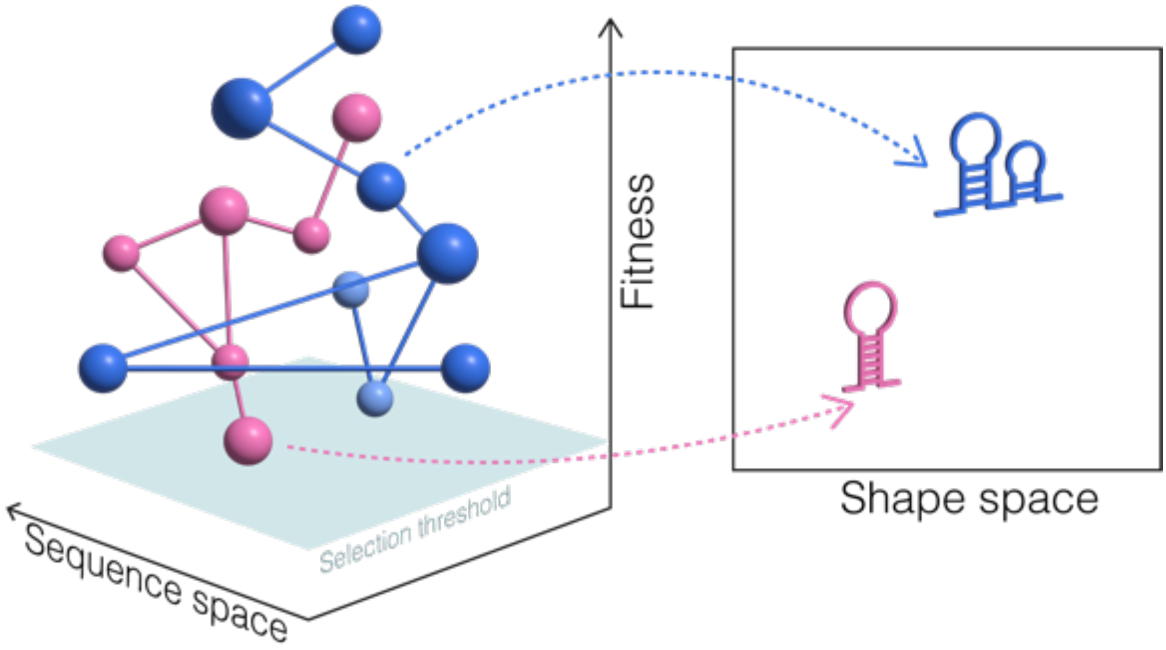


Figure 1-1: Visualization of RNA fitness landscape and neutral network. The RNA sequence space is the total possible combination of bases for a given length of sequence (4^L). The sequence space can be represented as a function of its fitness forming a fitness landscape. Many properties such as catalytic activity or ligand binding can be used as a proxy for fitness. Structure prediction algorithm suggests that many sequences in the fitness landscape that are connected by single mutations can be mapped onto the same secondary structure. This many-to-one mapping between the sequence space and the shape space resulted in a large neutral network that could facilitate evolution. In this schematics, neutral genotypes are represented by nodes with edges representing single mutations connecting each genotype.

With the recent advancement in DNA synthesis and sequencing technology, an increasing number of studies has been done towards empirical mapping of RNA fitness landscape. Two most notable attempts were made by the Chen group, where they map almost the entire sequence space of aminoacylating ribozyme (Pressman *et al.*, 2019) and a GTP binding RNA aptamer (Jiménez *et al.*, 2013). Both studies reveal a highly sparse landscape with isolated fitness peaks separated by extensive regions of inactivity or fitness valleys (Figure 1-2). In these landscapes, functional genotypes are rare and sparsely distributed. Furthermore, these functional genotypes are not connected by neutral mutational steps and large-scale neural networks, as predicted by theoretical studies, are absent. Studies of other RNA fitness landscapes have further supported these findings (Pitt and Ferré-D'Amaré, 2010; Hayden, Bendixsen and Wagner, 2015; Kobori and Yokobayashi, 2016; Li *et al.*, 2016; Domingo, Diss and Lehner, 2018; Bendixsen *et al.*, 2019). With all these findings two key questions emerge. Firstly, is there indeed any empirical evidence for the kind of neutral network predicted by the sequence-structure map. Secondly, how does natural or artificial evolutionary processes find functional genotypes during its navigation of the fitness landscape if neutral networks rare or completely absent.

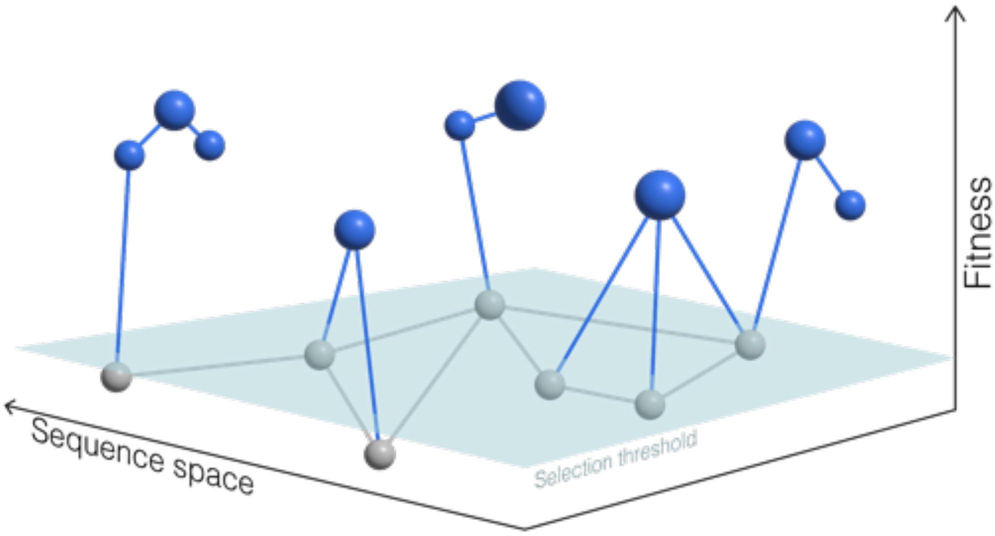


Figure 1-2: Visualization of sparse fitness landscape.
Recent experimental evidence suggests that many fitness landscapes have isolated fitness peaks where 'fit' genotypes are surrounded by regions of 'unfit' genotypes. Here, blue nodes represent 'fit' genotypes that can survive the selection threshold. To travel between any 'fit' nodes, evolution needs to pass by 'unfit' nodes (gray) which can lead to evolutionary constraint.

In this thesis, I present a collection of experimental and computational works that enable efficient identification of functional genotypes within the fitness landscapes of a small, artificial ligase ribozyme (Figure 1-3). Combining high-throughput experimental assay and advanced machine learning, this work culminates in the first empirical evidence of a large-scale RNA neutral network.

In chapter 2, I describe how, inspired by the concept of evolutionary algorithms in computer science, I used computational genetic processes to design libraries of ligase ribozyme sequences that are enriched in activity. This precise design process was enabled by the ability to synthesize custom ribozyme libraries with on-chip DNA synthesis. Sequencing-based high-throughput experimental assay provides a direct measurement of activity for tens of thousands of ribozyme sequences in parallel. Combining the experimental assay and computational design, this approach enables efficient exploration of the functional sequence space of the ligase ribozyme.

In chapter 3, I saw an opportunity in the dataset generated from the work done in chapter 2 and was inspired by the recent success of deep learning in solving biological problems (Angermueller *et al.*, 2016). I used the collected sequence-activity data to train a group of supervised machine learning models that learn the underlying fitness function governing the ligase ribozyme fitness landscape. I developed an evolutionary algorithm guided by a deep learning model to perform computational navigation of the ligase ribozyme fitness landscape. Using high throughput assay I confirmed that the algorithm was indeed able to identify functional genotypes in distant regions of the fitness landscape.

In the final chapter, I analyzed one of the mutants discovered by the evolutionary algorithm in chapter 3 and revealed that the accumulation of neutral mutation leads to an increase in mutational robustness in one of the structural modules of the ribozyme. Experimental screening of the entire combinatorial space between this mutant and the wildtype reveals an extensive neutral network. Many direct mutational paths connect the two genotypes while maintaining ligase activity. This is the first experimental evidence of a large-scale RNA neutral network discovered thus far. Further analysis of the mutational interactions within the network suggests that this network could increase ribozyme's fitness landscape navigability and predictability.

Fitness landscape is an elegant theoretical visualization of the evolutionary search space of organisms and biomolecules. The study and investigation of the theoretical fitness landscape has led to better understanding of many fundamental properties of evolution. The recent ability to push this study into the experimental realm has unveiled even more questions and surprising properties of evolution. The proof-of-concept studies that I have presented here could serve as a starting point for the use of high-throughput experimental techniques combined with state-of-art computational methods for guided exploration of the evolutionary search space. This could lead to even more significant discoveries such as the elusive RNA neutral network that I have discovered here almost 30 years since its first prediction.

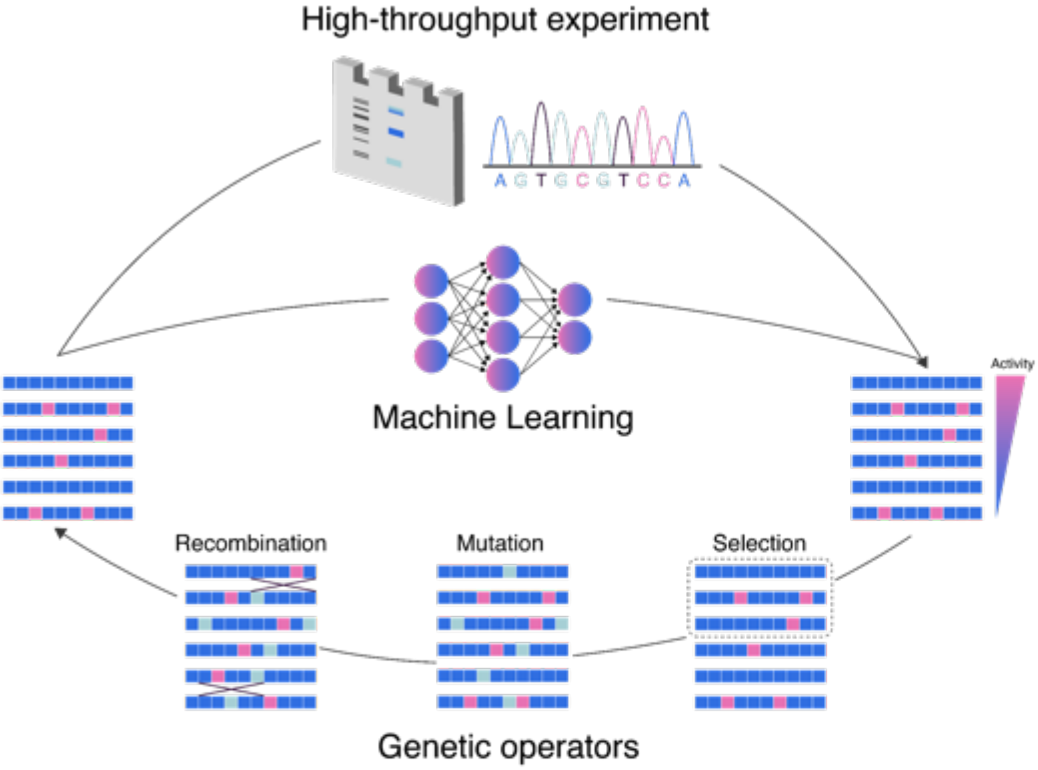


Figure 1-3: Overview of a hybrid evolutionary process combining computational model and high-throughput experiments.
Exploration of the RNA fitness landscape can be accelerated by combining sequence design using genetic operators and machine learning models trained on high-throughput experimental assay data. This approach can enable efficient crossing of fitness valleys towards new and higher fitness peaks.

Chapter 2: Identification of functional ribozymes with high-throughput experimental assay and genetic algorithm

Parts of this chapter, in particular the Methods section, have been duplicated or updated from a previously published article: Rotrattanadumrong, R. and Yokobayashi, Y. (2022) ‘Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning’, *Nature communications*, 13(1), p. 4847.

Background

In 1970, John Maynard Smith first proposed the idea of the protein sequence space as a network of discrete mutational paths for evolution to navigate (Maynard Smith, 1970). This extended the original idea of fitness landscape proposed by Wright (Wright, 1932) into the realm of molecular evolution. Smith's work introduces the idea that evolution proceeds on units of mutation and that natural selection occurs through a series of single mutations that maintain protein function. This idea of a discrete protein space transforms the study of fitness landscape into one of sequence-fitness maps. The discrete nature of the protein sequence space means the topography of the landscape can be understood by looking at how fitness changes along a mutational path. The accessibility of these mutational paths then becomes a major determinant of evolution. Early study of this idea was necessarily theoretical (Kauffman and Levin, 1987) due to limited experimental methods available to survey large regions of protein sequence space. This changes when Weinreich and co-workers show experimentally (Weinreich *et al.*, 2006) and theoretically (Weinreich, Watson and Chao, 2005) that evolutionary path along the protein sequence space can be severely limited by a particular form of mutational interaction or epistasis called 'sign epistasis. Epistasis occurs when effects of individual mutations combine in a non-additive way. Sign epistasis occurs when a mutation reverses the effect of another mutation (Figure 2-1a). In their seminal paper (Weinreich *et al.*, 2006), Weinreich and team experimentally showed that sign epistasis significantly reduces the number of accessible paths from a wildtype to a high-fitness β-lactamase enzyme (Figure 2-1b). They did so by synthesizing and evaluating all 2^L possible combinations of 5 mutations that give rise to the higher fitness variant. This enabled them to assess all 120 possible mutational paths between the two genotypes. The consequence of this work is twofold. First, it suggests that evolutionary paths are highly constrained which means that evolution could be highly reproducible and predictable. Second, it introduces the idea of studying fitness landscape and evolutionary paths through experimental construction of the combinatorial sequence space.

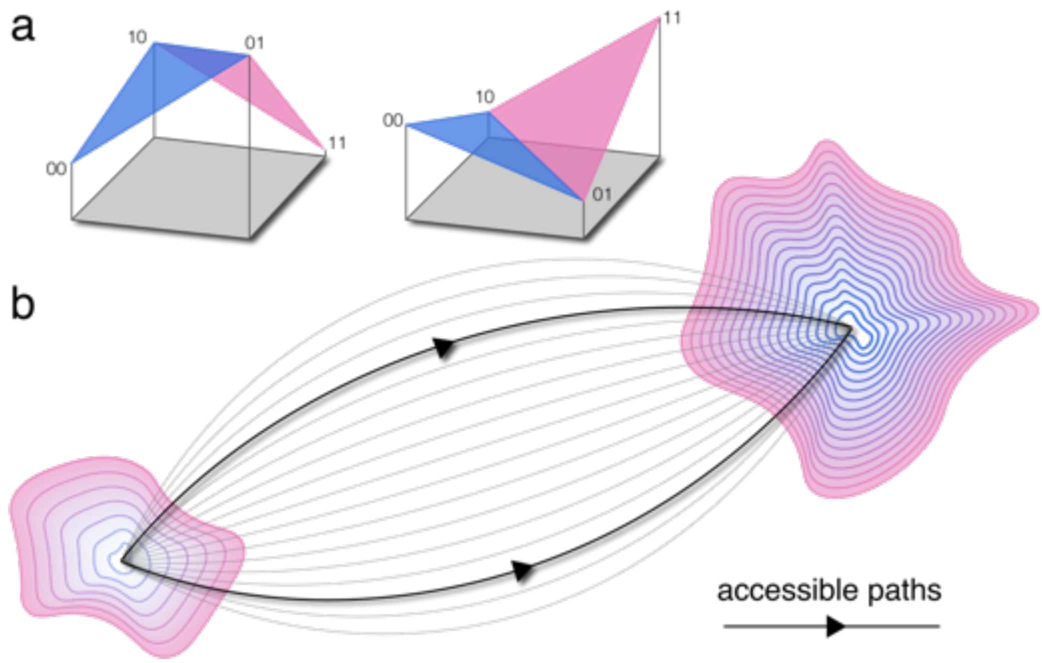


Figure 2-1: Sign epistasis can constrain evolutionary paths.
a) Pairwise reciprocal sign epistasis is observed when two genotypes differed by two mutations (00 and 11) both have higher or lower fitness than their two constituent single mutants (10 and 01) b) Accessible paths between different fitness peaks is defined as a smooth path involving mutational steps that either maintain or increase the fitness of the mutants relative to the previous step. Pairwise reciprocal sign epistasis can severely limit the number of accessible paths in the fitness landscape.

However, this method is experimentally challenging due to combinatorial explosion of the sequence space. In a combinatorial map, where only a specific set of mutations are considered, the total possible number of mutants is 2^L with L being the number of mutated positions. However, if the entire sequence space is to be considered where all possible substitution can occur then the total space becomes 4^L in the case of DNA or RNA (nucleotide sequences) or 20^L in the case of protein (amino acid sequences). In either case, the number of variants that needs to be assessed becomes experimentally intractable for any regular sized protein or oligonucleotide. However, this challenge has been alleviated somewhat with the advancement in DNA synthesis and sequencing. These technologies enable both generation of variants and fitness measurements in a rapid and high throughput way. It is now possible to obtain a large sequence-fitness dataset for many biomolecules. The first truly large-scale use of such a method to explore the biomolecular fitness landscape was not for protein but for an RNA enzyme. In 2010, Pitt and Ferré-D'Amaré constructed a fitness landscape consisting of ~10^7 variants of an RNA ligase ribozyme (Pitt and Ferré-D’Amaré, 2010). They did so through use of deep sequencing to monitor change in frequency of each mutant during selection (Figure 2-2). RNA was the ideal candidate for large scale fitness landscape mapping for many reasons. One, the sequence space of RNA is much smaller than protein (4^L vs 20^L). Two, the process of in vitro transcription and reverse transcription allow RNA mutants to be easily synthesized and directly analyzed through sequencing. RNA may not be the fundamental catalyst of the current life form as proteins are but their unique role in the evolution of life means that their fitness landscape can provide important understanding of the fundamental concept of evolution.

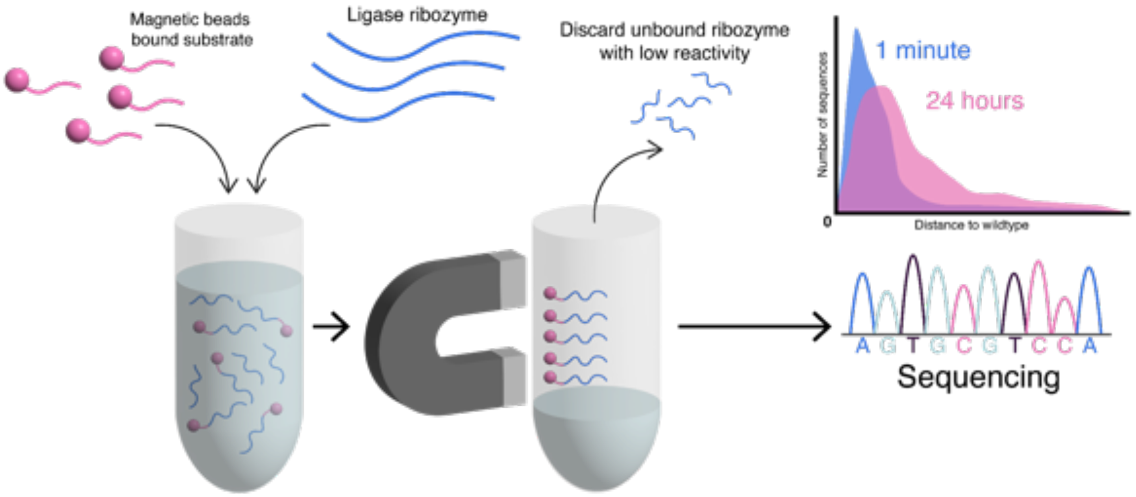


Figure 2-2: Schematic of select-and-sequence based high-throughput assay. Schematic of the deep sequencing based methods presented by (Pitt and Ferré-D'Amaré, 2010). Ligase ribozyme variants are mixed with magnetic bead bound substrates. Functional variants can self-ligate onto the substrate and can be selected using magnets allowing non-functional variants to be discarded. The frequency of each variant in the selected pool can be calculated by counting the number of reads from next-generation sequencing. Frequency of each variant can be compared from different reaction time points and change in frequency can be used as a proxy for fitness.

The term RNA world was first coined by Walter Gilbert (Gilbert, 1986). In this theory, RNA is suggested as the original biomolecules that sustained life prior to the emergence of DNA and proteins. The primordial organisms in the RNA world relied on RNA to both carry genetic information and to catalyze chemical reactions that are necessary to sustain life, including replication of genetic information. This concept was further strengthened by the work of Thomas Cech who reported for the first time that RNA can possess catalytic properties, specifically self-splicing (Cech, Zaug and Grabowski, 1981). Since then, RNA enzymes or ribozymes with a wide range of catalytic properties have been discovered in both natural and artificial settings. One particularly important discovery is the artificial creation of a ligase ribozyme. Bartel and Szotak first isolated a ribozyme capable of joining two RNA fragments together from a pool of random sequences (Bartel and Szostak, 1993). The ribozymes they discovered catalyze the ligation between a 3′-hydroxyl and 5′-triphosphate termini in a similar fashion to the modern RNA polymerase enzymes. The fact that this catalytic process, that is so essential to life, can be performed by RNA provides strong evidence that ligase ribozymes might be one of the first biological catalysts that emerged at the beginning of life. It is therefore fortuitous that a ligase ribozyme became the first biomolecule with a sequence space mapped in a large scale (Pitt and Ferré-D'Amaré, 2010).

Since then, many RNA sequence spaces have been empirically mapped. Two comprehensive RNA fitness landscapes were mapped by the Chen lab, one for a GTP binding RNA aptamer (Jiménez *et al.*, 2013) and another for a self-aminoacylating ribozyme (Pressman *et al.*, 2019). In the GTP aptamer study, they used in vitro selection to infer the fitness of around 10^{17} RNA sequences, representing the largest fitness landscape mapped at the time. However, this approach only infers fitness based on the survival of the variants through an artificial selection criterion. As a result, the dataset only offers an indirect measurement of fitness and does not assess quantitative activity of the variants. In their next study, using self-aminoacylation ribozyme as the subject, they used kinetic sequencing to measure the catalytic activity of around 70-99% of the entire sequence space of a 21-nucleotide ribozyme. However, this method still relies on in vitro selection and the activity of sequences with very low abundance after selection cannot be estimated. As a result, activity was only estimated for $8.9x10^6$ sequences while the rest of the 4^{21} possible sequences were deemed to be inactive. Both studies reveal, for the first time, the approximation of the entire fitness landscape of RNA and yielded an interesting observation. Both landscapes appear to be very rugged with sparsely distributed fitness peaks separated by valleys of inactivity. This observation implies that even when starting from a known fitness peak, evolution or navigation away from this peak through mutational paths would be very difficult. Most paths would lead to evolutionary dead ends and accessing other fitness peaks would be very challenging.

Other works have similarly shown that RNA fitness landscapes are populated mostly by inactive or deleterious genotypes (Pitt and Ferré-D'Amaré, 2010; Hayden, Bendixsen and Wagner, 2015; Li *et al.*, 2016; Domingo, Diss and Lehner, 2018; Andreasson *et al.*, 2020). A few mutational steps away from a known active genotype (wildtype) often leads to significant reduction in fitness. Development of better methods to identify functional genotypes within a sparse RNA fitness landscape has two important applications. Firstly, it will allow better understanding of the topography of the functional space within the fitness landscape and how evolution proceeds in this space during its search for a new fitness peak. This can elucidate how evolutionary adaptation or innovation can occur during the RNA world (Wagner, 2008) or even during the modern process of viral evolution (Lauring, Frydman and Andino, 2013). Secondly, directed evolution has been employed as a powerful experimental technique that can identify genotypes with new or improved functionality. RNA based technology such as aptamers and riboswitches have been engineered using this method. These synthetic RNA have been used as potential biosensors or genetic control devices for therapeutic and diagnostic application (Famulok, Hartig and Mayer, 2007). A better method to identify functional genotypes can greatly accelerate the discovery and engineering of novel RNA devices (Dykstra, Kaplan and Smolke, 2022) (Figure 2-3).

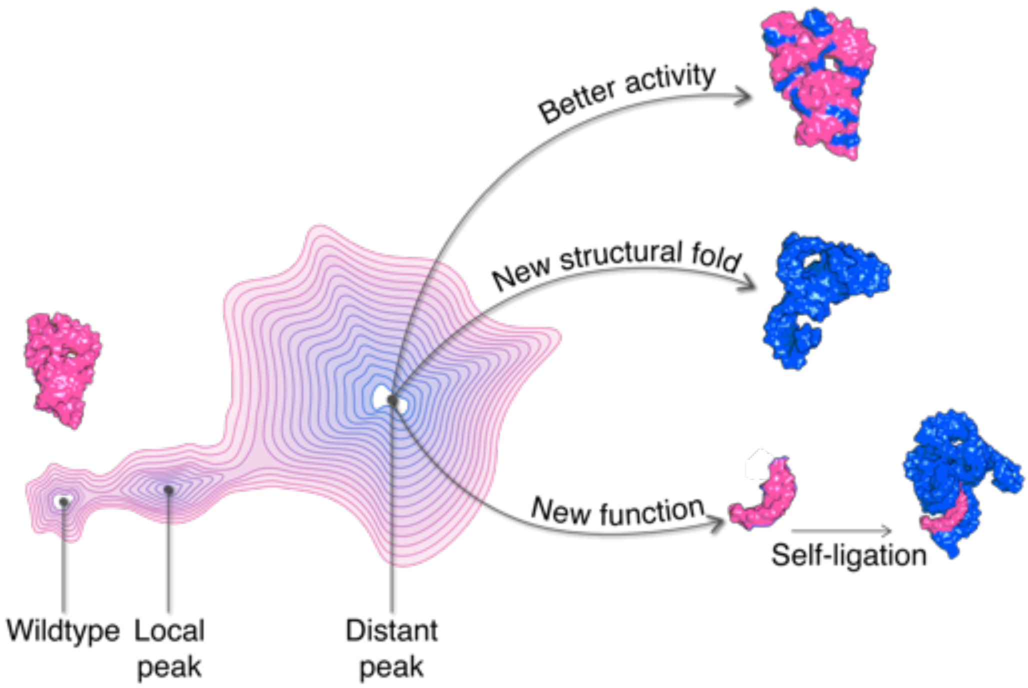


Figure 2-3: Benefits of identifying new peaks in the fitness landscape. Reaching new fitness peaks far away from the starting point (wildtype) can be beneficial for several reasons. Distant peaks can lead to discover of genotypes with improved enzymatic activity, discovery of new structural folds which could be useful for downstream engineering or identifying variants with completely new functions.

Comprehensive screening of fitness landscapes allows researchers to determine the best variants from all possible sequences without the need to optimize across the frustrated and rugged landscape. However, this approach is only possible for a relatively short RNA with a small sequence space. In order to identify ribozymes with improved or novel catalytic function, a larger sequence space of longer RNA must be explored. Furthermore, increasing length has been shown to correlate with an increase in informational and structural complexity (Carothers *et al.*, 2004). Therefore, the exploration of large RNA sequence space might yield new and better insight into the property of fitness landscape. However, the sequencing-based method is currently only experimentally tractable for comprehensive mapping of sequences with length less than around 27 nucleotides (Blanco *et al.*, 2019). Therefore, in order to explore sequence space of larger sizes, we need a method to collect and infer information from sparse sampling of the landscape. In order to maximize the information gain from these sparse sampling, the fitness estimation method must also be quantitative. As mentioned previously, in vitro selection only offers indirect measurement of sequence fitness. Furthermore, such a method leads to loss of low activity variants which provide important information. Therefore, the goal is to develop a method that can measure the activity or fitness of RNA variants in both a quantitative and high-throughput manner (Figure 2-4).

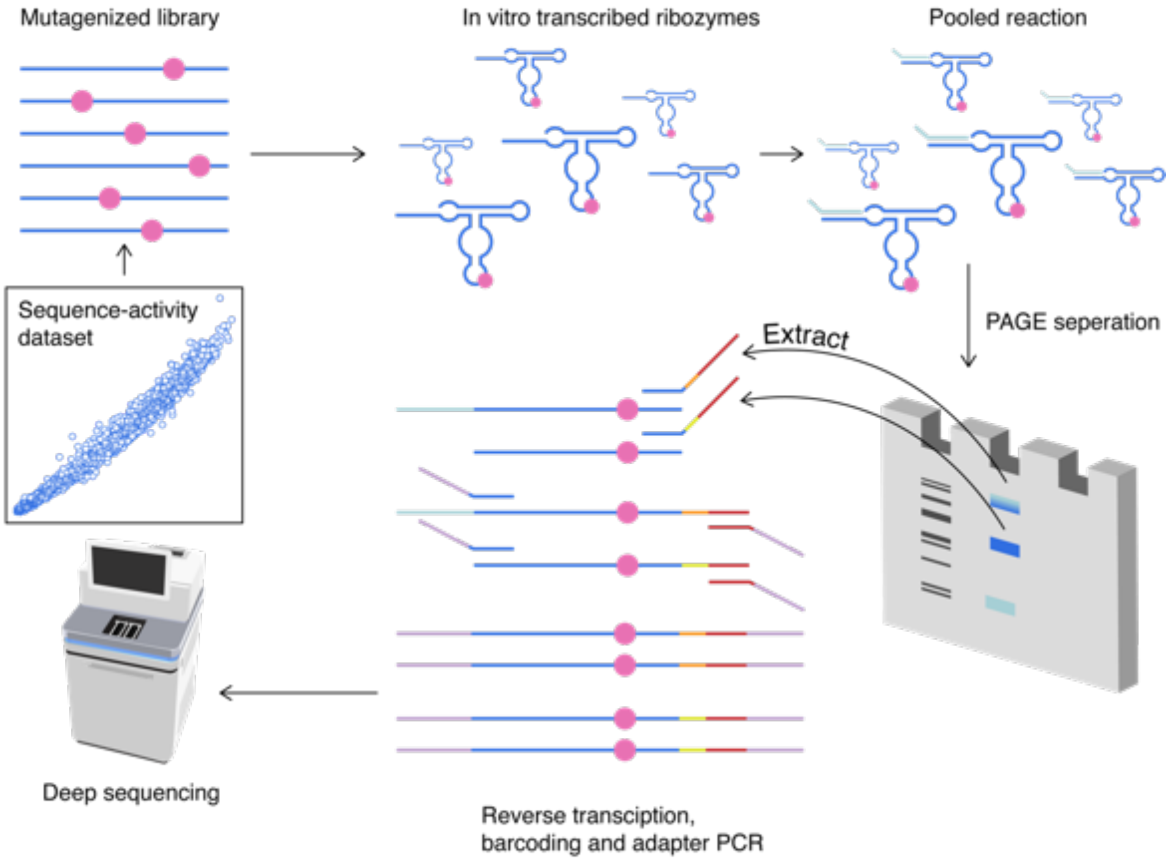

Mutagenized library
In vitro transcribed ribozymes
Pooled reaction
Sequence-activity dataset
Extract
PAGE seperation
Deep sequencing
Reverse transcription, barcoding and adapter PCR

Figure 2-4: Overview of ligase ribozyme high-throughput design cycle.
This figure is adapted from (Nomura and Yokobayashi, 2019). Diverse DNA sequence pool with T7 promoter is first synthesized by using oligo pool synthesis or saturation mutagenesis. In vitro transcription generates a pool of ribozymes that then undergoes ligase reaction in bulk. Ligated and unligated populations of variants can be separated based on size by using polyacrylamide gel electrophoresis (PAGE). The populations are then separately extracted from the gel and barcode can be added using reverse transcription. Different barcodes are used to identify each population during downstream sequence analysis. The cDNA libraries are then pooled together, and PCR is used to add a sequencing adapter to the library. The library is then sequenced with next-generation sequencing platforms such as Illumina MiSeq or NovaSeq. The sequencing reads counts for each sequence in ligated and unligated populations are then counted and used as a surrogate for activity. The resulting sequence-activity dataset is then used to guide the design of the next set of ribozymes.

Our lab previously developed a method to directly measure the activity of large ribozyme variant libraries in parallel (Kobori *et al.*, 2015; Kobori and Yokobayashi, 2016, 2018; Dhamodharan, Kobori and Yokobayashi, 2017; Kobori, Takahashi and Yokobayashi, 2017; Nomura and Yokobayashi, 2019; Yokobayashi, 2019). By comparing the sequencing read counts of reacted and unreacted ribozyme variants that were synthesized using custom on-chip DNA synthesis, we can directly measure the activity of each ribozyme variant (Figure 2-4). Our lab have applied this method to analyze both a natural self-cleaving ribozyme (Kobori and Yokobayashi, 2016) and an artificial RNA ligase ribozyme (Nomura and Yokobayashi, 2019). So far, this method has only been used to explore variants which are only a few mutational steps away from the wildtype. Novel or improved phenotypes can often be located much further away than single or double mutants. Identification of these distant genotypes using information from single or double mutants alone can be difficult due to the effect of higher-order epistasis (Weinreich *et al.*, 2013). It is a well-known phenomenon that accumulation of mutation leads to rapid reduction in activity due to prevalence of negative epistasis (Bank *et al.*, 2016; Bendixsen, Østman and Hayden, 2017). Random sampling of the sequence space will also yield mostly inactive variants due to the sparsity of fitness landscape as discussed above. Therefore, the aim of this study is to develop a method that can leverage the quantitative information gained from local sampling of a fitness landscape to identify functional genotypes from distant regions of the fitness landscape.

To do this, I was inspired by the field of evolutionary computation (EC) which in turn was inspired by natural evolution itself (Miikkulainen and Forrest, 2021). EC goal is to find the best solution to a global optimization problem starting from a set of initial candidate solutions using biologically inspired processes such as mutation, selection and recombination. EC has found successes in many domains of computer science and recently have also been applied to synthetic evolution of biological systems (Yoshida *et al.*, 2018; Boone *et al.*, 2021). Genetic algorithm (GA) is the most popular and most basic form of evolutionary computation (Figure 2-5). It starts with an initial population, which is then selected according to the fitness determined by a fitness function. The selected population undergoes a diversification process involving mutation and recombination to create a new population. This new population is then again subjected to selection and the process is repeated until an optimal solution is found. GA is a powerful optimization algorithm for finding optimal solutions within the vast combinatorial space of possible solutions. Its strength lies in the ability to precisely control several parameters including selection and mutation rate. This enable the algorithm to be adapted to different topologies of fitness landscape where different rate of selection and mutation at different stages of the algorithm can enable the algorithm to avoid local optimums while maintaining efficient search for the global optimum. Therefore, GA could potentially aid in the identification of functional genotypes in a ribozyme fitness landscape as well.

However, applying GA to solve optimization problems for real biological systems is limited by two main problems. First, identifying a computational fitness-function that best represents the real fitness of the system is difficult. There are no obvious properties of a ribozyme sequence that can be accurately calculated computationally that directly correlate with its experimental fitness. I solved this problem by using the high-throughput experimental methods to directly measure fitness. For every round of the algorithm, the new population is experimentally assayed, and the data are fed back into the algorithm, bypassing the need for a computational fitness function. This method offers a much more accurate reflection of the system’s fitness, allowing the algorithm to optimize the function from a real dataset. However, by estimating fitness directly with experiment a second problem occurred. How would I generate each new population with the precise positions of mutations that are required by the algorithm? Select-and-sequence methods often use doped oligonucleotide synthesis or error-prone PCR to create statistically mutated variant libraries. This only allows a global control of ribozyme mutation. However, in my systems, libraries are created using custom on-chip DNA synthesis that allows for the creation of variants with precise mutation. Therefore, I can control precisely where mutations will occur and can also control the location of recombination and crossover. Therefore, custom oligo pool synthesis and the direct quantitative screening method used here should offer a better way to fully exploit the analytical potential of a genetic algorithm than in vitro selection.

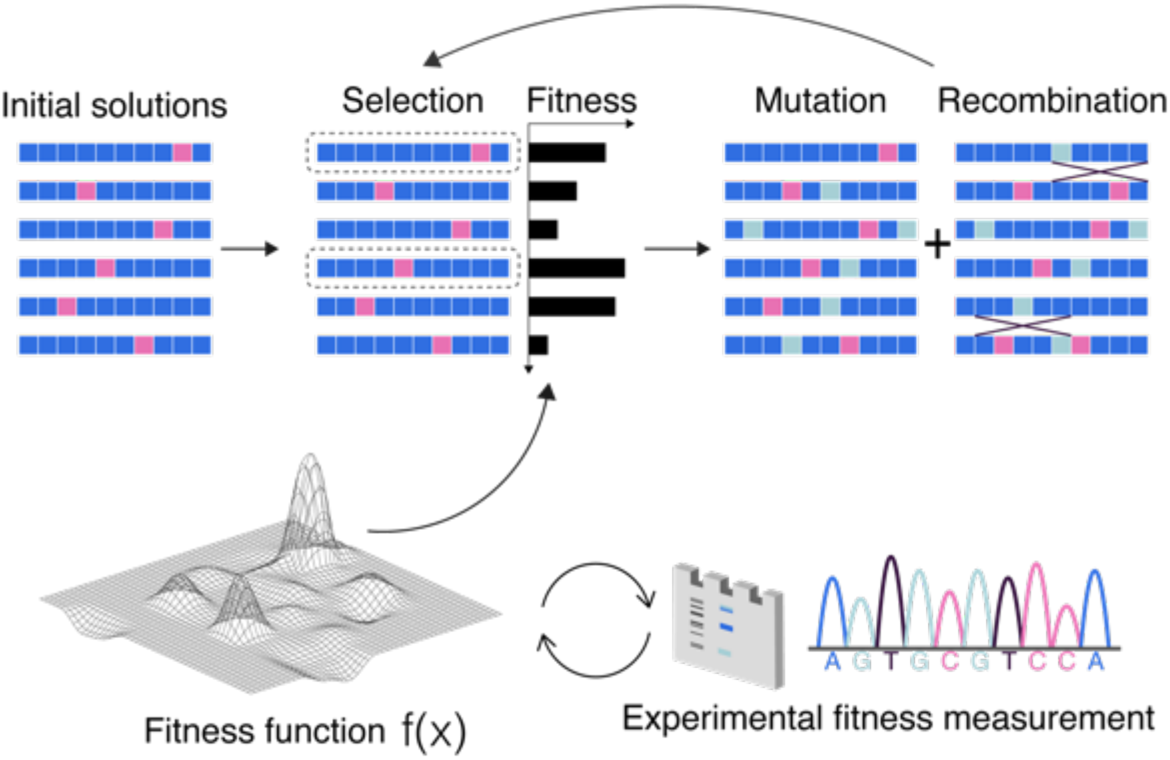


Figure 2-5: Overview of genetic algorithms.
Genetic algorithms (GA) belong to a broader class of evolutionary algorithms (EA), a family of computational processes inspired by natural selection that are used to solve combinatorial optimization problems. GA starts with an initial population of possible solutions that are represented as a collection of genetic components. The set of possible solutions can be evaluated with a fitness function which essentially measures how close each solution is to the target objective. Selection is used to pick favorable solutions which then undergo a diversification process. New combinations of possible solutions can be created by using the process of mutation or recombination, these are collectively known as genetic operators. Mutation changes the value of individual solution components such as flipping a bit or varying real-number value. Recombination or crossover is done by picking random positions in the solution sequence and swapping the solution components with another solution sequence. This process is analogous to sexual reproduction in natural evolution. Designing an appropriate fitness function for a given task can be challenging. Therefore, computational fitness functions can be replaced by real experimental measurement enabling a hybrid approach that performs combinatorial optimization based on real data.

For this study, I chose to use a particular variant of the RNA polymerase ribozyme, the F1 ligase, as a proof-of-concept (Figure 2-6). The F1 ligase ribozyme was first isolated by the Joyce lab (Robertson and Joyce, 2014) and was used to create the first highly efficient self-replicating RNA systems. The F1 ligase ribozyme catalyzes the phosphodiester bond formation through a nucleophilic attack by 3’-hydroxyl group of an RNA substrate on the α-phosphate of the 5'-triphosphate group of the ribozyme. This reaction is guided by the template region within the ribozyme and is analogous to the reaction carried out by the modern RNA polymerase enzymes. The F1 ligase is particularly efficient with a measured k_{cat} of 16.6 ± 0.4 min^{-1}. Thus, this ligase was used to create a self-replicating system that exhibit an exponential growth rate of 0.14 min^{-1} when measured with 10 μM substrate at 48 °C (Robertson and Joyce, 2014). Self-replication is a major catalytic process that is an important determinant of a living system. Therefore, creating self-replicating RNA systems is a major goal in the study of the RNA world. Because of this, many variants and structural motifs have been discovered that can catalyze RNA ligation (Wachowius, Attwater and Holliger, 2017). The F1 ligase was a result of multiple evolutionary lineages and many variants of it exists. Our lab has shown that the F1 ligase ribozyme is surprisingly robust to mutations and deletions (Nomura and Yokobayashi, 2019). The robustness of the F1 ligase ribozyme to mutations could be attributed to the fact that the original ribozyme it was derived from, the R3 ligase, was evolved from a pool of random RNA sequence lacking cytidine (Rogers and Joyce, 2001). The lower available chemical diversity could explain why this structural motif is more tolerant to mutation. The secondary structure of the R3 ligase, which it shares with the F1 ligase, was determined through chemical probing, 3'-terminal deletion analysis and site-directed mutagenesis (Rogers and Joyce, 2001). The analysis showed that the R3 ligase forms a three-way junction structure with the substrate attaching to the ribozyme through Watson-Crick base paring at a complementary region in the ribozyme's 3' terminus (Figure 2-6a). This aligns the 3′ end of the substrate with the 5′ end of the ribozyme at the ligation junction which is rich in purine bases. The study also suggests that substrate binding induces conformational change in the ribozyme causing some residues in the P3 stem to interact through hydrogen bonding. Although the 3D structures of the F1 or R3 ligase have not been solved, they do share a similar secondary structure to an unrelated L1 ligase ribozyme whose 3D crystal structure is available (Robertson and Scott, 2007). The structure of the L1 ligase shows an extensive network of tertiary interactions that stabilizes the transition-state and positions the functional groups for general base catalysis. These structural properties are akin to what are found in natural ribozymes suggesting that in vitro evolved ligase ribozyme could potentially be created through natural evolution. The existence of multiple sequences that share the same ligase activity and secondary structure implied that the sequence space of the F1 ligase ribozyme could be well populated with functional genotypes. Therefore, I chose the F1 ribozyme as the subject of this study to facilitate easier development of this method. Furthermore, dataset that reveals how functional genotypes are distributed within the fitness landscape of a system capable of efficient self-replication could provide insight into the evolutionary search space that led to such behavior in the beginning of life.

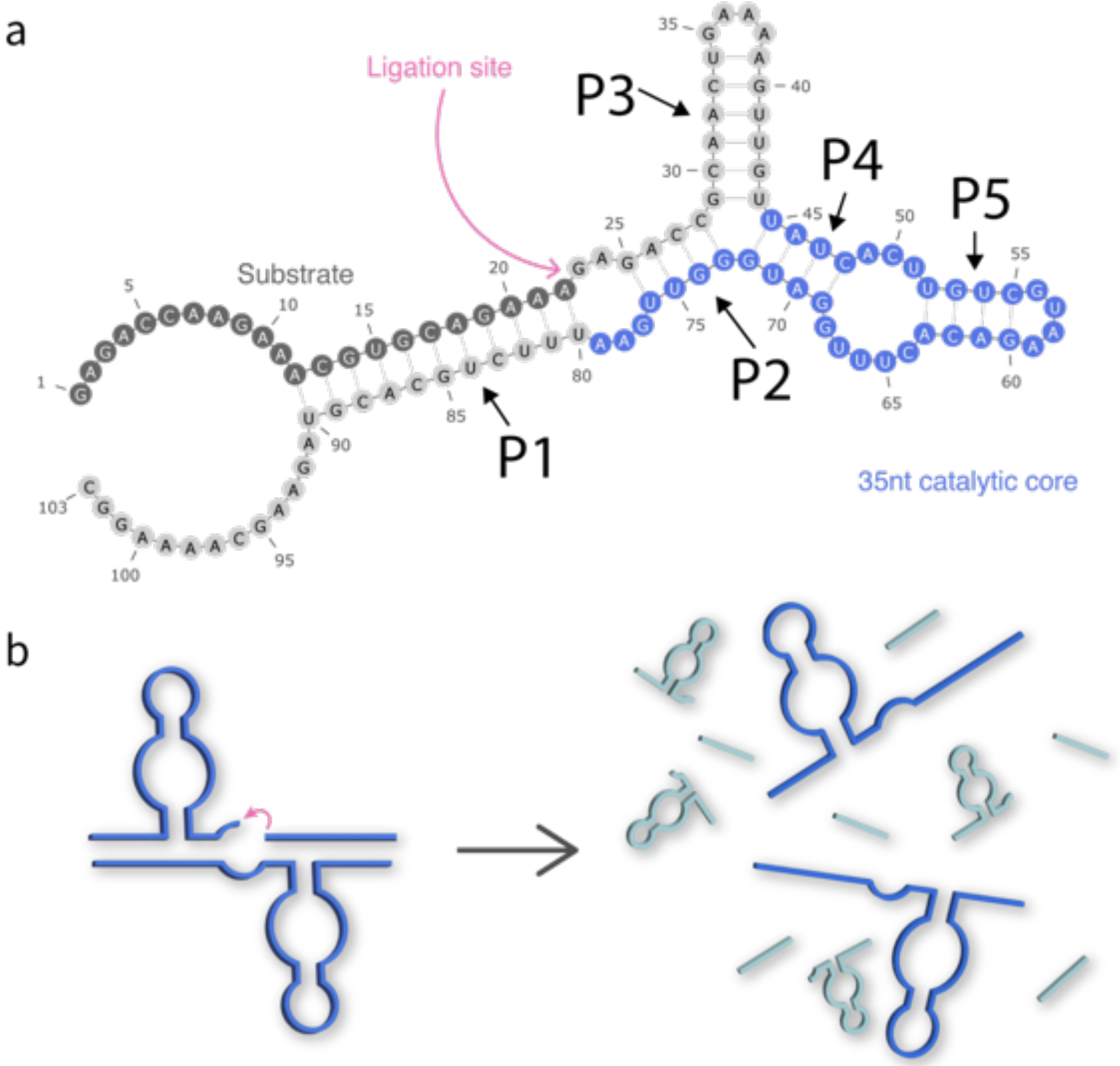
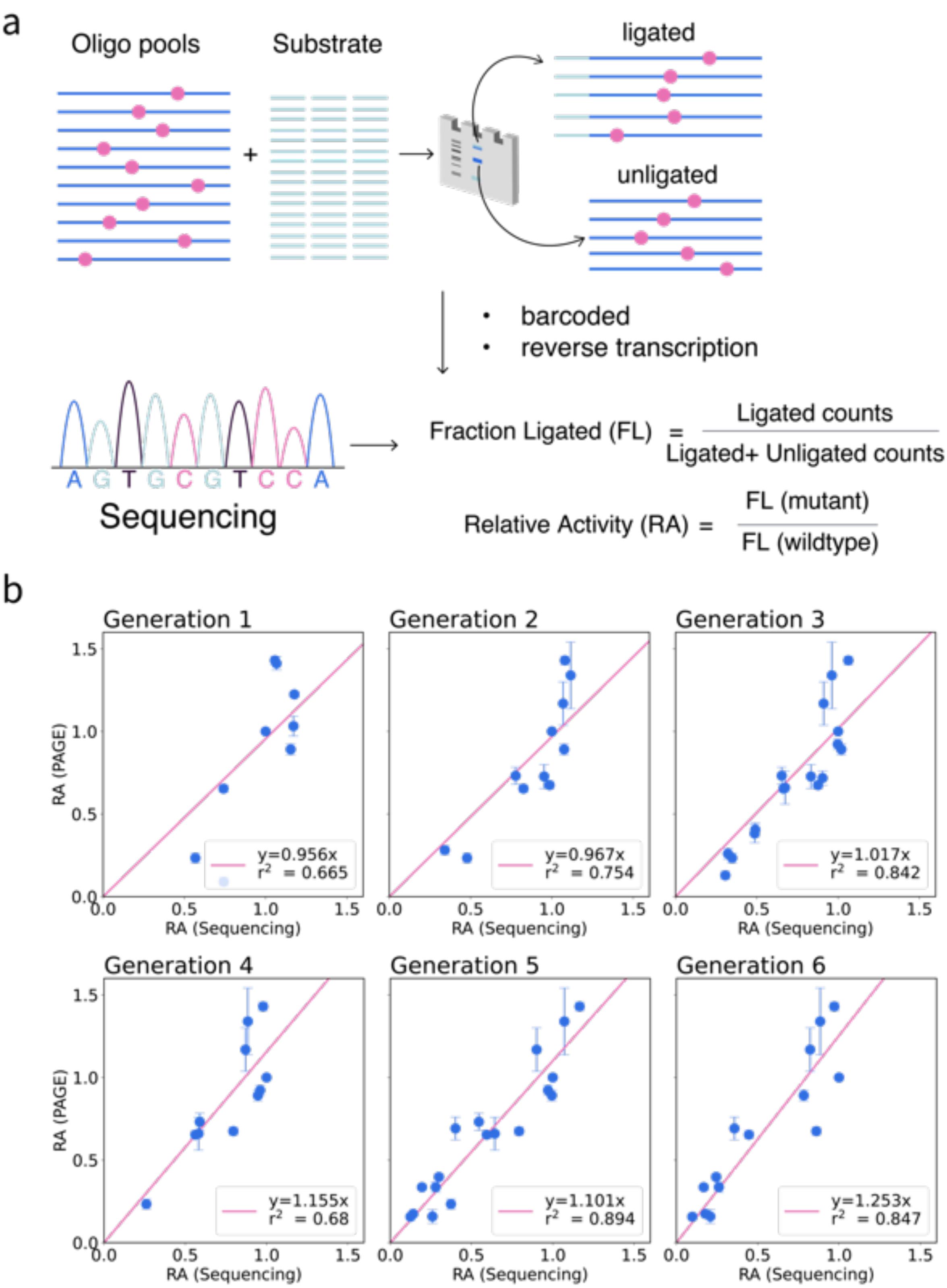


Figure 2-6: F1*U Ligase ribozyme and its secondary structure.
a) Secondary structure of a self-ligated F1*U ribozyme as predicted by ViennaRNA. The F1 ligase ribozyme was first engineered by (Robertson and Joyce, 2014) as a highly efficient self-ligating ribozyme. The ligated F1*U contains U22A and G80U substitutions relative to F1. These mutations were introduced in a previous study (Nomura and Yokobayashi, 2019) to allow the analysis of the regiospecificity at the ligation junction. b) Schematics of the F1 ligase ribozyme in its self-replicating form. Self-replication is a highly studied phenomenon in the RNA world hypothesis.

Results & Discussion

Local landscape analysis of F1*U ligase reveals mutationally robust positions

Using random sampling of the sequence space as initial population for the genetic algorithm would yield mostly inactive sequences due to the sparsity of the fitness landscape. Therefore, I chose to sample the local fitness landscape around the F1*U wildtype (WT) as a starting point. I generated a library of all 105 single, all 5355 double, and 4540 randomly chosen triple mutants of the 35 nt catalytic cores of the ribozyme (Figure 2-6a). The local landscape around the highly active WT should be more likely to contain functional variants. Furthermore, a double mutant map could reveal important information about the epistasis within the landscape. I experimentally measured the ligation activity of all the variants using next-generation sequencing. Fraction ligated (FL) can be calculated by comparing the read counts of each variant in the ligated and unligated population (Figure 2-7a). Then the FL of each variant was divided by the FL of the WT, which was included in every library as a control, to calculate the relative activity (RA). Two repeats of the sequencing assay were done for each experiment and the mean RA values were calculated for each variant (Figure S1-1). The mean of RA is then used as a proxy for variant fitness and is referred to only as RA for subsequent discussion in the rest of this thesis. After every round of sequencing assay, some mutants in each library were also randomly selected and were individually synthesized. The RA of these mutants was also measured using PAGE assay to confirm agreement with sequencing-based assay (Figure 2-7b).


a
Oligo pools
Substrate
+
ligated
unligated
barcoded
reverse transcription
A G T G C G T C C A
Sequencing
Fraction Ligated (FL) = Ligated counts / Ligated+ Unligated counts
Relative Activity (RA) = FL (mutant) / FL (wildtype)
b
Generation 1
y=0.956x
r² = 0.665
Generation 2
y=0.967x
r² = 0.754
Generation 3
y=1.017x
r² = 0.842
Generation 4
y=1.155x
r² = 0.68
Generation 5
y=1.101x
r² = 0.894
Generation 6
y=1.253x
r² = 0.847
RA (PAGE)
RA (Sequencing)
0.0
0.5
1.0
1.5

(Figure caption continues from the previous page)

Figure 2-7: Overview of ribozyme activity calculation and PAGE confirmation experiments.
a) Ligated and unligated sequences are separately extracted and reverse transcription is used to retrieve cDNA and to add a barcode. The frequency of each designed sequence is counted after deep sequencing. Ligated population, identified by the barcode, is divided by the total frequency (ligated + unligated) of that sequence to give fraction ligated (FL). The activity or fitness of each sequence is given as the FL divided by the FL of the wildtype sequence giving the relative activity (RA). b) For every population, a subset of sequences is selected, and PAGE is used to determine the RA of individual sequences (see Methods). The PAGE measured RA is compared with the RA measured using the sequencing method. The data points are presented as mean values +/− SD with n = 3 for the PAGE values and n = 2 for sequencing values. Square of Pearson's correlation coefficient (r^2) measures correlation between RA values determined by the two methods and indicates good agreement between the two assays.

I first examined the RA of each single mutant to evaluate the mutational robustness of the catalytic core (Figure 2-8 diagonal). Mutations within the P4 and P2 are not tolerated and result in completely inactive ribozymes, while mutations within the P5 stems are well tolerated. Surprisingly, mutations between position 75 and 79 are also slightly tolerated despite being close to the ligation junction. These results are also supported by looking at the double mutant map (Figure 2-8 lower triangle). The most well tolerated double mutants are compensatory pairs that restore base pairing within the P5 stem of the ribozyme. Double mutants within the loop at the end of P5 are also well tolerated. However, double mutants with mutations in P4 and P2 stems are not tolerated at all. Even compensatory pairs within P4 result in mostly inactive ribozymes. Looking at the pairwise epistasis (E_{AB}) calculated using the log-additive model ($E_{AB} = \log(RA_{AB}) - (\log(RA_A) + \log(RA_B)))$ from the experimental RA showed that strong positive epistasis occurs between positions that form the base pairing within the P5 stem (Figure 2-8 upper triangle). This suggests that the F1*U fitness is contingent on maintaining the P5 stem loop.

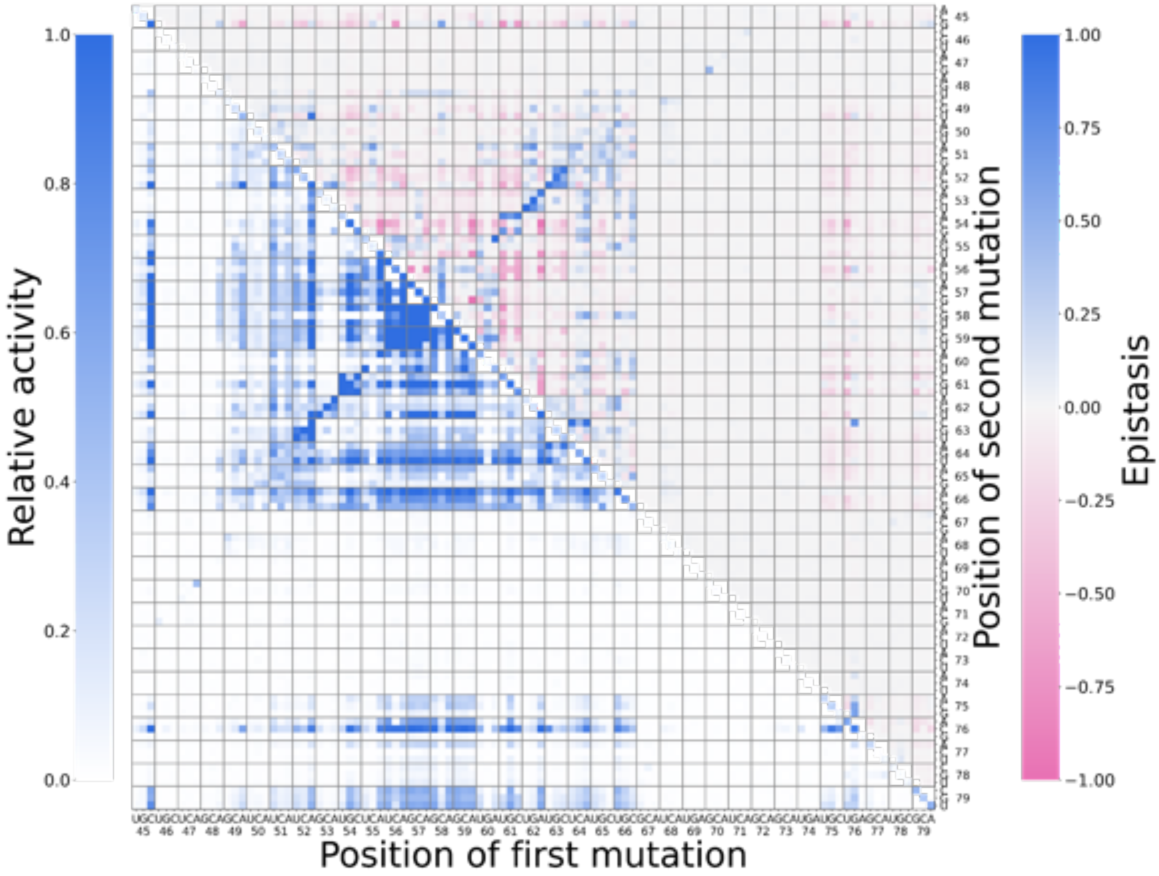


Figure 2-8: The fitness and epistasis measurements of double and single mutants of F1*U.

In the first generation, the relative activity (RA) of the complete set of single and double mutants of the F1*U ligase was measured. The RA of each variant is plotted in the lower triangle of the heatmap, with the diagonal showing the RA of single mutants. The upper triangle showed the pairwise epistasis (E_{AB}) values calculated from the RA measurements using the log-additive model ($E_{AB} = \log(RA_{AB}) - (\log(RA_A) + \log(RA_B)))$.

So far, I have only looked at the minimum free energy (MFE) of the ribozyme. However, the ensemble of possible alternative structures can reveal how stable the MFE structure is and could potentially explain the observed mutational effects. I calculated the base-pair probabilities of all possible base pairs for the F1*U ligase using ViennaRNA web server (Figure 2-9). Overlaying the probabilities onto the MFE structure showed that the P4 stem has lower probabilities compared to the rest of the structure (Figure 2-9a). This explains why the P4 stem has low tolerance to mutation, even for compensatory double substitutions, as any substitution could easily disrupt the base-pairing leading to overall loss in activity. The rest of the ribozyme has relatively stable secondary structure. The dot plot shows little alternative structures around the MFE structure suggesting that the F1*U ligase occupies a single peak within the structural landscape (Figure 2-9b). This suggests that identifying functional variants within the ribozyme sequence space could be done by preserving the WT secondary structure. The agreements between experimental mutational effects and predicted secondary structure suggested that even a very local sampling of the fitness landscape can provide important information about the structural constraints of the F1*U ligase ribozyme which can hopefully be exploited for subsequent design processes.

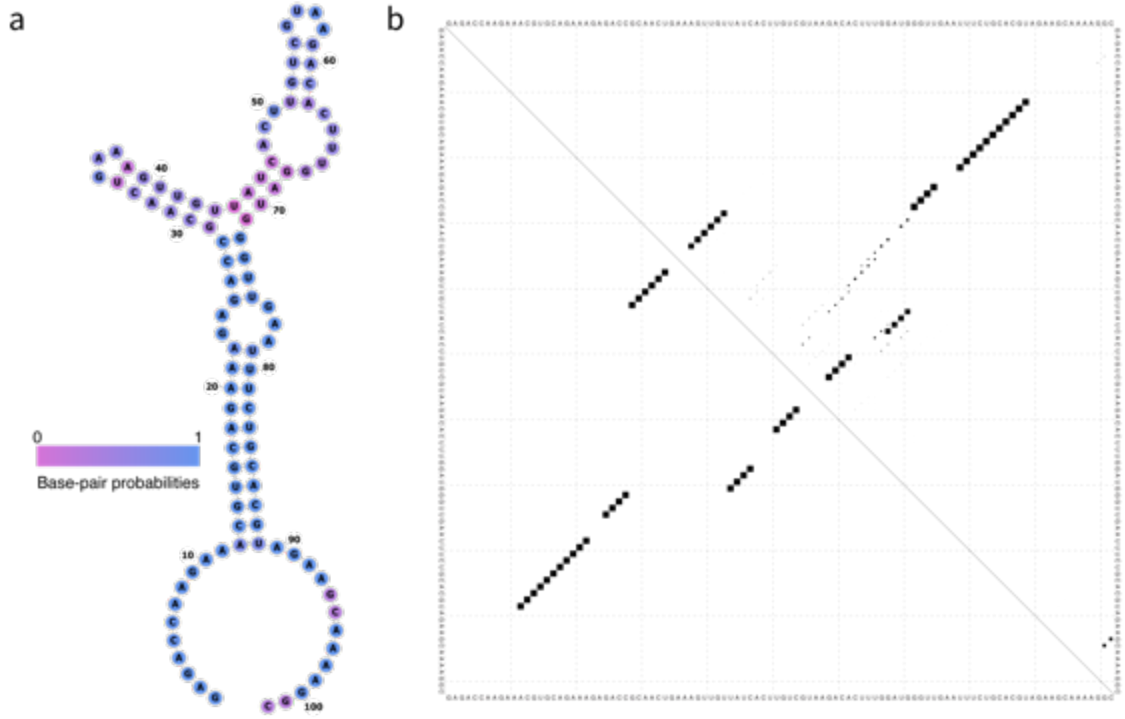


Figure 2-9: Base-pair probabilities of the F1*U ligase ribozyme.
a) Base-pair probabilities were calculated using RNAfold WebServer and is overlayed as a colormap on the minimum free energy structure. B) Dot plot of the F1*U ligase ribozyme secondary structure with the lower triangle showing the minimum free energy structure and the upper triangle showing the probability of all possible base pairs with the area of each dot proportional to the pairing probabilities.

Epistatic information can be gained from inspection of the local landscape

As mentioned earlier, the local landscape around the WT, although representing a minuscule fraction of the total sequence space, can contain rich information about the level of epistasis within the landscape. Epistasis is observed when the effects of mutation combine in a non-linear manner. High level of epistasis increases the ruggedness of the landscape, constraining evolution and reducing the predictability of the landscape by models that only learn from small samples. Therefore, I evaluated epistasis level within the local landscape of the WT to assess how well the genetic algorithm will be able to traverse it. A simple measure of epistasis is to see how well effects of multiple mutations can be predicted from single mutants. I used the log-additive model of epistasis, where the expected ln(RA) of a mutant is the sum of ln(RA) of all its constituent single mutants. Coefficient of correlation (R^2) is then calculated between the expected and observed RA, where a value of 1 indicates complete absence of epistasis.

Looking at Figure 2-10a, the R^2 of the double mutants indicates that most double mutants can be well predicted by single mutants ($R^2 = 0.759$). The R^2 between the triple mutants are also relatively high ($R^2 = 0.618$, Figure 2-10b), however this is probably due to mostly inactive variants present in the population. Strong negative epistasis is a common feature within the RNA fitness landscape (Bendixsen, Østman and Hayden, 2017) where increasing the number of mutations leads to rapid reduction in fitness. However, the low level of pairwise epistasis suggests that the landscape around F1*U could be smoother than other observed landscapes. This could mean that genetic algorithms might be able to find evolutionary paths away from the WT peak.

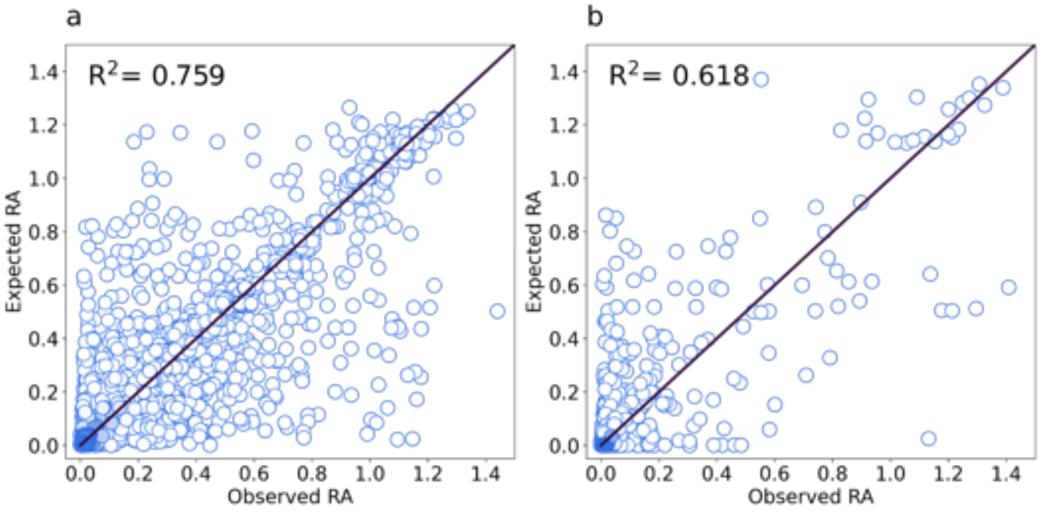


Figure 2-10: Comparison between experimentally measured relative activity (RA) and RA expected under the epistasis-free model.

In an epistasis-free landscape, RA values of any variant can be calculated by summing the RA of all constituent single multinational effects in the log space. The RA values calculated from this epistasis-free approach (Expected RA) are compared to the experimentally measured RA values (Observed RA) of all a) double and b) triple mutants present in generation 1.

Genetic algorithm can increase proportion of neutral mutants

Next, I investigated whether genetic algorithms (GA) can be used to design ribozyme variants and effectively identify neutral genotypes within the landscape. I defined neutrality as having an RA of at least 0.2. This is based on the following reasoning. First, most of the genotypes within the sequence space is assumed to be completely inactive. Rarity of functional genotypes have been observed multiple times in previous experimental studies (Bendixsen, Østman and Hayden, 2017). Similarly, I observed that most of the variants within this study have RA below 0.2 (Figure 2-11a). Therefore, activity at the level of 20% of the WT is considerably active given the rarity of these variants and evaluating the performance of the genetic algorithms at this threshold is considered sufficient. Secondly, evaluation of the sequencing data reveals that most of the variants has standard deviation below 0.2 when calculated from two replicates (Figure 2-11b). Furthermore, it was assumed that deviation between RA calculated from two replicates would correlate with the number of reads from the sequencing assay. This would mean that less active variants would have higher deviation due to lower frequency of reads in the ligated population. However, this is not the case, with standard deviation only weakly correlated to relative activity or the mean read counts (Figure 2-11c & d). Considering these analyses, any variants with RA above 0.2 could be considered to have a reliably detected activity.

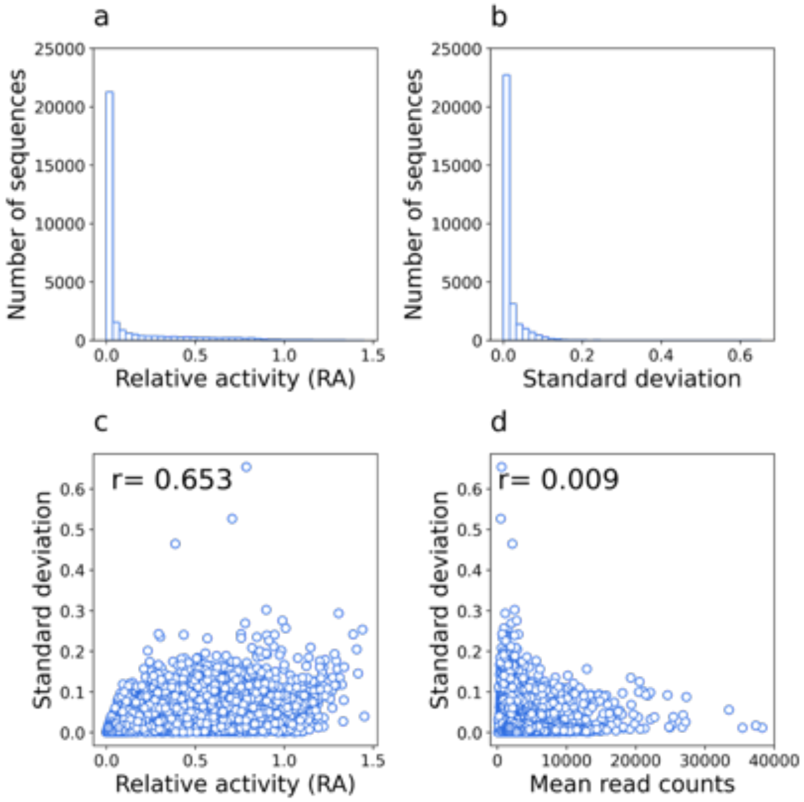


Figure 2-11: Analysis of sequencing measurement errors for generation 1-6. Histogram showing the distribution of a) the mean relative activity (RA) and b) standard deviation (SD) of all 29,887 sequences in generation 1 to 6 after filtering for read counts and sequence quality. The SD was also plotted against the mean RA and mean read counts of each variant. The mean RA, standard deviation, and mean read counts are calculated from two independent sequencing assays. The total read count of each variant is calculated as the sum of the ligated and unligated read counts.

The choice of threshold for neutrality could be considered to be too arbitrary. However, the notion of neutrality and fitness is highly context dependent. Fitness of a variant is relative to the selective pressure imposed on it. Under the scenario where selection is done at the molecular level, then any variants with lower activity would be rapidly removed from the population. However, when the catalytic step is limited by another step, for example when the population exists in protocells which require replication of the compartment. Then any activity above a certain threshold would be considered selectively neutral. Correlation of activity to fitness in different evolutionary scenarios is an important topic of future studies. However, for the current study, where the goal is to assess how well a genetic algorithm can identify rare functional variants within vast sequence space, a minimal reliable threshold of 0.2 is considered sufficient.

I used a genetic algorithm to design 5 more populations of ribozyme variants named generation 2 to 6 with each generation subjected to experimental assay. Briefly, I used RA as a proxy for variant fitness and starting from generation 1, I used computational selection, recombination, and mutation to successively design each generation which were then experimentally screened to determine fitness for the next round. (See Methods for details.) Selection is done using a method called tournament selection which maintains a small proportion of variants with low to medium activity. In generation 2 to 5, new variants are created from a mixture of recombination and point mutation (substitution) of selected parent variants (Figure S1-2). However, as visible on Figure 2-12a, this method leads to a low fraction of neutral mutants in the population. Therefore, I changed the strategy in generation 6, allowing a higher number of mutants that were created by recombining selected parents to be passed on to the final population without undergoing random substitution (Figure S1-3). Some recombinants and selected mutations from the previous generation still undergo point mutation to maintain diversity. This change in strategy leads to significant increase in the proportion of neutral mutants in generation 6 (Figure 2-12a). The increase in the proportion of neutral mutants as a result of the increased contribution of recombination supports earlier observations from directed evolution experiments which shows recombination can better preserve structure and function of proteins than random substitution (Drummond *et al.*, 2005).

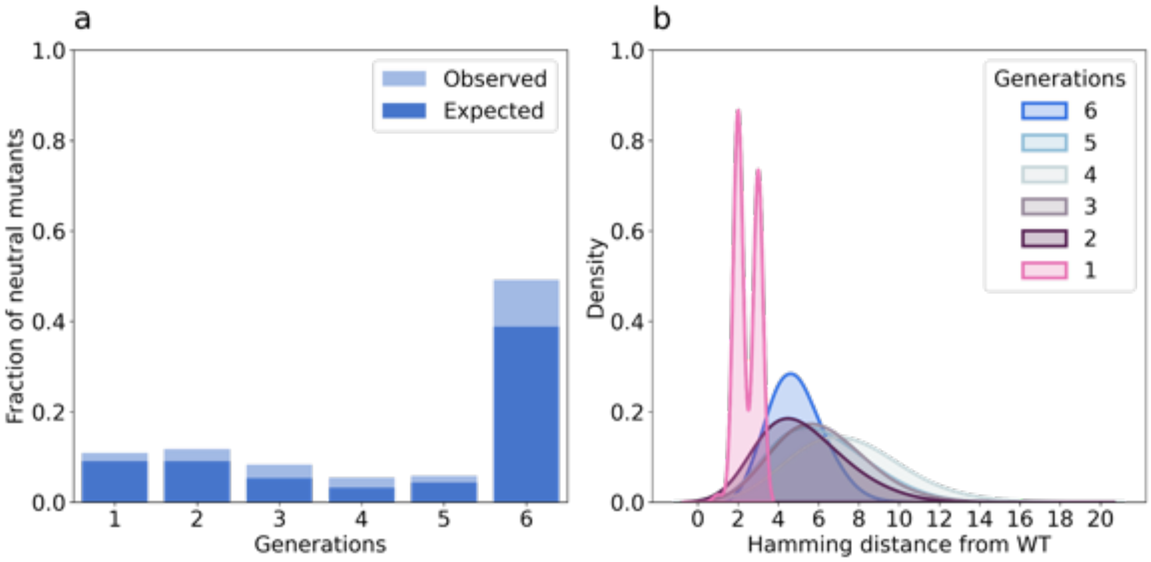


Figure 2-12: Fraction of neutral mutants and distribution of mutant diversity in generation 1 to 6.
a) Mutants are considered neutral if their relative activity (RA) is higher than or equal to 0.2. Fractions of neutral mutants are calculated from all mutants in each generation from 1 to 6. The dark blue area indicates the neutral mutants that are also neutral when RA values are calculated from the epistasis-free log additive model. b) The Hamming distance of mutants in each generation from the wildtype (WT) are plotted as probability distributions.

I also calculated the expected activity of each variant in all the generations using the log-additive model. Mutants that have the activity calculated by log-additive model above 0.2 are classified as expected neutral mutants. As shown in Figure 2-12a, most neutral mutants identified in each generation are also expected to be neutral under the log-additive model. This means that the genetic algorithm is more efficient at identifying mutants that exhibit lower levels of epistasis. These properties could be intuitively explained considering that recombination of selected neutral mutants is more likely to result in another neutral mutant if the combination of each mutant exhibits a linear or non-epistatic effect. Neutral mutants that are a result of mutational effects that combine non-linearly would be harder to identify using only iterated rounds of selection and recombination. I also measured the diversity of each generation by looking at the Hamming distance of each sequence from the WT (Figure 2-12b). A change in strategy during generation 6 results in slightly lower diversity compared to previous generations. This is expected, as recombination generates less overall diversity compared to random substitution. However, looking at the distribution of RA as a function of Hamming distance to the WT shows that generation 6 was able to identify variants with higher Hamming distance that retain considerably higher activity than mutants found in the same distance in previous generations (Figure 2-13). The strategy used in generation 6 was able to identify variants with 7 to 10 mutations that have much higher activity than all other mutants found with the same range of mutations.

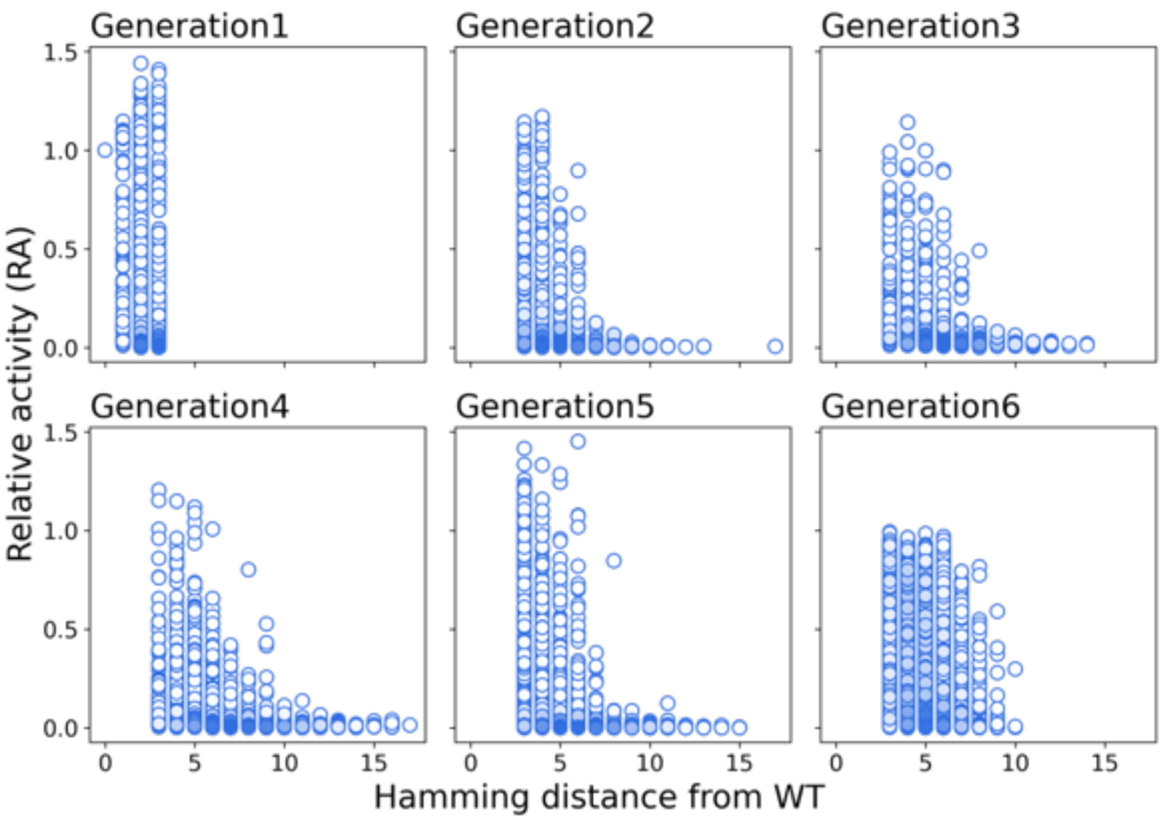


Figure 2-13: The fitness landscape explored by the genetic operators. The relative activity (RA) of each variant is plotted against its Hamming distance from the wildtype (WT). Each generation explored mutants further and further away from the wildtype while increasing the proportion of variants with high activity.

A strategy to increase the efficiency of the genetic algorithm could be a recombination method that maintains the secondary structure of the ribozyme. It is widely accepted that secondary structure of an RNA is important for its activity and function. I calculated the frequency of the MFE structure in the thermodynamic ensemble for the F1*U ligase using ViennaRNA package. The calculated frequency is 26.04% which indicates that there are few alternative structures for the F1*U sequence and that maintaining the MFE structure could be important for retaining catalytic activity. My current approach represents the ribozyme as linear sequence of bases, and no structural or base pairing information is given. In this representation, crossover occurs at a random position within the linear sequence and can occur in a way that breaks a local structural motif such as in the middle of a stem. This could be avoided if the ribozyme is represented in a way that provides information about substructures within the sequence. RNA secondary structure can be represented in a tree-like graph where each node represents an unpaired region such as internal loop and edges represent the base pair stem connecting these loops (Shapiro, 1988). Using this representation, the crossover function can be specified to only occur between corresponding nodes preventing disruption of stem regions. This will enable efficient recombination of structural motif while preserving the overall secondary structure of the ribozyme. In fact, a similar approach has been used by earlier works to discover common structural motifs in genomic RNA sequences (Hu, 2002, 2003; Michal *et al.*, 2007). The GeRNAMo system utilizes RNAsubopt function within the ViennaRNA package to predict all possible structures of subsequences of a set of genomic RNA sequences (Michal *et al.*, 2007). The predicted structure is represented as a tree and an evolutionary algorithm was then used to find a set of common motifs amongst these genomic sequences. Although this approach has been successfully used for motif discovery, it has not been used for RNA sequence design. Future work could improve upon the approach I presented here by evaluating how different sequence representation strategies affect the efficiency of the genetic algorithm.

Conclusion

In this chapter, I presented a new approach to directed evolution that utilizes the precision of custom oligo pool synthesis combined with smart library design using genetic operators. Genetic operators are borrowed from the field of evolutionary computation (Miikkulainen and Forrest, 2021). Here, I showed that a combination of these processes with real experimental evaluation can be used to improve the efficiency of directed evolution processes. Directed evolution are powerful experimental tools that have been widely used to engineer new or improved proteins for biotechnology application. However, the ruggedness of the fitness landscape limits the success of this technique. Fitness valleys and isolated fitness peaks often constrain directed evolution towards local maxima. Optimization away from these maxima is difficult as these regions of low activity can be quite pervasive. Because most variants are inactive this means that most of the datapoints collected from high throughput assay provide little information about the functional regions of the sequence space and do not provide any starting point for further rounds of optimization. The works I have presented here provided a novel method that can increase the availability of these functional variants which can be used for further optimization. The pipeline I have developed can be used to efficiently collect functional variants as initial population for further directed evolution. Increasing selection stringency could lead to higher chance of getting more active ribozymes. This approach could potentially reduce the cost and experimental burden of directed evolution. Furthermore, the sequence-activity data collected in this chapter provide information about the topology of the F1*U ligase ribozyme fitness landscape. The statistical patterns of mutation within the neutral and deleterious population could be used to inform directed evolution to further increase its speed and efficiency. In the next chapter, I will discuss how I use machine learning to extract information from this dataset and create a predictive model for informed ribozyme library design. I investigate how this model can be used to augment experimental directed evolution for the exploration of distant functional regions within the fitness landscape.

Methods

Preparation of ligase ribozyme libraries

I used commercially available custom oligo pools from Twist Biosciences to construct dsDNA templates of ribozyme libraries for in vitro transcription. I ordered the oligo pools with the T7 promoter and ribozyme sequence and then amplified them through PCR using Phusion High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs (NEB)) and primers Ligase-lib-f and Ligase-lib-r (Table S2-1). Afterward, I purified the PCR product using the DNA Clean & Concentrator-5 kit (Zymo Research).

For in vitro transcription, I performed the reaction using a purified dsDNA template with a ScriptMAX Thermo T7 Transcription Kit (Toyobo) in a 10 µL volume. Once the transcription reaction was complete, I incubated the solution for 10 min at 37 °C with a DNase-I (NEB) solution consisting of 2 µL DNase I (2 U/µL), 2 µL 10× DNase I Reaction Buffer, and 6 µL nuclease-free water. Finally, I purified the RNA product using the RNA Clean & Concentrator-5 kit (Zymo Research).

Ligation reactions of ribozyme libraries

I mixed the ribozyme pool (0.8 µM) with substrate F1*subA (Table S2-1) at 8 µM in nuclease-free water, creating a reaction volume of 24 µL. After that, I heated the solution to 72 °C for 3 min, and then cooled it down to 4 °C for 5 min. To continue, I separately incubated the RNA solution and the 4× reaction buffer (200 mM EPPS pH 7.5, 2.0 mM $MgCl_2$, 8 U/µL RNase Inhibitor, Murine (NEB)) at 37 °C for 3 min. Once ready, I initiated the reaction by adding the RNA solution to 8 µL of reaction buffer, followed by an incubation of 60 min at 37 °C. Finally, I terminated the reaction by adding 72 µL of a cold stop solution (25 µL 0.5 M EDTA and 65 µL RNA Loading Dye (2×) (NEB)) and kept it on ice.

Preparation of sequencing templates

I heated the reaction solutions to 95 °C for 3 min and separated them on a 12% urea polyacrylamide gel. Then, I stained the gels with SYBR Gold (Thermo Fisher) and visualized them using a blue light transilluminator. I excised and crushed the ligated and unligated ribozyme bands and extract the RNA in Tris/NaCl buffer (30 mM Tris-HCl, pH 7.5, 30 mM NaCl) by shaking at 1200 rpm and 4 °C for 18 h. Next, I precipitated RNAs by ethanol using Quick-Precip Plus Solution (EdgeBio), washed them twice with 70% ethanol, and resuspended them in nuclease-free water. Then, I dissolved the ligated and unligated RNA in 10 µL of nuclease-free water and used 5 µL for reverse-transcription reactions. I performed reverse-transcription reactions in a 10 µL volume with Maxima H Minus Reverse Transcriptase (Thermo Fisher) according to the manufacturer's instructions. I used R1-[barcode]-F1-lig (Tables S2-1) as the reverse-transcription primer, with different barcodes for unligated and ligated ribozymes. The reactions were allowed to proceed for 30 min at 65 °C, and the enzyme was inactivated at 85 °C for 5 min. I added 1 µL of 20 U/µL exonuclease I (NEB) to the reverse-transcription solution to remove the primers and incubated it for 30 min at 37 °C followed by 15 min at 85 °C. I combined and diluted the solutions (ligated and unligated) for PCR analysis. Then, I used Primers R2-F1-lig and R1-f2 (Tables S2-1) to amplify the cDNA mixture using the Phusion High-Fidelity PCR Master Mix with HF Buffer. I diluted the PCR product and used it in a second PCR with TruSeq-i7-UDI000# and TruSeq-i5-UDI000# primers (Tables S2-1). Different UDIs were used to identify different replicates if they were sequenced simultaneously. Finally, I purified the PCR products by agarose gel electrophoresis using the Zymoclean Gel DNA Recovery Kit (Zymo Research). I measured DNA concentration by real-time PCR (StepOnePlus, Thermo Fisher) using the NEBNext Library Quant Kit for Illumina (NEB). The DNA library was analyzed using Illumina NovaSeq or MiSeq by the Sequencing Section at OIST.

Sequencing data analysis

I used custom Python scripts to analyze the sequencing data. The script sorted each read in the FASTQ file into ligated or unligated pools based on the barcode sequence. Quality filtering is done by checking that all base calls within the variable catalytic core region have $\mathrm{QS} \geq 20$. For the F1*$\mathrm{U^m}$ and WT/Mut libraries, a maximum of one base call in the variable region was allowed to have $\mathrm{QS} < 20$. For each variant, the reads were counted for the ligated (N_{ligated}) and unligated sequence ($N_{\mathrm{unligated}}$), which were then used to calculate the FL ($\mathrm{FL} = N_{\mathrm{ligated}}/(N_{\mathrm{ligated}} + N_{\mathrm{unligated}})$). Relative activity (RA) was calculated for each variant by dividing the FL by that of the WT, which was included in every generation as control. Each generation was assayed in duplicate, and variants were discarded if the total read count ($N_{\mathrm{ligated}} + N_{\mathrm{unligated}}$) in either replicate was below 30 for the F1*$\mathrm{U^m}$ library or below 100 for all other libraries. The mean RA was calculated from the two measurements for each variant and is referred to as the RA for subsequent analysis.

PAGE analysis of individual ligase ribozymes

I constructed DNA templates for individual ribozyme variants by annealing and extending two oligonucleotides using OneTaq 2X Master Mix with Standard Buffer (NEB). (Tables S2-2) I then purified the PCR products using DNA Clean & Concentrator-5 columns and the PCR products were transcribed in vitro as described above. I performed the ligation reactions using the same procedures as described above, except for using excess ligase ribozyme (2 μM) over the FAM-labeled substrate (FAM-F1*subA, 0.1 μM, FASMAC). I imaged the polyacrylamide gels using a Typhoon FLA9500 (GE Healthcare) and the band intensities were quantified using ImageJ 2.3.0 software.

Computational selection, mutation, and recombination

The evolutionary pipeline consisted of iterative cycles of oligo pool synthesis, experimental assay, computational selection, computational recombination, and computational mutation. Flowcharts describing the steps in the algorithm are shown in Figure S2-2 and S2-3. I used tournament selection as the selection method. In this method, a predetermined number of variants from the population are randomly selected, and the variant with the highest RA within this set is retained as a parent. The remaining variants (losers) are returned to the population, and the process is repeated until a predetermined number of variants are selected as parents. Using tournament selection allows a small percentage of medium to low RA variants to be selected along with high RA variants for the next generation. This approach could account for the epistatic nature of the fitness landscape, where less-active mutants might become more active later with additional mutations.

In the first design of the algorithm (Figure S2-2), two parental sequences were picked at random and recombined using one-point crossover at a random position, and one of the resulting recombinants was randomly selected for substitution. Each position in the sequence has a 1/35 chance of mutating to one of the three remaining bases with an equal probability. Therefore, on average, each recombined mutant had one substitution. This process was repeated until the total number of offspring was reached. Mutants were selected only if they had not been previously selected. Finally, some mutants were randomly replaced with the controls. This strategy was used to design generations 2–5.

For generation 6, parents were picked by tournament selection. A set of pure recombinants was then generated from the parents. Next, a random variant was selected from the pool of recombinants and parents. This variant was then randomly mutated, and the process was repeated to create another set of point mutants that were generated by random substitution of parents or recombinants. The new generation consisted of a combination of pure recombinants and random point mutants (Figure S2-3). I prespecified the number of pure recombinants and random point mutants in the population.

The parameters used in the computational algorithm, including tournament size, number of parents, number of pure recombinants, number of random mutants, and total population size, are listed in Table S2-2. From generation 3 onward, I increased the total population size to increase the chance of finding neutral mutants during each round of experimental screening. From generation 7 onward, I reduced the tournament size and increased the number of selected parents to account for the increased fraction of neutral mutants. This led to an overall reduction in selection stringency to ensure that some variants with lower activity were still being selected

Chapter 3: Computational evolution of diverse ribozyme sequences with supervised machine learning models

Parts of this chapter, in particular the Methods section, have been duplicated or updated from a previously published article: Rotrattanadumrong, R. and Yokobayashi, Y. (2022) ‘Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning’, *Nature communications*, 13(1), p. 4847.

Background

Advances in DNA sequencing and synthesis have had a wide-ranging impact across different fields of experimental biology. It has led to developments of ultra-high throughput experimental methods that lead to an unprecedented explosion of biological datasets ranging from splicing data (Jaganathan *et al.*, 2019) to protein binding specificity (Alipanahi *et al.*, 2015). Studies of fitness landscapes is amongst the fields that has benefitted from these technical advancement the most (Blanco *et al.*, 2019). Many works have recently been published that shed light on the high dimensionality and ruggedness of RNA fitness landscapes (Jiménez *et al.*, 2013; Li *et al.*, 2016; Puchta *et al.*, 2016; Domingo, Diss and Lehner, 2018; Li and Zhang, 2018; Pressman *et al.*, 2019). My result from the previous chapter shows that navigation of such high-dimensional and rugged space can be made more efficient by leveraging a computational genetic process of selection, mutation and recombination. However, these processes maximize the chance of finding functional variants by combining mutations already known to be functional in a semi-random manner. The genetic operators themselves do not learn any rules, properties or topology of the underlying fitness landscape and therefore have very little predictive power when it comes to distant unseen regions of the sequence space. Because the combinatorial space is extremely large, it would be impossible to comprehensively investigate this space using just the genetic operators and high-throughput experiments. Complex epistatic relationships within the ligase ribozyme sequence space means that the dataset collected thus far contain rich information about the properties of fitness landscapes. A predictive model that can capture the complex properties of such high-dimensional and non-linear dataset could improve experimental evolutionary process and provide important insights into the fundamental properties of evolution.

Machine learning (ML) has emerged as current state-of-the-art methods for analyzing patterns and relationships within complex high dimensional dataset (Greener *et al.*, 2022). Machine learning is a broad class of computational algorithms that 'learn' to perform a predictive task from a set of sample data (Figure 3-1). These input data usually have a large number of features. ML models try to fit a function that can describe the relationship between these features and the output that indicate the specific task being modeled. The process of function fitting is done through minimization of a pre-specified loss function depending on the tasks which are most often categorized as either regression or classification. Regression tasks are one that make a prediction of continuous value such as predicting enzyme catalytic rate from sequence (Li *et al.*, 2022). While classification tasks predict a class label either binary or multicategorical labels such as predicting carcinogenicity from skin lesion images (Esteva *et al.*, 2017). The biggest bottleneck in utilizing ML models is data availability, which in the biological domain can be expensive and labor intensive to obtain.

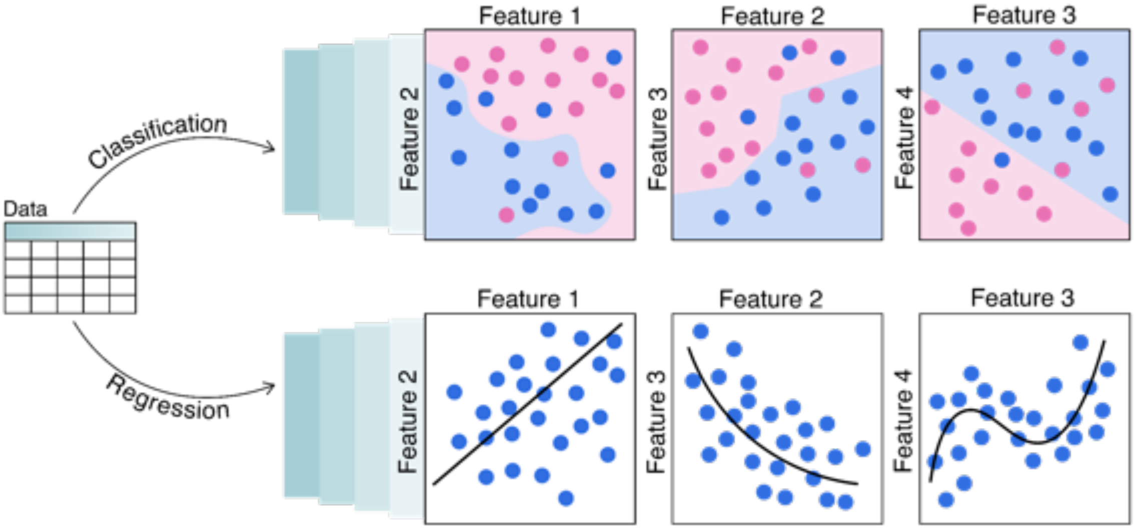


Figure 3-1: High-level overview of machine learning.
The field of machine learning (ML) aims to make predictive models about data to solve a variety of tasks which can be broadly categorized into either classification or regression. Regression involves predicting a continuous value like protein activity range. Classification task aims to assign discrete labels to a set of variables such as predicting protein families based on amino acid sequences. ML models perform these tasks by establishing relationships between features of the dataset the model is trained on. These models are trained to fit the function that best describes the feature relationship by using a training process that involves minimizing a task-specific loss function.

With the current era of big data, machine learning models, especially deep neural networks, have led to significant advancement in computer vision and natural language understanding. Similarly, with the arrival of high-throughput experimental methods for data collection, major breakthroughs in biology have also been solved with machine learning, with a notable example being the massive success of AlphaFold 2. AlphaFold 2 is a supervised deep neural network model that leverages the protein structure database as well as genomic data to predict protein three dimensional structure from sequence at experimental accuracy (Jumper *et al.*, 2021). This technology represents a paradigm shifting method that is already transforming our understanding of cellular biology, drug discovery and protein engineering. The success of machine learning models in such wide-ranging fields meant that adoption of these models for exploring the molecular fitness landscape could lead to similar breakthroughs and success.

Multiple approaches can be taken for the applications of machine learning to the study of fitness landscapes. The approach taken by different machine learning algorithms can be broadly categorized into supervised or semi/unsupervised models. Supervised models are trained on data with continuous labels such as protein thermostability or categorical labels such as enzyme family. Much more work has been done in using machine learning to explore the sequence space of proteins compared to RNA. The Arnold lab pioneered the use of machine learning combined with experimental evaluation to guide the directed evolution of proteins (Figure 3-2a). They applied a kernel based method called Gaussian process (GP) to explore the fitness landscape of cytochrome P450 (Romero, Krause and Arnold, 2013). The model was trained on hundreds of data points to guide a combinatorial library toward sequences with higher thermostability. Later, similar models were used to engineer channelrhodopsin protein with higher light-sensitivity and acyl-ACP reductase with double the fatty alcohols yield compared to the wildtype (Bedbrook *et al.*, 2019; Greenhalgh *et al.*, 2021). Gaussian processes were used for these works due to its data-efficiency. With Gaussian processes, a limited number of data points (in the hundreds) can yield successful model training and the Bayesian nature of GP models lend itself well to protein engineering tasks where knowledge of uncertainty can help guide experimental decisions (Hie, Bryson and Berger, 2020). Other supervised models, including deep neural networks (DNN) have also been successful in understanding protein sequence space (Wu *et al.*, 2019; Gelman *et al.*, 2021; Gonzalez Somermeyer *et al.*, 2022). However, DNN models required very large experimental datasets that are only available for a few classes of proteins due to experimental limitations. To circumvent these limitations, there has recently been a surge in the development of large language models trained on all known protein sequences (Alley *et al.*, 2019; Bepler and Berger, 2021; Rives *et al.*, 2021; Brandes *et al.*, 2022; Ferruz, Schmidt and Höcker, 2022). These models use an unsupervised or self-supervised learning regime where millions of unlabeled natural protein sequences are used to train an extremely large language model. These models learn to encode the sequences into latent or hidden spaces that capture the protein sequences as continuous representation that has been shown to correspond to features like structure, phylogeny and biochemical properties (Figure 3-2c). These representations have been used to augment limited experimental dataset and improve prediction accuracy for a wide range of protein engineering and structure prediction tasks (Rao *et al.*, 2020; Biswas *et al.*, 2021; Meier *et al.*, 2021; Chowdhury *et al.*, 2022; Hsu *et al.*, 2022). Because these models capture continuous representation of natural protein properties, an inverse design process has also been done to generate novel synthetic proteins by sampling from these natural representations (Madani *et al.*, 2023). This generative approach can be used to design novel functional proteins such as antibodies (Shin *et al.*, 2021) for therapeutic and biotechnological application (Figure 2-3d) (Ferruz and Höcker, 2022).

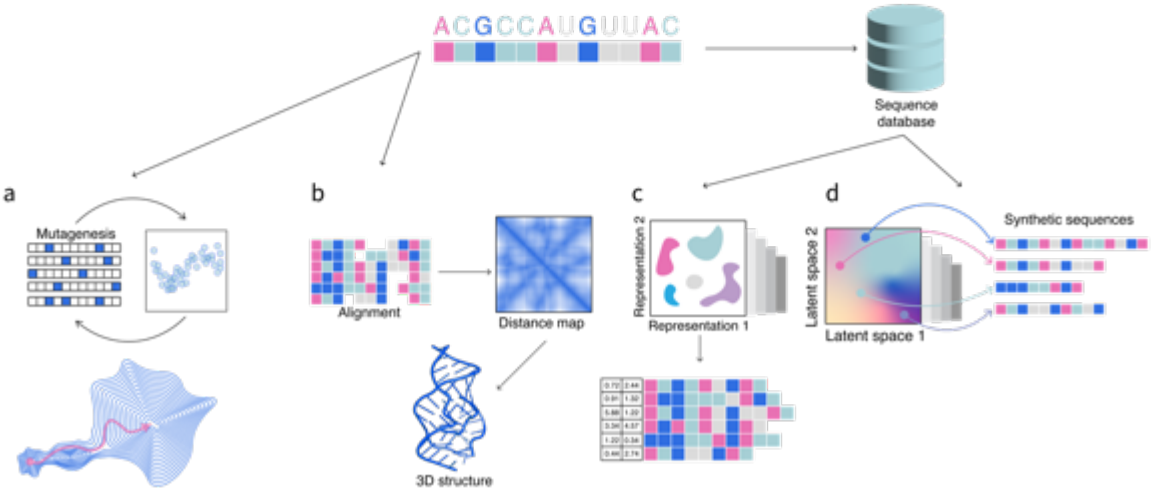


Figure 3-2: Applications of deep learning for biological sequence modeling. a) Supervised machine learning models have been used to fit a sequence-activity model by training on experimental mutagenesis data. The model can be used to prioritize sequences for directed evolution experiments. (Wu *et al.*, 2019) b) Several deep learning models have been used to predict RNA and protein 3D structure from sequence (Jumper *et al.*, 2021; Pearce, Omenn and Zhang, 2022). c) Large language models have been trained on protein or RNA sequence databases in a self-supervised manner. These models can encode sequences into a continuous representation or embedding space. These embeddings can be used to increase performance of downstream prediction tasks such as mutational effects prediction (Biswas *et al.*, 2021). d) Sampling from these large language models can be used to design new functional protein sequences not found in nature (Ferruz, Schmidt and Höcker, 2022; Madani *et al.*, 2023).

Although machine learning based methods have seen great success for tasks in protein sequence analysis, much less work has been done into utilizing a similar model for RNA or ribozyme sequences. The ribozyme system offer a major advantage over protein in that much larger library of mutants can be constructed and experimentally labeling these sequences with activity value is relatively easy using high-throughput sequencing based assay (Pitt and Ferré-D'Amaré, 2010; Blanco *et al.*, 2019). Therefore, ML based analysis of fitness landscapes could potentially benefit more through access to larger areas of the sequence space offered by the ribozyme systems. One of the earliest applications of the ML model to predict functional RNA behavior was from the Suess lab (Groher *et al.*, 2019). They used a random forest model combined with a convolutional neural network trained on a combinatorial library of a tetracycline riboswitch to predict a variant with high dynamic range, achieving over 40-fold increase in switching activity. Later works from the Church and Collins lab used a dataset of over 90,000 toehold riboswitch sequences to train CNN and long short-term memory (LSTM) recurrent neural network models that can accurately predict switching behaviors (Angenent-Mari *et al.*, 2020; Valeri *et al.*, 2020). More recent works have also used deep neural networks to predict the activity of self-cleaving ribozymes. Schmidt and Smolke used a CNN model to predict the in vivo regulatory activity of hammerhead ribozymes (Schmidt and Smolke, 2021) by incorporating predicted secondary structure information and using single-cell fluorescence measurement. Another study evaluated the performance of random forest and LSTM for the prediction of in vitro self-cleaving activity of CPEB3 ribozyme (Beck *et al.*, 2022). In this study, the author showed that the predictive performance of the models decreases for sequence with higher Hamming distance from the wildtype. The likely reason for degradation in performance is the lack of data points for mutants with higher number of mutations due to combinatorial explosion. This highlights the major problem in using supervised models to predict the properties of RNA and protein sequences, that despite the performance gains from experimental dataset, that can be as much as hundreds of thousands of sequences, this still represent a minuscule portion of the total sequence space of any RNA or protein with meaningful length. How well the models trained on such a limited dataset will generalize to distant regions further away from the wildtype remain to be seen. The studies highlighted here used a dataset generated from a semi-random combinatorial library (Figure 3-3). Therefore, most of these libraries represent very local and sparse sampling of the sequence space. Random sampling of the sequence such as this will mostly yield deleterious variants that provide little information about the functional space of the landscape. Therefore, getting information about functional mutants further away from the wildtype is difficult without a smart strategy that can guide the sampling of the fitness landscape.

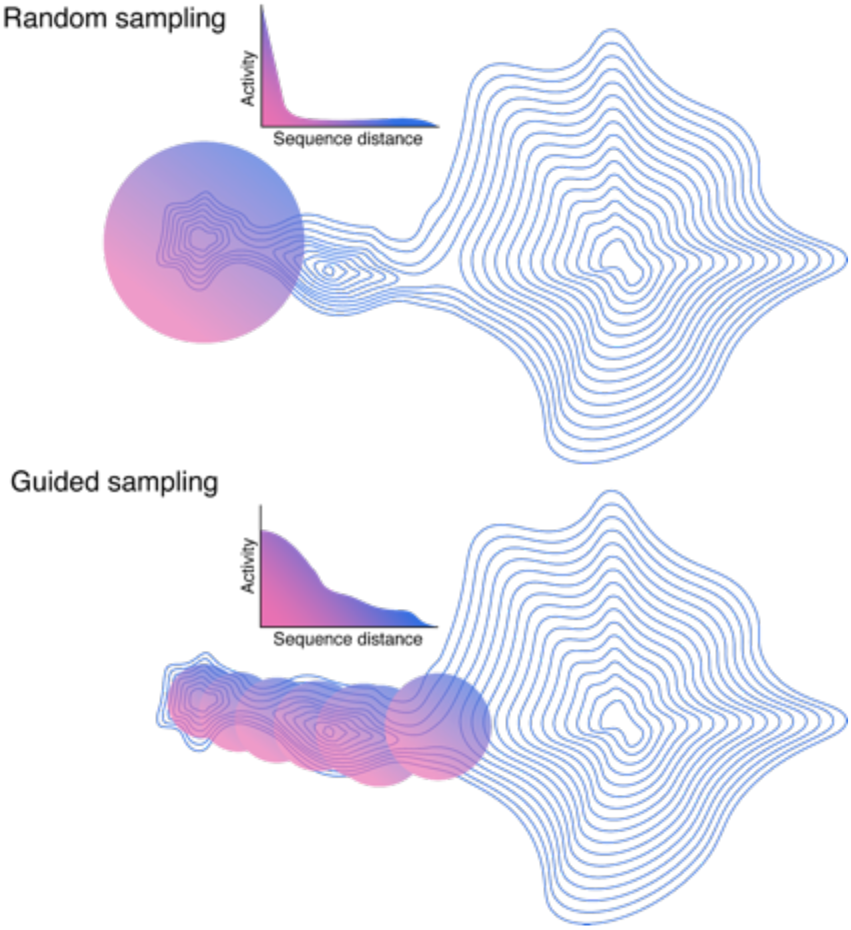


Figure 3-3: Guided-sampling can generate a more balanced sequence-activity dataset of a fitness landscape.
Because peaks are isolated in the fitness landscape, random sampling can lead to an extreme case of class imbalance where most sequences have almost zero activity. Guided sampling of the fitness landscape using approaches such as experimental directed evolution or genetic algorithm can lead to a more balanced dataset. These datasets are more enriched in sequences with appreciable activity and can be used to train machine learning models more effectively.

Better ability to predict functional variants further for the wildtype is important for two reasons. Firstly, a better understanding of the overall topology of the fitness landscape can provide important insight about molecular evolution. Secondly, in molecular engineering variants with higher or novel activity might be in regions currently out of reach for current ML models. In the last chapter, I used genetic operators combined with high-throughput sequencing assay to obtain functional variants of the F1 ligase ribozyme. This dataset represents a much more balanced and informative sampling of the sequence compared to random sampling. The dataset contains a high proportion of functional mutants including ones with a high Hamming distance to the wildtype. Therefore, supervised model trains on this dataset should be able to capture information about distant regions better than models trained on random combinatorial libraries. Furthermore, the superior combinatorial optimization capability of genetic algorithms lend itself well to guiding ML predictions towards combinatorial regions more likely to contain the optimal solutions. There have been a few works that combined genetic operators with prediction of ML models to guide the exploration of biological sequence space. Some examples include designing 5'UTR with high ribosome load (Sample *et al.*, 2019), predicting the effectiveness of DNA promoter sequences (Vaishnav *et al.*, 2022) and antimicrobial peptide discovery (Boone *et al.*, 2021). However, no studies so far have applied such a strategy to explore the fitness landscape of ribozymes.

In this chapter, I evaluate the performance of supervised machine learning models on capturing the functional landscape of the F1*U ligase ribozyme. Using the large screening data obtained from the works done in chapter 2, I trained several supervised ML models to perform a binary classification task, categorizing unseen ribozyme sequences into neutral or deleterious mutants. I then validate the model predictions with further experimental assay. After experimentally confirming that the model could make predictions with high accuracy, I combined a deep neural network model with the genetic operators introduced in the previous chapter. By replacing the experimental assay step with model prediction, I computationally evolved the ligase ribozyme population and experimentally show that this population contains several functional ligase ribozymes that have as many as 17 mutations compared to the wildtype.

Results & Discussion

Machine learning models can identify functional ribozyme variants

In order to gain more insight into the functional space of fitness landscapes, we need an approach that can efficiently identify functional variants within distant regions. Although genetic operators can increase the proportion of functional mutants in the population, as shown in the previous chapter, the process is limited to identifying variants closer to the starting point, the wildtype. Genetic operators alone require certain elements of randomness where random combinations of mutations are created and expected to be functional based on the fact that their parents are functional. Genetic operators alone do not evaluate the degree of epistasis within the landscape. Epistasis is observed when mutational effects combine in a non-linear way resulting in rugged landscape with reduced predictability (Domingo, Baeza-Centurion and Lehner, 2019). Epistasis can occur at multiple orders according to how many mutations are interacting. Using deep sequencing, information about local epistasis such as pairwise epistasis can be evaluated relatively easily for small sequence spaces such as the catalytic core of the F1*U ligase ribozyme. However, several pieces of evidence have been presented that higher-order epistasis, involving 3 or more mutational interaction, is prevalent in fitness landscapes and contribute to the overall navigability of the sequence space (Weinreich, Watson and Chao, 2005; Poelwijk *et al.*, 2011; Weinreich *et al.*, 2013; Domingo, Diss and Lehner, 2018). Therefore, in order to identify functional variants in the distant region of the fitness landscape, a navigation process needs to be able to evaluate epistasis at higher order. Relying on genetic operators alone, I would need several repeated rounds of experimental screening in order to gain enough activity information to assess higher-order epistasis (Figure 3-4). This can quickly become experimentally costly and intractable. Furthermore, because genetic operators have no notion of global fitness, it can quickly become trapped in a local optimum if that optimum is particularly well isolated. Therefore, better evaluation methods are needed that can identify patterns of epistasis from relatively local dataset that can hopefully be used to identify functional variants at distant regions.

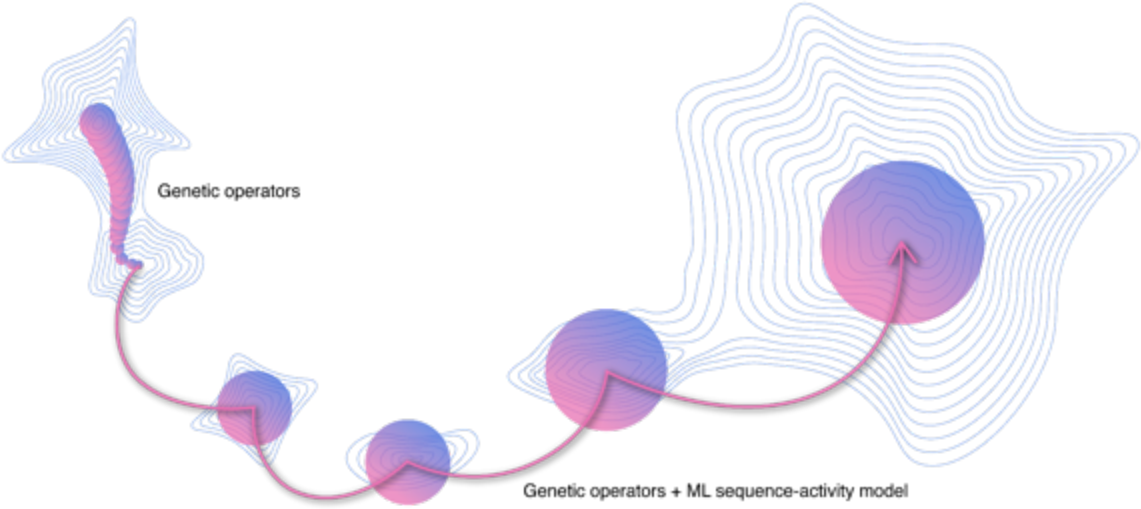


Figure 3-4: Machine learning model can guide evolutionary process away from local optimums.
Genetic operators do not have knowledge of the global fitness of the landscape. Therefore, the algorithm does not know how to sacrifice short term gain for long term fitness, as the fittest solution is only relative to the solution it has seen. As a result, genetic algorithms tend to converge to a local optimum if that optimum is particularly well isolated. In order to guide the genetic operators away from the local optimum and to minimize the amount of expensive experimental assay, machine learning models can be used as a fitness function. A machine learning model could possibly capture complex relationships between sequences and predict distant variants with higher fitness, guiding the algorithms towards a global fitness peak.

Machine learning methods have been shown to capture complex high-dimensional patterns in images, language and protein sequences given enough training data. With my current dataset comprising 29,887 unique ribozyme sequences and their activity label collected over 6 generations, I should have enough data to effectively train and evaluate supervised machine learning models. Because it is difficult to know at the beginning what type of models are suited to the type of fitness landscape being analyzed, I decided to evaluate several popular supervised models, with different degrees of complexity, for their ability to predict unseen parts of the sequence space using a held-out dataset. Five models were evaluated in total (Figure 3-5). As a baseline, logistic regression (LR) and support vector machine (SVM) with linear kernels were used. These models can only capture independent mutational effects and were selected as simple benchmarks. The k-nearest neighbor (k-NN) and gradient-boosted decision trees (GBDT) are powerful nonlinear models that can learn complex interactions, such as epistasis in the data. Finally, an MLP is a neural network model that can potentially learn complex nonlinearities, such as higher-order epistasis. These models were trained as binary classification models, where mutants are categorized into neutral or deleterious using a relative activity (RA) threshold of 0.2, as described in the previous chapter. I decided to train the models as binary classifiers as the goal of the study is to identify functional variants within the fitness landscape. As long as the variants have appreciable activity their absolute activity does not matter; that is why a regression model was deemed unnecessary. Each model was trained using a training dataset consisting of 20,920 unique sequences and evaluated on a held-out dataset consisting of 8,967 sequences. Representing a 30/70 random split of the entire dataset from generation 1 to 6.

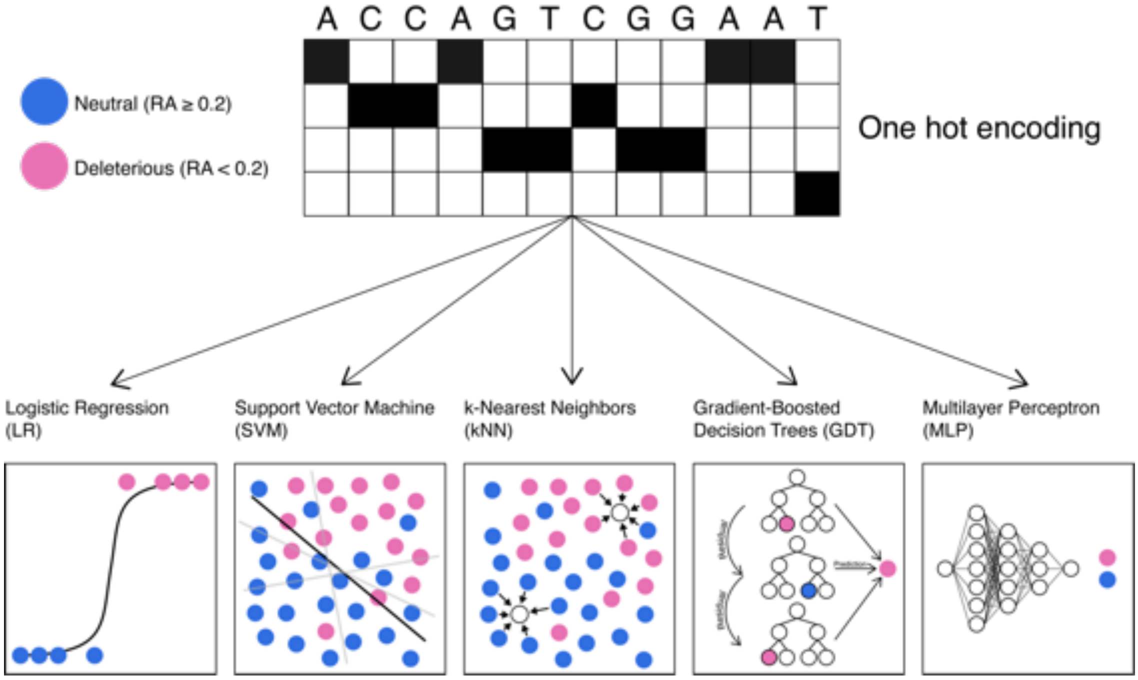

A C C A G T C G G A A T
Neutral (RA ≥ 0.2)
Deleterious (RA < 0.2)
One hot encoding
Logistic Regression (LR)
Support Vector Machine (SVM)
k-Nearest Neighbors (kNN)
Gradient-Boosted Decision Trees (GDT)
Multilayer Perceptron (MLP)

Figure 3-5: Overview of 5 machine learning models evaluated in this study. Schematic of the machine learning model evaluated in this study. Ribozyme sequences are first one-hot encoded into a 4x*L* vector. Logistic regression (LR), support vector machine with linear kernel (SVM), k-nearest neighbors (k-NN), gradient boosted decision trees (GBDT) and multilayer perceptron (MLP) were trained to to perform binary classification of the sequence into neutral or deleterious.

To select the model to be incorporated into the algorithm, precision and recall were used as performance metrics. Precision is the fraction of positive (neutral) predictions that are true positive. This represents the probability that the variants predicted to be neutral are actually neutral when tested experimentally. Recall is the fraction of neutral mutants in the experimental data identified by the model. The precision-recall curve of each model is plotted in Figure 3-6a. SVM, GBDT and MLP all have comparable performance to each other when looking at the precision-recall curve. Looking at the precision and recall scores of each model when the test dataset is separated by Hamming distance to the WT revealed that MLP narrowly beats other models at recalling mutants with higher Hamming distances while maintaining good precision (Figure 3-6b). Because of the rarity of functional mutants, especially at higher Hamming distances, I focused on recall as the main metrics while sacrificing precision. Therefore, the MLP model was selected for further evaluation. Although other models also perform comparably well, especially GBDT. The models evaluated here have varying degrees of complexity and use different approaches to model the dataset. The models were evaluated using the default hyperparameters from the scikit-learn package, and the performances were compared using train-test split strategy. This approach does not guarantee that the selected MLP model is the best type of model for ribozyme landscape analysis. It is possible that the other four models that were evaluated simply have a default set of hyperparameters that is not yet optimized for ribozyme landscape analysis. Further optimization of these hyperparameters might yield improved performances that are comparable to the MLP. However, the goal of this current study is to find the best performing model that can guide evolutionary process towards distant functional region rather than a complete evaluation of ribozyme fitness landscape modelling strategy. Therefore, further optimization of the other models was deemed unnecessary at this point. Furthermore, direct comparison between the five models with vastly different properties and complexity is a challenging task. If the goal is to select the least complex model that can best describe the fitness landscape, then probabilistic model selection criteria such as Akaike or Bayesian Information Criterion can be used to measure the tradeoff between model performance and complexity. Although these metrics can be used to select amongst parametric models such as logistic regression, its use for non-parametric models such as Decision Trees is not obvious. Decision Trees models are not fitted using maximum likelihood and do not have a well-defined number of parameters like a logistic regression model, where the number of parameters is simply the number of coefficients. The number of parameters and the likelihood function are required to calculate the Akaike or Bayesian Information Criterion. Therefore, direct application of these criteria to the five models presented here would not provide a very meaningful comparison. In the later section of this chapter, I provide an evaluation of three different neural network models that vary in the degree of complexity. I optimized the hyperparameters of the three models using an exhaustive grid search. This provides an initial comparison and evaluation of the effects of model complexity on the ability to capture the ribozyme fitness landscape. Nevertheless, the different models presented in this section could potentially be used to guide the navigation of fitness landscape as well and future works should focus on a systematic evaluation of model architectures which can best capture fitness landscape properties.

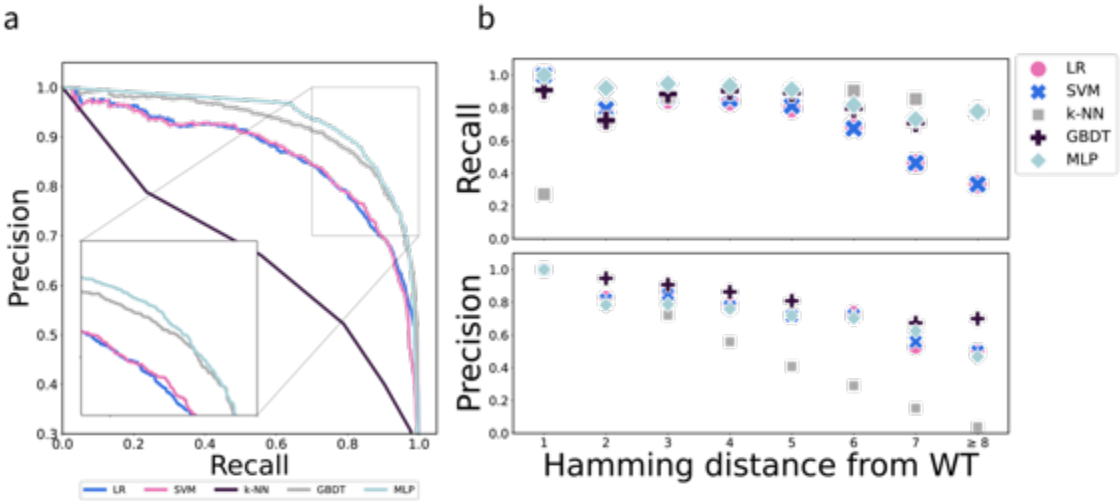


Figure 3-6: Evaluation of machine learning models performance.
a) Each model was trained with 20,920 variants from generations 1–6. Precision and recall curve was plotted by evaluating the model on a held-out testing set of 8,967 variants. b) Precision and recall calculated separately for sequences in the held-out testing set sorted according to the Hamming distance from the wildtype.

Next, I proceed to experimentally validate the MLP performance at identifying functional mutants within the fitness landscape. I designed three populations of variants named generation 7a, 7b and 7c. Generation 7a was designed using the genetic operators like generation 6 but with a higher proportion of variants created from recombination alone without any point substitution (See Methods). In generation 7a, 80% of the population were pure recombinants compared to 66% in generation 6. The reason for this is that I was confident that recombination would lead to a higher proportion of functional mutants. Generation 7b was created in the same way as generation 7a, except each variant was only picked if it was predicted to be neutral by the MLP model trained on the previous 6 generations. Generation 7c was created by taking 7b and shuffling (recombining) the population with itself with an average of 10 recombination events per variant. Each variant in generation 7c was again only picked if predicted to be neutral by the MLP. Generation 7c was created to test the performance of MLP at identifying variants at higher Hamming distances.

These 3 populations were then experimentally evaluated in the same way as the previous generations. As indicated in Figure 3-7a, the fraction of neutral mutants increased in generation 7a to 74%. This increase in neutral variants could potentially be explained by the higher pool of functional parents in generation 6 and increased in the proportion of pure recombinants in the population. The use of MLP increased the fraction of neutral mutants even further, to 89% in generation 7b. Impressively, even when the average Hamming distance in the population increases in generation 7c (Figure 3-7c), the fraction of neutral mutants is still very similar to generation 7b (Figure 3-7a), indicating that the MLP model was able to maintain performance and generalize well to unseen parts of the landscape. Furthermore, Figure 3-7b shows that the distribution of RA in generation 7a is bimodal with many mutants having RA values close to zero. In comparison, the variants in generation 7b and 7c have RA values closer to one, with much less ribozyme variants having RA values closer to zero. These results show that MLP models can be effectively used to eliminate low activity variants from the population.

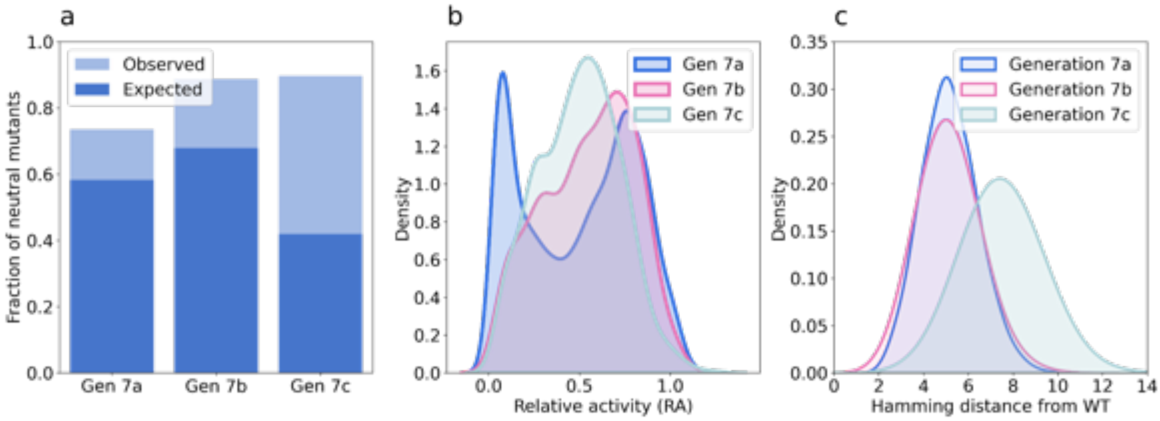


Figure 3-7: MLP-augmented genetic operators identify higher proportion of neutral variants which are distant from the wildtype.
a) Fraction of neutral (RA ≥ 0.2) mutants identified by different design approaches. In Generation 7a, only computational selection, recombination, and mutation were utilized. In Generation 7b, on top of the computational genetic process, MLP model trained on generation 1 to 6 was used to select only variants predicted to be neutral. Generation 7c involved shuffling Generation 7b with a 10x recombination rate, and only variants predicted to be neutral by MLP were selected. The darker area indicates the fraction of neutral mutants that was also identified by the epistasis-free log-additive model. b) The distribution of RA in each of the populations indicates that MLP can significantly reduce the proportion of mutants with very low activity. c) Comparison of the diversity of the three populations as indicated by the distribution of sequences according to their Hamming distance from the wildtype (WT).

Accessing distant regions with ML-guided evolution

Because of the enormity of the sequence space, navigating it by using only experiments would become costly and intractable. Therefore, it is desirable to have an automatic computational algorithm that can accelerate this process. The scale of the sequence space (4^{35}) means that simply using the model to predict all possible sequence combinations and picking the functional variants would be computationally intractable. Therefore, I need an optimization strategy that will guide the search towards the sequence space most likely to be functional. Genetic algorithms lend themselves naturally well to this task. Originally inspired by natural evolution itself, GA can balance the sequence diversification process of mutation, selection and recombination to adapt the searching process heuristically to regions of the sequence space it is exploring. Preserving sequence and structural constraint while exploring potentially new sequences. I have shown in the previous chapters that genetic operators can indeed be used to explore the landscape more effectively. Therefore, I chose to see if genetic operators can be combined with the trained MLP model and create a completely computational evolutionary algorithm to explore vast regions within the fitness landscape (Figure 3-8).

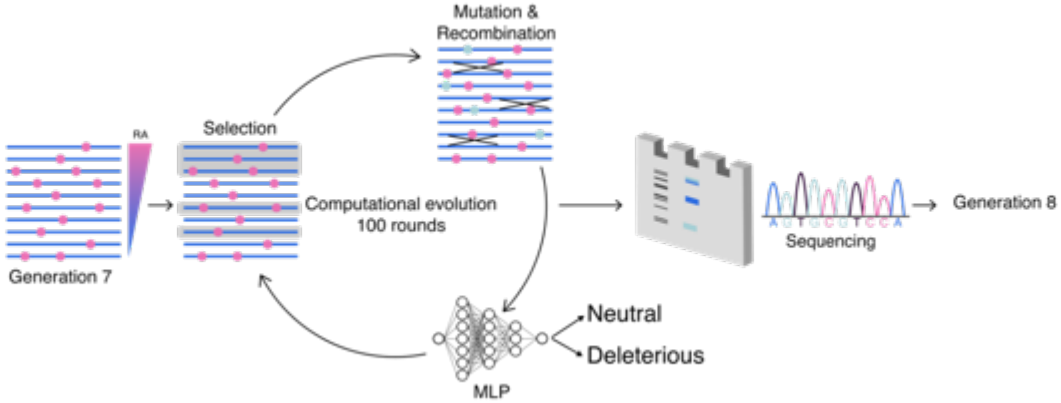


Figure 3-8: Overview of MLP-guided computational evolution.
Schematics of the computational evolutionary algorithm. Generation 7 was used as the initial population which then underwent 100 rounds of computational selection, mutation, recombination and classification by the MLP model. The final population is selected by ensuring all sequences are predicted to be neutral by the MLP model. This population was then experimentally evaluated and named generation 8.

The MLP model was retrained using all the data collected from the previous 7 generations. I used the trained MLP model to replace the experimental evaluation step in the current procedure. Mutants are evaluated by the prediction of the MLP model, and the binary classification is used by the selection process (Figure 3-8). The detailed flow chart of this algorithm can be found in Figure S3-1. Using generation 7a, 7b and 7c as the starting population I performed 100 rounds of selection, recombination, mutation and MLP classification. In the final round, I picked 12,000 sequences classified as neutral by the MLP and experimentally evaluated this new library, naming it generation 8. Experimental results show that the fraction of neutral mutants in generation 8 is 0.28 (Figure 3-9a). This is a decrease in performance from generation 7. Furthermore, the average RA between variants in generation is also much lower than generation 7 (Figure 3-9b). However, generation 8 occupies a sequence space that is much more distant than the previous generations, with an average Hamming distance to the WT of 13 mutations (Figure 3-9c). The model was trained primarily on variants with average hamming distance of 8 and has only seen functional variants with a maximum of 12 mutations. The fact that the model was able to identify functional variants with as much as 17 mutations represents a good generalization capability (Figure 3-9d).

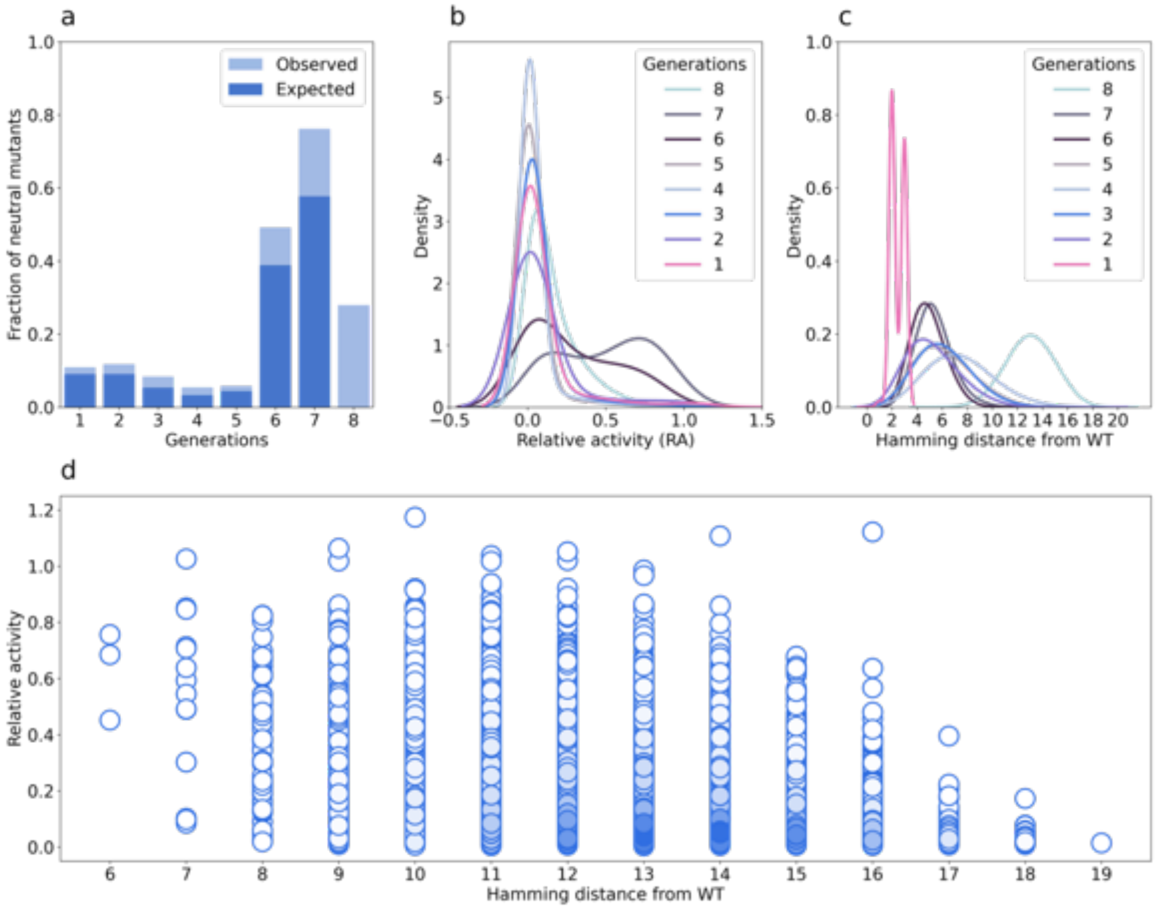


Figure 3-9: Computational evolution identified neutral variants further away from the wildtype.
a) The fraction of neutral (RA ≥ 0.2) mutants in generation 8 compared to all previous generations. b) The distribution of activity in each generation of ribozymes as determined by the sequencing assay. c) The diversity of each generation indicated by the distribution of each sequence according to their Hamming distance from the wildtype (WT). d) The RA of each sequence in generation 8 plotted against their Hamming distance from the wildtype.

Additionally, I assessed the model performance by looking at the level of epistasis present within the landscape that was explored. Using a log-additive model introduced in the previous chapter, I calculated the fraction of neutral mutants within generation 7 and 8 that are also expected to be neutral if epistasis is absent. I can see that the fraction of expected neutral variants is high in generation 7 (Figure 3-9a). However, the expected fraction is almost zero in generation 8. This suggests that the presence of epistasis would have limited the ability of the genetic operators alone to identify neutral variants at further distance of the landscape without periodically feeding it more information from experimental assay. The introduction of MLP reduces this experimental burden and can help push the genetic operators toward further regions within the sequence space while maintaining good efficiency. The fact that the variants identified by the MLP in generation 8 are not possible to identify with the log additive model also suggest that deep neural networks can capture higher-order epistatic information within the dataset to a certain degree. This offers a promising strategy to use deep neural networks to access regions within fitness landscapes not accessible with the current experimental capability.

Post-hoc analysis of neural networks performance in capturing ribozyme sequence-activity relationship from experimental data

The ability to predict activity of RNA or protein enzymes from sequence alone is important for both biotechnology and evolutionary biology. Deep learning models trained by examples obtained from experimental mutational scan data have shown promising results in reaching this goal (Romero, Krause and Arnold, 2013; Angenent-Mari *et al.*, 2020; Valeri *et al.*, 2020; Aghazadeh *et al.*, 2021; Gelman *et al.*, 2021; Luo *et al.*, 2021; Schmidt and Smolke, 2021; Song *et al.*, 2021; Beck *et al.*, 2022). However, many studies, including this work, evaluate their models' performances by using random split of data into training, validation and testing sets. This means that the high performance as reported by these studies were often achieved by evaluating the model with sequences having the same distribution as the training data. Some studies attempt to evaluate their model generalization to unseen mutations either by training and testing on a specific subset of mutants with different Hamming distances to the reference sequence (Luo *et al.*, 2021; Beck *et al.*, 2022). Another approach is to test the model on a set of variants containing particular mutations or mutated positions which are not in the training set (Gelman *et al.*, 2021). However, even these approaches fail to properly evaluate the generalization capability of the model because the available dataset often has relatively low diversity. Most mutants in the test dataset used by these studies have low Hamming distance to wildtype and the rare variants with high Hamming distance are often completely inactive. This lack of diverse and balanced dataset, as a result of the sparsity of fitness landscape, means that relatively little work has been done to assess how well supervised models generalize to diverse regions of the fitness landscape.

The dataset I have produced during 8 rounds of high throughput experimental assay represent a diverse library of ribozymes with high quality and quantitative experimental labels. The guided process using genetic operators also helps the dataset to be relatively more balanced than many other sequence-activity datasets published thus far. Particularly, generation 8 contains a high proportion of active variants with average mutational distance much further away from other generations, providing a high-quality out-of-distribution testing set for supervised models. Therefore, I use this opportunity to do a proof-of-concept study on how well a neural network model can generalize beyond its training dataset. I follow a similar approach to the work done by Gelmen et al. on evaluating neural network performance on protein deep mutational scanning data (Gelman *et al.*, 2021). Briefly, three different models were evaluated which are logistic regression (LR), multilayer perceptron (MLP) and convolutional neural network (CNN) representing varying degrees of complexity and capacity (Figure 3-10). Logistic regression model is used as a baseline as it can be thought of as a neural network with a single node connected to all the features in a variable (in this case is the identity and position of a one hot encoded ribozyme sequence). Because logistic regression is a linear model, it can only capture additive effects of mutations and would not perform well if there is a high prevalence of epistasis in the fitness landscape. On the other hand, multilayer perceptron can capture epistatic effects by using multiple layers of fully connected neural nodes with a non-linear activation function in each node. Finally, convolutional neural networks can capture high-level patterns of mutation by using and sharing convolutional filters. Convolution filters can capture general local patterns of mutation and the fully connected layers can integrate this information to form a high-level understanding of the pattern of mutation in the dataset. In this way, CNN model could potentially generalize to unseen part of the sequence space than logistic regression or multilayer perceptron.

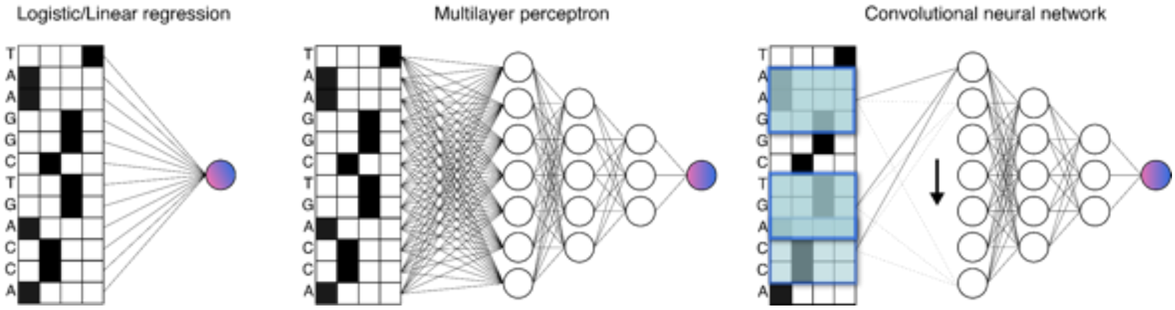


Figure 3-10: Supervised neural network models.
Schematics of the different architecture of three neural network models. Logistic/linear regression models capture additive effects of each feature which is the nucleotide identity in each position in a sequence. Multilayer perceptron integrates all the information and apply non-linearity through multiple hidden layers. Convolutional neural networks used sliding windows of convolution followed by pooling to extract high level features from the sequence.

The three models were first designed by tuning hyperparameters such as number of layers, number of neural nodes or learning rate using a grid search. The total hyperparameter search space is listed in Table S3-1. Each set of hyperparameters is evaluated using stratified 5-fold cross validation on a binary classification task, which was predicting neutral (RA ≥ 0.2) vs deleterious mutants. The training data consist of all the sequences in generation 1 to 7 and mean F1 score (F1=2 * (precision * recall) / (precision + recall)) across the 5 folds was used to pick the best performing set of hyperparameters. After hyperparameter tuning, each model was retrained on the entire training dataset using the selected hyperparameters. Finally, the models were evaluated on the held-out testing set consisting of the entire generation 8. Evaluation of the model performances is shown in Figure 3-11. Evaluation of the model using area under curve (AUC) for precision-recall curve and receiver operating characteristic (ROC) curve shows that CNN outperforms all the other models (Figure 3-11 a, b and c). The CNN model outperforms the MLP model by a small margin when looking only at the AUC scores. However, when looking at the accuracy score the CNN model performs better than MLP by almost 20%. This can be explained by evaluating the confusion matrices for both models (Figure 3-11e & f) The confusion matrices show that MLP model has higher rate of false positives than CNN model leading to small overall accuracy. The CNN model's higher abstraction capacity could potentially enable it to generalize better to unseen dataset which could potentially explain its higher accuracy compared to the MLP. The confusion matrix for the logistic regression model also reveals a very high rate of false negatives. Logistic regression could only model the additive contribution of mutational effects. Considering that most mutations in a fitness landscape led to complete loss in activity, it is conceivable that a logistic regression model would predict most mutants to be non-functional. Additionally, I binned the testing dataset (generation 8) by Hamming distance to the wildtype and evaluated the model’s performance using F1 score on each subset. As evident in figure 3-11g, the F1 score decreases as a function of Hamming distance indicating that the models struggle to generalize to distant parts of the fitness landscape. The F1 score of CNN model is higher than the other models at almost all Hamming distances, confirming its superior predictive performance.

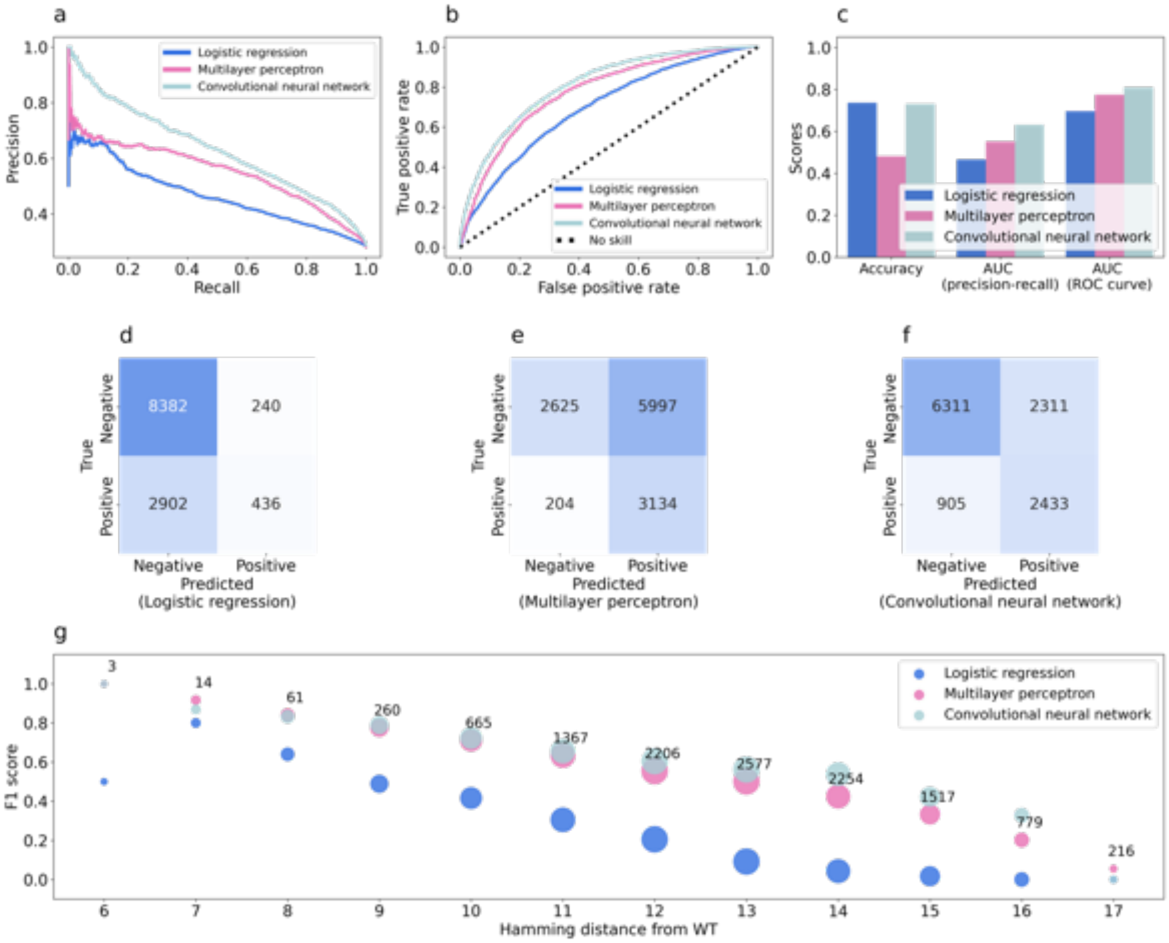


Figure 3-11: Evaluation of classification neural network model performance. Models were trained on data from generation 1 to 7 to perform binary classification of sequence into neutral (RA ≥ 0.2) or deleterious. Each model was evaluated on the entire generation 8 which was held out as a testing set. a) Precision-recall and b) receiver operating characteristics (ROC) curve were plotted for each model using varying classification thresholds. c) Comparison of the accuracy and area under curve (AUC) for both the precision-recall and ROC. Confusion matrices were plotted using a classification threshold of 0.5 for d) LR e) MLP and f) CNN. g) Sequences in generation 8 were binned according to their Hamming distance from the wildtype (WT). F1 score were calculated separately for each bin of sequences. The marker size and the annotation indicate the total number of sequences in each bin. Calculation for sequences which are 18 or 19 Hamming distances away from the WT are not shown as no neutral sequences are found for these Hamming distances.

In many cases, there is a need to distinguish between sequences with very similar activity. For example, finding the most highly active variants in enzyme engineering tasks. In this scenario a regression model will be more useful than a binary classification model. Therefore, I also evaluated the three models on a task of directly predicting the relative activity (RA) from sequence. The same parameter tuning scheme as for the classification model was used to select the best regression model. The models' performances were evaluated using Spearman's rank correlation when comparing the predicted RA and experimentally measured RA. Spearman's rank correlation was used as a performance metric to reflect the potential use case of these models for enzyme engineering tasks. In these tasks the main goal is to find the best variants relative to other variants and therefore prediction of absolute activity is not necessary. Looking at Figure 3-12 a, b and c shows that the CNN model performs the best. Looking at the mean squared errors when the test data is sorted into Hamming distance to the wildtype again show that performance decreases with Hamming distance, with the CNN model performing best at almost all Hamming distances (Figure 3-12d). However, the overall regression model performance is quite poor with Spearman's rank score of 0.47 for the CNN model.

The variants present in generation 8 have higher Hamming distance range than the training data and their activity is likely influenced by a complex network of higher-order epistasis which is hard for the models to capture. Some studies have been done which try to improve the generalization ability of machine learning models towards distant regions. One interesting approach incorporate the sparsity of epistasis directly into the loss function (Aghazadeh *et al.*, 2021). Another recent work uses information on how correlation of activity changes as a function of mutational distance as statistical priors for a model (Zhou *et al.*, 2022). This gives the model prior understanding of how the same mutation has different effects in different genotypic backgrounds. Both works provide important steps towards better predictive performance for higher order mutations. But as mentioned previously, these models lack high quality and diverse dataset that provide information about the distant part of the landscape. The approach I have presented here using genetic operators combined with high-throughput assay offers an effective strategy to collect better and diverse dataset. Generation 8, although occupying a much more distant region of the sequence space than the generation that was used to train the models, was still designed based on mutations present in the training dataset. Therefore, generation 8 might not be a truly out-of-distribution testing data. However, a more systematic utilization of genetic operators could produce a focused but diverse dataset that can be used to effectively train and evaluate models. This could lead to significant improvement in our ability to predict fitness landscape based on sparse sampling of the sequence space.

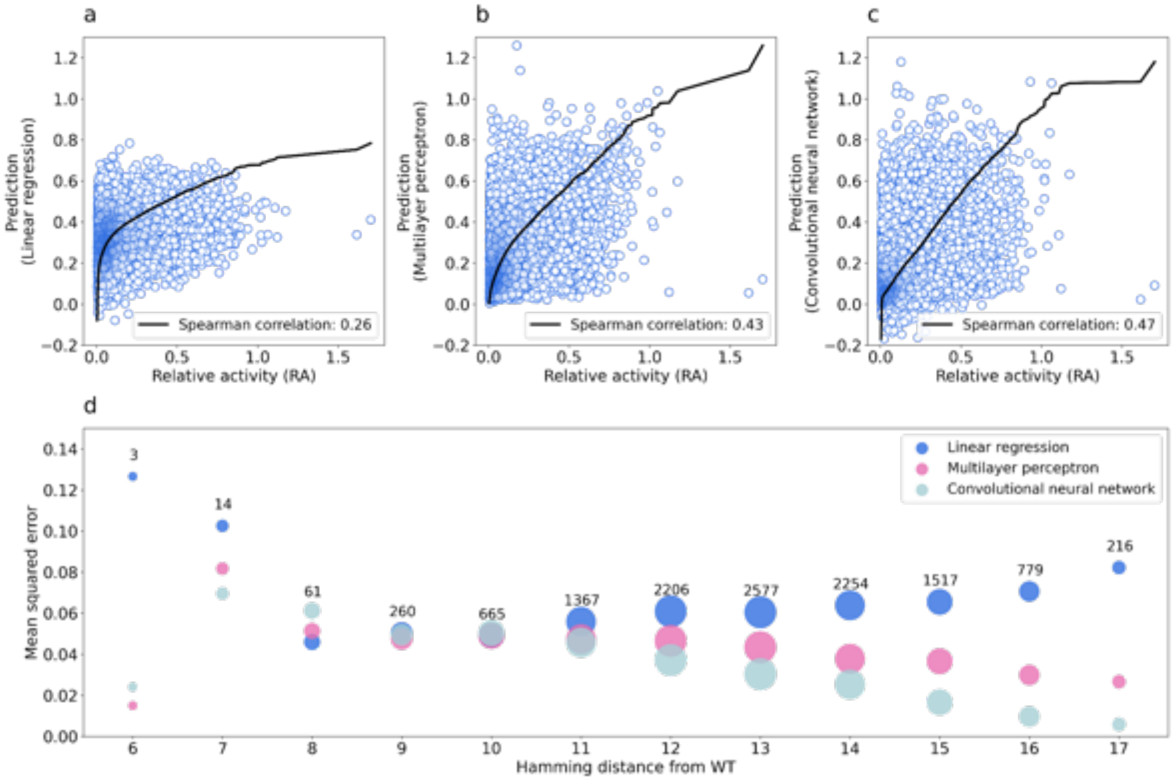


Figure 3-12: Evaluation of regression neural network model performance. The relative activity (RA) predicted by a) a linear regression model b) a multilayer perceptron and c) convolutional neural network are plotted against the experimentally measured values. The models were trained on data from generation 1 to 7 to perform a regression task of predicting RA from one-hot encoded sequence. Each model was evaluated on the entire generation 8 which was held out as a testing set. d) Sequences in generation 8 were binned according to their Hamming distance from the wildtype (WT). Mean squared errors were calculated separately for each bin of sequence. The marker size and the annotation indicate the total number of sequences in each bin. Calculation for sequences which are 18 or 19 Hamming distances away from the WT are not shown as no neutral sequences are found for these Hamming distances.

Conclusion

In this chapter, I evaluated the use of supervised machine learning models to learn the sequence-activity relationship of the F1 ligase ribozyme. I showed that by using genetic operators as a search strategy, deep neural networks can be used to efficiently navigate the fitness landscape of RNA by learning epistatic patterns within the dataset. These results add to the growing list of work that showed that machine guided directed evolution can be used to optimize protein and RNA engineering efforts for biotechnology applications. However, in this work I have also shown that this process can be used to explore the general properties of the fitness landscape. An ability to map the entire sequence space of a full-length RNA or protein can offer major insights in the evolutionary process and can lead to better understanding of how life originally evolved. One major goal in the study of the fitness landscape is to see how well we can predict future trajectories of evolution. Having this ability can solve many problems such as predicting cancer cell metastasis or to see how viruses will evolve and even predicting the trajectory of a pandemic. The works I have shown here that deep neural networks can be generalized beyond its training data towards distant epistatic regions of the sequence space offer an important step towards this goal. The predictability of fitness landscape is influence by many of its fundamental properties including evolvability, robustness and the size and distribution of its neutral network. In order to successfully adopt machine learning strategy within this domain, the field need a better understanding of how these different properties increases or limit the accuracy of the model predictions. In the next chapter, I investigate the connectivity of the mutational paths underlying the region explored by the ML-guided evolutionary algorithm. I evaluate the degrees of epistasis within this region and discuss how higher-order epistatic interactions effects the model accuracy and influence landscape navigability within the context of evolvability and robustness.

Methods

Machine learning models

Five machine-learning models for binary classification were trained using data from generations 1–6. All models made predictions by trying to fit a function that describes the relationship between the input features, which in this case is the position and identity of the nucleotide in each ribozyme sequence, and the class labels that are either neutral or deleterious. Each model is briefly described below.

Logistic regression (LR) is a linear model that assigns different weights to the input features. Predictions are made using a linear combination of the input features and their weights, followed by a sigmoid function that outputs the class probability. LR can only model the additive contribution of each mutation to fitness and therefore cannot model nonlinear interactions between positions.

A support vector machine (SVM) with a linear kernel assumes that the classes in the data are linearly separable in the feature space and attempts to draw a boundary line to separate them. The optimal solution was achieved by maximizing the distance between each class and the boundary line. SVM with a linear kernel can only model a linear combination of mutations, similar to LR, although the tendency to overfit with SVM is lower. Overfitting is observed when a model accurately predicts the training data but creates poor predictions for new data points.

The k-nearest neighbor (k-NN) method does not assume a linear separation of classes. The prediction for a new input is made based on the majority class of the *k* number of neighboring training points closest to the new input in the feature space. This allows k-NN to model nonlinearity better than SVM or LR, but it is more affected by noise in the data and is more likely to become overfit.

Gradient-boosted decision trees (GBDT) makes predictions by constructing a group of "trees" that branch out each time a condition for a feature is met (e.g., is position 23 in the sequence a guanosine?). The tree depth determines the complexity of these conditions for making the final decision regarding the class label. Gradient boosting is a technique that builds a large group of weak trees with shallow depths in an iterative fashion based on residuals from the previous tree. The final prediction is made by votes from the ensemble of trees on the final class label. This typically enables a higher prediction accuracy and less overfitting than individual trees or a small group of very deep trees like a random forest. The GBDT can model more complex nonlinear interactions than LR and SVM.

Multilayer perceptron (MLP) is a simple neural network model. A neural network consists of a group of individual "neurons" that take an input value and transform them using a nonlinear activation function. These neurons are arranged in fully connected layers, meaning that the output of one neuron becomes the input of the other neuron. This architecture allows a neural network to approximate any function. This means that MLP can potentially model mutational interactions or epistasis at a very high order better than the other models.

However, neural networks require substantially more data to accurately learn a function without overfitting.

Additionally, linear regression and convolutional neural networks were also evaluated in a post-hoc analysis of the dataset. Linear regression has essentially the same architecture as the logistic regression model but without the sigmoid activation function. Convolutional neural network (CNN) is essentially a regularized version of MLP that employs convolution operations to extract higher-level data from a dataset. Combining multiple layers of convolution followed by fully connected layers allows CNN model to achieve an increasingly abstract understanding of the sequence dataset. In this way CNN models could potentially achieve better generalization on unseen sequences than MLP models.

For training, the sequences were one-hot encoded and flattened (except in the case of CNN where the data is fed as a 35 × 4 binary vectors) into 1 × 140 binary vectors. Thirty percent of the dataset was used as the testing set, and the rest was used as the training set. The LR, k-NN, SVM, and GBDT were trained using the Python scikit-learn package. k-NN, LR, and SVM were trained using the default hyperparameters for the binary classification of sequences into neutral (RA ≥ 0.2) or deleterious (RA < 0.2). The GBDT was trained in the same manner using a maximum tree depth of 10. The MLP was written using the TensorFlow 2 Python library. The model consisted of three dense layers with rectified linear unit (ReLU) activation, batch normalization, and 20% dropout. The dense layers consisted of 128, 64, and 32 neurons, respectively. This was followed by a final dense layer with sigmoid activation for the classification output. The model was compiled using the Adam optimizer, with a learning rate of 0.005. Binary cross-entropy was used as the loss function. During training, 10% of the training set was used as a validation set, and the model was trained for 100 epochs with a batch size of 1024. All the trained model performances were evaluated on the test dataset using precision and recall as metrics. All codes were written in Python 3.9. The software libraries used were pandas 1.4.4, numpy 1.21.2, tqdm 4.62.3, scipy 1.7.3, matplotlib 3.5.1, seaborn 0.11.2, scikit-learn 1.0.2 and tensorflow 2.8.0.

Computational evolutionary algorithm

The MLP model was retrained using data from generations 1–7 in the same manner as described for **Machine learning models.** Model performances were also tested using 10-fold cross-validation (Figure S3-2). To account for class imbalance, I adjusted the prediction threshold using the receiver operating characteristic (ROC) curve. Prediction threshold that produced the largest geometric mean ($\sqrt{True\,positive\,rates \times (1 - False\,positive\,rates)}$) was used for subsequent classification by the model. Generation 7 was used as the starting parent population or the computational evolution. Tournament selection was used to select variants as parents. If more than one variant in the tournament was classified as neutral, then a random variant was selected. In each generation, 80% of the variants were created by recombination, and the remainder were created by point mutations. These variants were classified as neutral or deleterious mutants using MLP (Figure S3-1). This was repeated over 100 rounds, and the average Hamming distance in each round was tracked to ensure an increase in diversity (Figure S3-3). After 100 rounds of evolution, the mean Hamming distance plateaued at around 13. Increasing the number of rounds of computational evolution might lead to an increased average Hamming distance; however, this was slowed by the increased search space and a likely increase in false positives. For the last round, the total number of variants was increased to 12,000 to maximize the coverage of the sequence space for experimental screening, and only variants that were predicted to be neutral by the MLP were selected as generation 8. The parameters of the computational evolutionary algorithm are listed in Table S2-2.

Hyperparameter tuning for post-hoc analysis of neural network

Linear/logistic regression (LR), multilayer perceptron (MLP) and convolutional neural network (CNN) were designed for binary classification and regression tasks. Model hyperparameters were selected using the scikit-learn function GridSearchCV. In this scheme, the entire possible combination of selected parameters was tested for each model. The parameters and range used in the grid search are listed in Table S3-1. The combinatorial sets of parameters tested totaled 9 for LR, 81 for MLP and 72 for CNN. For each set of parameters each model was fitted 5 times in a 5-fold cross validation scheme. The training dataset, which includes the entire data from generation 1 to 7 after sequence reads and quality filtering totaling 39,864 unique sequences, was randomly split into 5 subsets. For each training the model is trained on 4 of the data subsets and validated on the one remaining subset. F1 and mean squared error score were used to validate the model performance for classification and regression task respectively. The best set of parameters was selected as the set with the highest average score across all 5 folds. The architecture of the final logistic regression model is one Flatten layer, one output Dense (Fully connected) layer with one hidden unit , learning rate of 0.001 and batch size of 128. The linear regression model has the same architecture as logistic regression except the one hidden unit in the output Dense layer has a sigmoid activation and the best learning rate was identified as 0.0001. For the MLP classification model the best architecture found was one Flatten layer, one hidden Dense layer with 128 hidden units with ReLU activation function, a Dropout layer with 20% dropout rate, an output Dense layer with one hidden unit and sigmoid activation, learning rate of 0.0001 and batch size of 32. For the MLP regression model the best architecture found was one Flatten layer, three hidden Dense layers with 128 hidden units with ReLU activation function, each hidden Dense layer is followed by a Dropout layer with 20% dropout rate, an output Dense layer with one hidden unit, learning rate of 0.0001 and batch size of 128. For the CNN classification model, the best architecture found was one 1D convolution layer with 128 filters of width 6 and ReLU activation function, one Max Pooling layer, one Flatten layer, one hidden Dense layer with 100 hidden units and ReLU activation, one Dropout layer with 20% dropout rate, one output Dense layer with sigmoid activation, learning rate of 0.0001 and batch size 128. Finally, the best architecture found for the CNN regression model was two 1D convolution layer with 128 filters of width 6 and ReLU activation function, each conclusion filter is followed by one Max Pooling layer, one Flatten layer, one hidden Dense layer with 100 hidden units and ReLU activation, one Dropout layer with 20% dropout rate, one output Dense layer, learning rate of 0.0001 and batch size 32. Each model training was done with an early stop callback with a patience of 15 epochs and minimum change in the loss (binary cross entropy for classification and mean squared error for regression) of 0.0001. The maximum epoch was set to 300 for all model training. The final model is generated by retraining the model using the selected parameters on the entire training dataset. Finally, each tuned model performance was tested and compared using a held-out dataset consisting of the entire generation 8 totaling 11,960 unique sequences.

Experimental measurements of ribozyme activity

The same experimental protocols were used as described in chapter 2 Methods section. The reproducibility between sequencing repeats is shown in Figure S2-1 and reproducibility between sequencing and PAGE are shown in Figure S3-4.

Chapter 4: Experimental RNA neutral network reveals many accessible paths towards a robust genotype

Parts of this chapter, in particular the Methods section, have been duplicated or updated from a previously published article: Rotrattanadumrong, R. and Yokobayashi, Y. (2022) ‘Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning’, *Nature communications*, 13(1), p. 4847.

Background

In the previous chapter, I investigated the use of ML-guided evolution to navigate the fitness landscape of a ligase ribozyme. The success of this process or any evolutionary process, artificial or natural, is highly dependent on the evolvability of the molecule which in turn is influenced by the topology and connectivity of the underlying fitness landscape. In particular, the presence or absence of a 'neutral network' has historically been an important indication of the accessibility of an RNA fitness landscape. In this chapter, I seek to understand how these key properties influences the predictability and navigability of the F1*U ligase ribozyme fitness landscape and whether the regions explored by the evolutionary algorithm in the previous chapter provide any important new insight into the general properties of RNA fitness landscape.

Evolvability can be defined as the ability of biological systems to produce selectively viable phenotypes for adaptive evolution (Payne and Wagner, 2019). The more evolvable a system is the more likely it is to be able to adapt or innovate. Evolvability research is very broad, with implications in fields ranging from cancer biology to population genetics. Understanding the cause and consequences of evolvability can lead to better ways to tackle antibiotic resistance (Sánchez-Romero and Casadesús, 2014) or engineering new biocatalysts (Bloom, Romero, *et al.*, 2007). Here, I focused on the role of evolvability in biomolecular evolution. As mentioned in the beginning of this thesis, molecular evolution can be conceptualized as mutational walks along the fitness landscape. In this scenario, the concept of evolvability is intrinsically linked to mutational robustness (Figure 4-1a). A molecule like RNA is considered robust if, given a certain selective pressure imposed by environmental conditions, it possesses many selectively accessible mutations within its fitness landscape. A mutation is accessible if it is connected to other mutations that maintain fitness via a single mutational step. In evolutionary theory study fitness refers to reproductive success. In the molecular realms, properties such as catalytic activities can be a proxy for 'fitness' if selective advantage is influenced by the ability to perform those catalysis. A population of biomolecules can gain selective advantage in a dynamic environment if they can rapidly acquire beneficial mutations for adaptive change. In a fitness landscape, this means that the population can quickly find and climb fitness peaks within the landscape (Figure 4-1b). The chance of finding fitness peaks can be increased by diversifying the population across the landscape, occupying various regions within the sequence space. By spreading out across different combinatorial regions a population could prepare for future adaptive events by maximizing the chance of locating close to a fitness peak. Therefore, an evolving molecular population that is robust should also have higher evolvability.

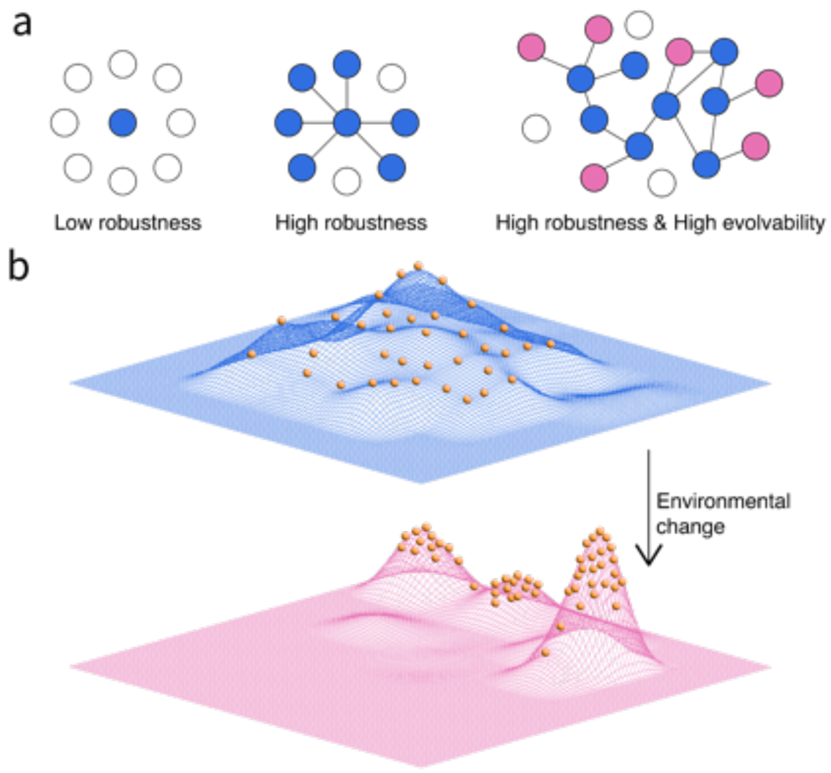


Figure 4-1: Robustness and evolvability enable adaptation in fitness landscape.

a) Schematic of the link between mutational robustness and evolvability adapted from (Whitacre, 2010). Each node represents a genotype and edge represent single step mutation. Node color represents different phenotypes. White nodes represent non-functional genotypes. When a functional genotype is surrounded by non-functional genotypes this means that the system has low robustness. While a highly robust genotype is surrounded by many functional genotypes that can be reached by single mutation. Higher robustness can lead to higher evolvability if functional genotypes form a single mutational network that can reach other new phenotypes without crossing a non-functional genotype. b) A large fitness plateau representing a robust network of functional or 'neutral' genotypes enables a population of evolving biomolecules to reach diverse regions of the landscape. When the environment changes, genotypes that are close to the new fitness peaks can rapidly adapt to this new landscape.

However, in order to access distant regions within the landscape, the population needs to acquire mutations that do not adversely affect its fitness in order to remain viable under the current selective environment. These mutations that are neither deleterious or beneficial, but maintain the current fitness are called 'neutral' mutations. The concept of neutral evolution, first formalized by Motoo Kimura (Kimura, 1968), postulates that evolution occurs at the molecular level and that most mutations are neutral. This theory emphasized the role of random genetic drift in driving genetic diversity. Genetic drift is an important cause of evolvability as a more diverse gene pool will be more ready for adaptation upon changes in the environment. Given that gaining neutral mutations directly affects the diversity of a population, understanding the availability and accessibility of neutral mutations in a fitness landscape therefore has important implications for the study of evolvability.

Fitness landscape provides a useful framework for the study of neutral mutations. Given a starting point in the sequence space, a population can travel far from this starting point by sequentially acquiring neutral mutations. These neutral mutations are accessible if they are connected to each other forming a 'neutral network' (van Nimwegen, Crutchfield and Huynen, 1999). The size, quantity and distribution of neutral networks determine how accessible a fitness landscape of a given molecule is and how easily diversification can happen. The larger and more connected the neutral networks are, the more quickly fitness peaks can be reached and the more evolvable the molecule becomes. In this way, the concept of neutral network reconciles the theory of neutral evolution, evolvability, robustness and fitness landscape in a useful and meaningful way (Wagner, 2008).

To map the neutral network, many genotypes need to have their fitness values assigned. Doing this experimentally for even a small protein or RNA is difficult and can quickly become intractable due to the combinatorial explosion of the sequence space. Therefore, until very recently, the study of neutral networks has been limited to computational and theoretical studies. An important technological milestone in this field is the invention of accurate RNA secondary structure prediction algorithms (Zuker and Stiegler, 1981). Given the sequence structure activity relationship of RNA, an assumption can be made that evolution that maintains the activity of a ribozyme also acts to maintain its structure. Therefore, the structure of an RNA can be reasonably used as a proxy for its fitness. RNA secondary structure prediction algorithms and packages such as ViennaRNA (Hofacker *et al.*, 1994) enable rapid and relatively accurate prediction for a large number of RNA sequences. This enables the determination of fitness for large regions of RNA sequence space, establishing RNA sequence-structure map as an powerful model for the study of neutral networks (Schuster *et al.*, 1997; van Nimwegen, Crutchfield and Huynen, 1999).

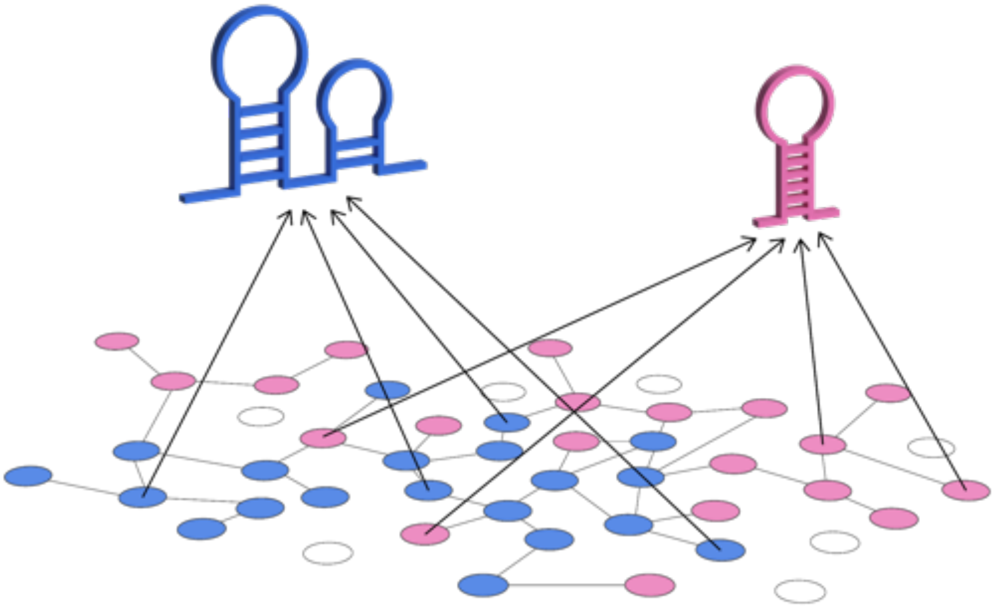


Figure 4-2: RNA sequence space forms connecting and overlapping neutral network that shares the same secondary structures.
Invention of relatively accurate and fast RNA secondary structure prediction algorithms enable large scale mapping of RNA sequence-structure maps. Many RNA sequences (represented by nodes in the schematics) are connected by single mutation (represented by edges) to other sequences that are predicted to fold into the same secondary structure (represented by node color).

Schuster and Fontana pioneered the use of predicted RNA sequence structure maps to test and simulate many theories of evolutionary and fitness landscape predictability. They show that many sequences in an RNA sequence space fold into similar secondary structures and that these sequences are connected by single mutations forming a percolating neutral network in the sequence space (Schuster *et al.*, 1997) (Figure 4-2). RNA neutral networks have led to better understanding of how robustness, plasticity and innovation influence evolution (Ancel and Fontana, 2000). Kun et al. used a secondary structure model of real ribozyme to provide a plausible solution to the Eigen paradox (Kun, Santos and Szathmáry, 2005). The Eigen paradox involves the concept of error threshold, which suggests that there is an upper limit to the length of genetic molecule, like RNA, can reach before mutation will destroy the information encoded in the sequence. The Eigen paradox suggests that because replication is error-prone, early self-replicating systems can only maintain a relatively short RNA. In order to evolve more genetic complexity, high-fidelity replication systems have to be evolved (Eigen, 1971). But to evolve this sophisticated machinery a large genome is needed to encode its information, larger than the error threshold. Kun et al. suggests a way to overcome this error threshold by using a predicted neutral network of real ribozymes (Kun, Santos and Szathmáry, 2005). Given that many ribozymes' sequences fold into the same secondary structure, many mutations that could lead to information meltdown according to Eigen's theory, in fact have no deleterious effects on the structure and are neutral. The neutrality of the RNA sequence space provides a buffering effect to mutational meltdown and leads to a more relaxed error threshold. This could enable early ribozymes to maintain larger genome size than previously thought, from few hundreds to over 7000 bases, large enough to encode rudimentary error correction mechanisms. The work form Kun et al. highlight the influence RNA secondary structure model has on the study of key evolutionary concepts.

Experimental studies of neutral networks present several key challenges. As highlighted several times in this thesis, the enormous size of the sequence space limits the comprehensive mapping of the fitness landscape to only very small RNA. As a result, early experimental works on RNA neutral networks were done in low throughput manners. In a seminal work, Schultes and Bartel present one of the earliest experimental evidence that two functionally and structurally distinct ribozyme folds are possibly connected through a neutral network (Schultes and Bartel, 2000). In this work, an RNA sequence was designed that satisfied the secondary structure of both an HDV self-cleaving ribozyme and a ligase ribozyme. This intersecting sequence was shown to possess both ligation and self-cleaving activity. Furthermore, they showed that a long mutational path can be designed that change more than half of the bases in the two prototype ribozymes that maintain considerable activity. This mutational pathway brings the two ribozymes very close in the sequence space, whereby wildtype level activity for either function is separated by 14 mutational steps that go through the intersecting sequence that possesses both functionalities, albeit with significant reduction in either activity. This work suggests that neural networks can facilitate evolutionary innovation and adaptation by bringing the evolving population closer to regions that lie close to another neutral network that can lead to rapid acquisition of new functions or adaptive traits upon changes in the environment. Following the advancement of deep DNA sequencing technology, a larger region of the RNA fitness landscape could be studied for the first time (Pitt and Ferré-D'Amaré, 2010). This led to works that both reaffirms and denies earlier theory of RNA neutral networks. Hayden et al. works present high-throughput evidence that neutral drift can lead to accumulation of 'cryptic' genetic variation in Azoarcus group I intron ribozyme (Hayden, Ferrada and Wagner, 2011). These cryptic mutations had no effects in the current chemical environment except to make the population more diverse. However, upon changes in the environment, where the ribozymes were presented with a new substrate, the population that contains higher cryptic variation were able to adapt more rapidly to this new environment compared to population with lower level of cryptic variation. Precise, high throughput experiments enable the group to reconstruct the mutational pathway that led to this new adaptation and was able to show that the adaptation is a direct consequence of the previously cryptic mutation. This work further highlighted the important role of neutral mutation and neutral network in molecular adaptation.

Although much evidence supporting the existence and role of neutral network in RNA sequence space have been presented using high throughput experiment, many works have also pointed toward the absence of large neutral networks. Petrie and Joyce subjected two ligase ribozyme populations to continuous evolution and assessed how far genetic drift alone can be used to access distant regions of the sequence space (Petrie and Joyce, 2014). They observe that both populations remain close to the original fitness peaks with no more than 16 or 12 mutations away from the type with most mutants having less than 7 to 10 mutations. These results point toward the idea that RNA fitness landscapes are sparse with isolated fitness peaks and many deleterious mutants. This idea is further supported by two works from the Chen group. They published comprehensive sequence-activity maps of a ribozyme (Pressman *et al.*, 2019) and an aptamer (Jiménez *et al.*, 2013). These works represent the first time the entire sequence space of a reasonably sized biomolecule was mapped and remains the largest empirical fitness landscape mapped so far. Despite large-scale mapping, no evidence of large neutral networks was presented in either landscape. Many fitness peaks are well separated by large expanse of deleterious combinatorial space. Another recent study comprehensively mapped the entire combinatorial space between the two HDV and ligase ribozymes differed by 14 mutations originally identified by Schultes and Bartel and was previously mentioned (Schultes and Bartel, 2000; Bendixsen *et al.*, 2019). This work shows the direct mutational path between the two functions is inaccessible and therefore evolutionary innovation might still require a way to leap over large fitness valleys. These works show that despite suggestions by early theoretical works, RNA sequence spaces do not always contain large percolating neutral networks. If the existence of RNA neutral networks is not given, then questions remain as to where and when neutral networks can be formed and what kind of evolutionary mechanisms are employed in the absence of such networks. To even begin to answer these questions, we must first confirm whether these neutral networks can even exist in the RNA sequence space at all. Therefore, in the final part of my thesis work, I seek to present the first ever experimental evidence of an RNA neutral network.

In this chapter, I present the results of a comprehensive assay of the combinatorial space between the wildtype F1*U ligase ribozyme and a mutant with 16 substitutions identified by the computational evolutionary algorithm presented in the previous chapter. I showed that these two mutants are functionally neutral with comparable activity and are connected by an extensive network of neutral mutations. Many direct accessible pathways exist that enable the ribozyme population to traverse large regions of the fitness landscape using this neutral network. Furthermore, the high-quality dataset enabled by the sequencing assay allows me to directly analyze the complete epistatic interactions within the network. I performed a quantitative measurement of mutational interactions up to 16th order, the largest analysis of its kind. These measurements reveal that the topography of the neutral networks is largely governed by second and third order epistasis suggesting a higher level of predictability within this region of the fitness landscape. Finally, I assayed the complete single and double mutants' landscape of this new variant and revealed that this variant evolved a more mutational robust structural module. This supports an early theoretical work that evolution along a neutral network leads to increased mutational robustness (van Nimwegen, Crutchfield and Huynen, 1999). Overall, these results present the first experimental evidence supporting the existence of a large neutral network within RNA sequence space and suggest the role of epistasis in determining the accessibility and predictability of RNA fitness landscape.

Results & Discussion

Exploring the neutral network between F1*U and F1*U^m variants through combinatorial library analysis

In the last generation of the evolutionary algorithm, many mutants were identified with comparable activity to the wildtype F1*U ligase. To investigate the outcome of this evolutionary process, I picked a mutant with the highest Hamming distance (16) that retains comparable catalytic activity to the WT (RA = 0.63). I individually tested the activity of this mutant, named F1*U^m using PAGE and confirmed its relative activity to be 0.7 (Figure 4-3a). The predicted secondary structure of this mutant using ViennaRNA also reveals a very similar secondary structure to the WT ribozyme (Figure 4-3b). The mutated positions of the F1*U^m are almost exclusively within the P5 stem-loop, with one mutation at position 76 close to the ligation site. The robustness of the P5 stem region is consistent with previous reports that this region can be replaced with an aptamer sequence or completely remove and still retain catalytic activity (Lam and Joyce, 2009; Nomura and Yokobayashi, 2019).

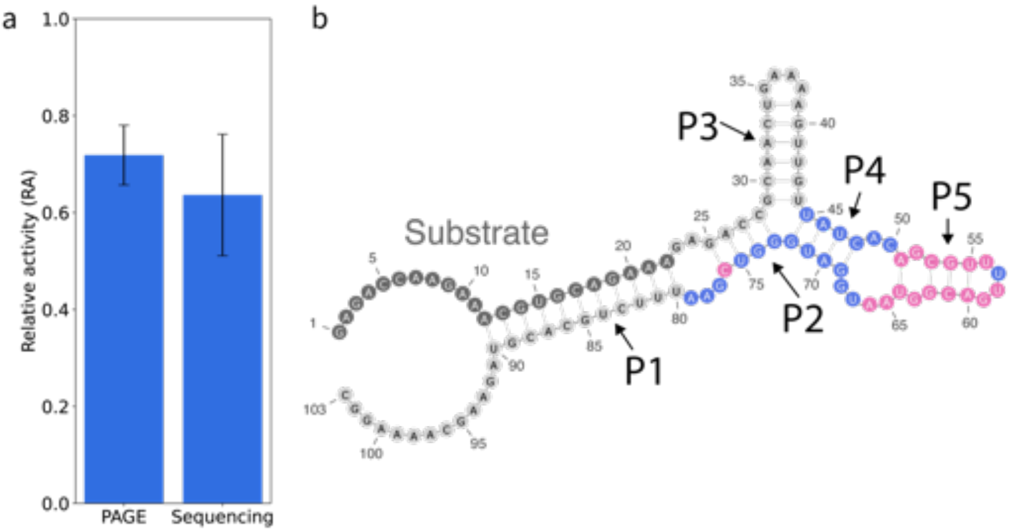


Figure 4-3: Secondary structure and activity of F1*U^m ligase ribozyme.
a) RA values of F1*U^m determined by PAGE and sequencing. PAGE experiments were performed in three replicates, and sequencing experiments were performed in duplicate. Data are presented as mean value +/− SD. b) Predicted secondary structure of F1*U^m ligase ribozyme as predicted by ViennaRNA. F1*U^m contains 16 mutations as indicated by position colored in pink. Blue color indicates the other variable positions.

Because the F1*U and F1*U^m are structurally and functionally neutral, it is possible the two variants lie within the same neutral networks and are connected by many accessible mutational pathways. In order to confirm this, I generated the complete combinatorial library between the two variants containing 65,536 (2^{16}) sequences using on-chip oligo pool synthesis. High-throughput experimental assay reveals that this combinatorial library has much higher density of variants with high relative activity compared to generation 1 which explore the local landscape (mostly single and double mutants) around the WT (Figure 4-4a). Looking at the fraction of neutral mutants using the RA threshold of 0.2, same as during the evolutionary process, also shows that the fraction is 0.6 in this library compared to 0.11

in generation 1 (Figure 4-4b). Many variants in this library across all Hamming distances also have similar RA values to the WT (Figure 4-4c). Because I have a complete set of mutational combinations, I can in theory map every possible direct mutational pathway between the two variants. However, because the library is so large, mapping every possible path (16!) would be computationally intractable. Therefore, I randomly sampled 10^6 paths. Under the strong selection, weak mutation regime, where at each generation the population can only gain one mutation which is fixed if the RA is above 0.2, almost 10% of the paths sampled would be selectively accessible (Figure 4-4b). Even after increasing the selection threshold to 0.6, I still found 39 accessible paths (Figure 4-4d). These results confirmed that the variant F1*U^m is indeed connected to the WT by an extensive neutral network allowing evolution to traverse the landscape without significant loss in activity.

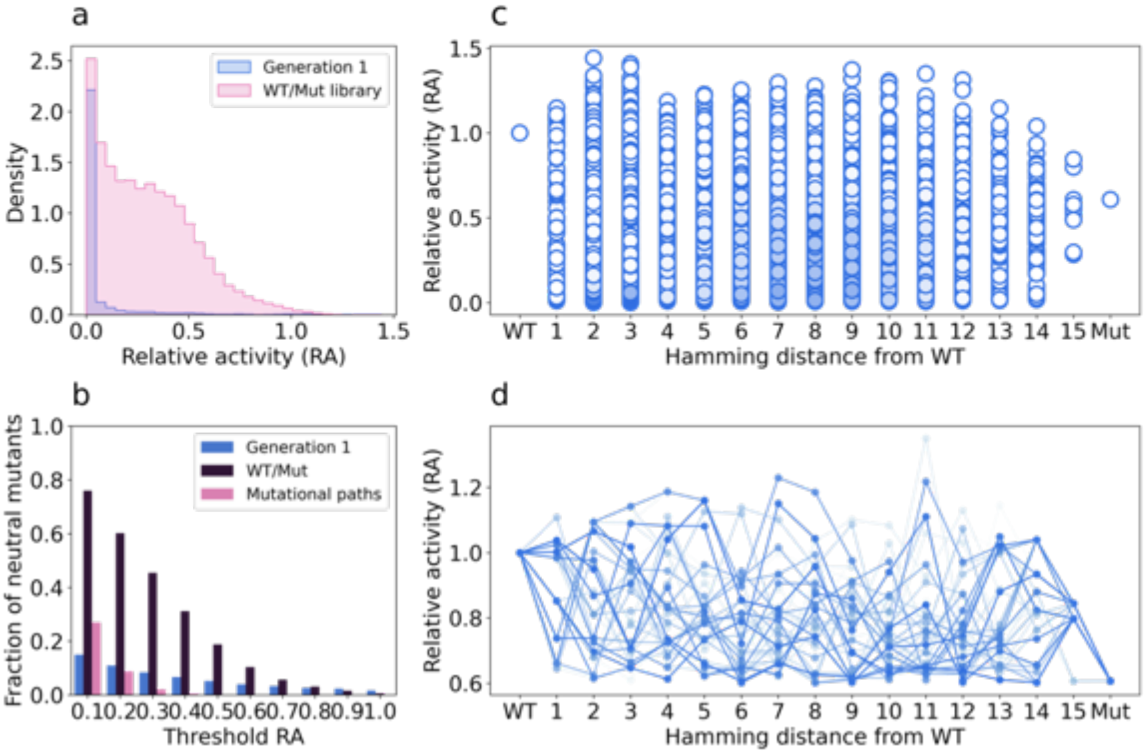


Figure 4-4: Combinatorial space between F1*U and F1*U^m contains many neutral mutants and many accessible mutational paths.

a) The distribution of relative activity (RA) of all sequences in generation 1 compared to the combinatorial library (WT/Mut) between F1*U and F1*U^m. b) Fraction of neutral mutants in the WT/Mut library and in generation 1 according to different neutrality thresholds. A total of 10^6 unique mutational paths were randomly sampled that could transform the WT sequence to F1*U^m in 16 direct mutational steps. Paths were considered neutral if none of the steps resulted in a mutant with RA lower than the neutrality threshold. c) The relative activity of all sequences in the combinatorial library plotted as a function of Hamming distance from the F1*U wildtype. d) All 39 paths identified from the 10^6 randomly sampled paths that maintain RA values above 0.6. Each node represents a single mutational step and the opacity of paths are varied for better clarity.

F1*U^{m} neutral network was predictable by multilayer perceptron

Because of the increased accessibility of this neutral network compared to the rest of the landscape, it is possible that the multilayer perceptron (MLP) model could identify the F1*U^{m} variants more efficiently by using these paths. If this is the case, then the original model should be able to predict the variants in the combinatorial library with high accuracy. To confirm this, I used the original model trained on generation 1 to 7 with no additional training to make predictions for the combinatorial library. Using the RA threshold of 0.2 as in the original evolutionary regime, the classification accuracy of the model is 0.71, much higher than the accuracy for the prediction of generation 8 (0.28 accuracy) (Figure 4-5a). The high accuracy of the model despite having only 441 of the mutants within this combinatorial library (~0.7%) presented in the training set, suggests that the model was able to generalize well to this neutral network. It is also possible that the model overfitted to these combinations of mutants as the mutants in the P5 stem regions are particularly well tolerated as evident from the first generation of assay. However, the accuracy is higher than the null accuracy which is 0.6 (the null accuracy is the accuracy if the model predicts all variants to be the majority (positive) class). The model also achieved balanced recall and precisions (F1 score = 0.77) suggesting that the models are discriminating well between neutral and deleterious combinations of mutations even within this particularly well tolerated region. The performance of the MLP also remains consistent when evaluated on variants sorted by Hamming distance to the WT (Figure 4-5b & c). The improvement in accuracy in this combinatorial library could also be attributed to the reduced diversity of this library compared to generation 8 (average Hamming distance = 8 vs. 13).

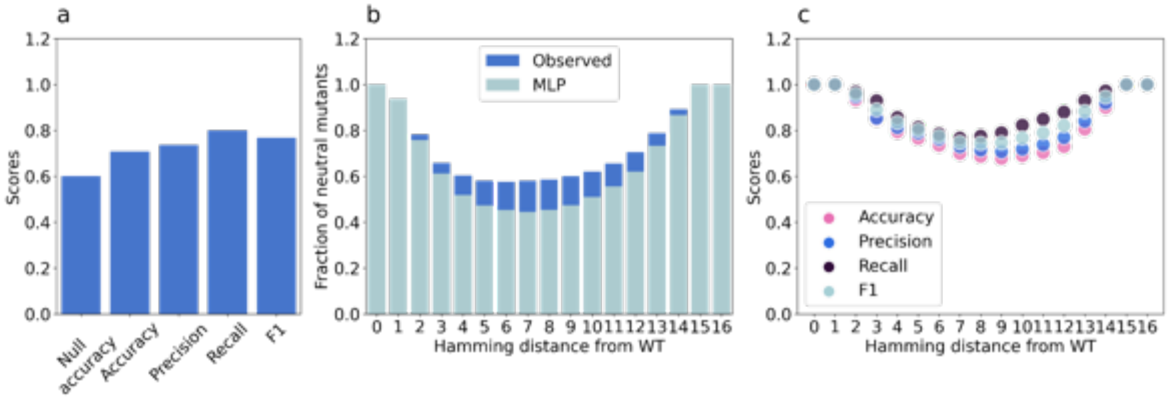


Figure 4-5: The multilayer perceptron (MLP) model trained on generation 1 to 7 was able to predict most neutral variants in the WT/Mut combinatorial library.

a) Performance metrics of MLP trained on data from generations 1–7 when used to classify the entire WT/Mut library. b) Fraction of neutral mutants at each Hamming distance in the WT/Mut library as measured by experimental assay (dark blue). The light blue area indicates the fraction of the neutral mutants that was also identified by MLP. c) Performance metrics of MLP when evaluated on sequences that are binned according to Hamming distance from the wildtype (WT).

F1*U^m neutral network is less rugged than other regions sampled by the evolutionary algorithm

Predictability of the fitness landscape is largely determined by the level epistasis presents (de Visser and Krug, 2014; Bank *et al.*, 2016). Epistasis is the nonlinear effects observed when different mutations are combined. High level of epistasis means that mutational effects change depending on the genotype background these mutations are introduced, preventing prediction of these effects without data of the complete sequence space and the landscape is rugged. Smoother landscapes have lower levels of epistasis, meaning that combinatorial mutation effects can be predicted by linear combination of individual mutations. Epistasis can be categorized in hierarchical order depending on the number of interacting mutations. Higher-order epistasis involving three or more mutations have been shown to influence RNA landscape ruggedness (Domingo, Diss and Lehner, 2018) and are important to determining evolvability and predictability of the landscape. Therefore, I decided to evaluate the amount of epistasis in the WT/Mut combinatorial library and landscape surrounding it to assess whether a neutral network is more accessible and predictable because of the lower level of epistasis or not. In theory, higher-order epistasis can be quantitatively measured if we have the complete combinatorial dataset like the one I have for the F1*U^m. However, generation 1 to 8 is a biased and incomplete sampling of the fitness landscape by the evolutionary algorithm. Therefore, direct comparison of higher-order epistasis between the two libraries is not possible. A good estimation of level of ruggedness can be achieved, however, by looking at the level of pairwise reciprocal sign epistasis (Szendro *et al.*, 2013; Song and Zhang, 2021). Pairwise reciprocal sign epistasis occurs when a double mutant and the background variant both exhibit lower or higher fitness than their two intermediate single mutants (Figure 4-6a). Reciprocal sign epistasis reverses the effects of individual mutational effects effectively forming local minima within the evolutionary path. Indeed, this form of epistasis has been shown to constrain evolution and is the direct cause of multiple peaks in the fitness landscape (Weinreich, Watson and Chao, 2005; Weinreich *et al.*, 2006; Poelwijk *et al.*, 2011). Given the effects on ruggedness, I can use the fraction of reciprocal sign epistasis to compare the accessibility and predictability between generation 1 to 8 and the combinatorial library.

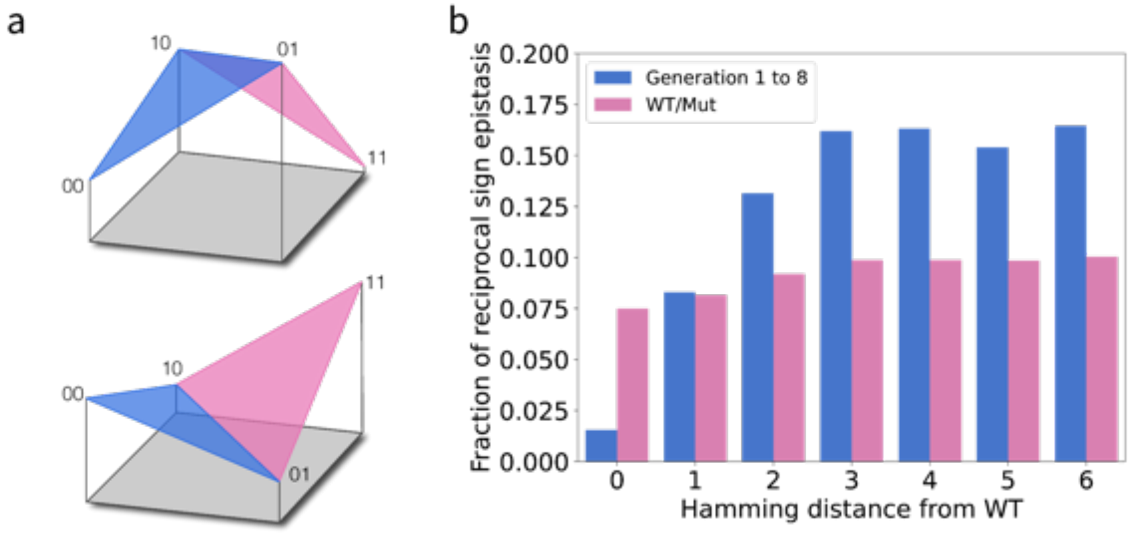
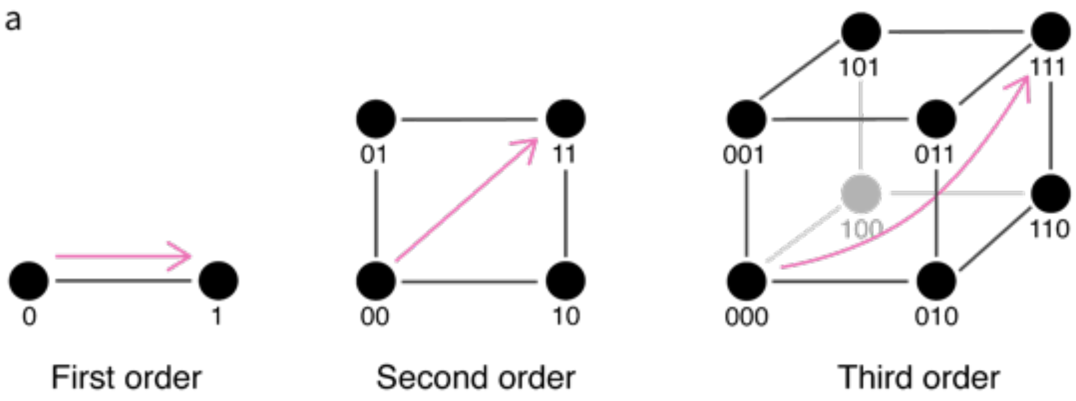


Figure 4-6: Reciprocal sign epistasis increases the ruggedness surrounding the neutral network.
a) Illustration of reciprocal sign epistasis involving two genotypes that differ by two substitutions (00 and 11). Reciprocal sign epistasis is observed when both the reference genotype (00) and the double mutant (11) has higher or lower fitness than both intermediate single mutants (01 and 10). This leads to landscape ruggedness and can restrict evolutionary paths. b) Fraction of pairwise reciprocal sign epistasis in generation 1 to 8 (blue) and WT/Mut library (pink) categorized by the Hamming distance of the reference genotype 00 from the WT. Higher fractions indicate higher ruggedness of the landscape.

To measure fraction of reciprocal sign epistasis, all 2^2 subgraphs representing a double mutant, its background sequence and two constituent single mutants are searched in both libraries. In total 214,068 subgraphs were found for generation 1-8 and all 3,932,160 possible subgraphs were obtained from the combinatorial library. The Hamming distance between each background sequence to the wildtype F1*U was used to assess the level of epistasis at each mutational order. The fraction of reciprocal sign epistasis is calculated by dividing the number of subgraphs exhibiting reciprocity to the total number of subgraphs found at each Hamming distance. The calculations showed that generation 1-8 exhibit higher fraction of reciprocal sign epistasis at higher Hamming distance compared to the combinatorial library (Figure 4-6b). This suggests that evolution within the region sampled by the evolutionary algorithm would be more constrained, while the F1*U^m neutral networks have smoother paths.

The epistatic landscape of the F1*U^m neutral network is sparse with lower-order interactions dominating its predictability

The results from pairwise epistasis measurement suggest that the neutral network should be more predictable using mostly information from lower order mutational effects. To assess this possibility further, I decided to calculate the level of epistasis at all orders within the neutral network. Mutational effects can be quantified by mapping the RA values of each mutant into the log space. In this space, a combination of mutations that increase or decrease the activity of the WT will have a log(RA) higher or lower than zero respectively. As mentioned, epistasis is observed when there is a non-linear combination of mutational effects. Therefore, the degree of pairwise or second-order epistasis can be calculated by measuring the difference between the observed log(RA) of a double mutant and the sum of the log(RA) of the two constituent single mutants. To calculate the level of third order epistasis, I can take the difference between the observed log(RA) of a triple mutant and the sum of all second order epistasis terms and the three constituent single mutant effects (which can also be considered the first order epistasis terms). This calculation is visualized in Figure 4-7a. This calculation can be extended to an arbitrary order of epistasis if there is a combinatorial complete dataset. For the F1*U^{m} neutral network, I can calculate the epistasis level up to 16th order. The calculation for an arbitrary order of epistasis can be generalized as a Walsh-Hadamard (WH) transform (Figure 4-7b). This transformation essentially performs the same calculation as outlined above except that for each order of epistasis, instead of calculating the epistatic terms relative to one background, the WH transform averages these terms across all genotype backgrounds. These background-averaged epistatic terms are more informative than using a single reference method because they show epistatic terms that have the strongest effects across all backgrounds. Another important feature of the WH transform is that it is a linear operation that converts RA values into the non-additive effects of mutations or epistatic terms. This operation can be inverse to retrieve the RA values back from the epistatic values. By controlling which epistatic terms are used for these reconstructions, I can assess the influence of epistatic terms on the predictability of the whole landscape.



b

$1/2^L$	X	Hadamard matrix	X	W	=	E	Binary string ordering
1/8	X	+1 +1 +1 +1 +1 +1 +1 +1	X	73.9	=	73.70	000
		+1 −1 +1 −1 +1 −1 +1 −1		77.5		-1.525	001
		+1 +1 −1 −1 +1 +1 −1 −1		71.2		1.275	010
		+1 −1 +1 +1 +1 −1 −1 +1		74.5		0.000	011
		+1 +1 +1 +1 −1 −1 −1 −1		73.0		0.575	100
		+1 −1 +1 −1 −1 +1 −1 +1		75.5		-0.200	101
		+1 +1 −1 −1 −1 −1 +1 +1		70.6		0.150	110
		+1 −1 −1 +1 −1 +1 +1 −1		73.4		-0.075	111

Figure 4-7: Visual diagram of the Walsh-Hadamard transform for the analysis of arbitrary order of epistasis.
a) These schematics are adapted from (Poelwijk, Krishna and Ranganathan, 2016) and shows the mutational effects relative to a single background genotype. First order epistasis (E_1) is equal to the effects of a single mutation on the reference genotype as denoted by the pink arrow. Second order epistasis (11) equal the double mutational effects minus the sum of the two single mutational effects (01 and 10) in the log space. This calculation can be extended to third and higher order terms if there is a complete combinatorial data available. b) This schematic, adapted from (Weinreich *et al.*, 2013), shows the calculation of third order epistasis using a Walsh-Hadamard transform. A Hadamard matrix can be constructed for any sequence of length, L and can be multiplied by the fitness vector, W (this example shows the thermal stability of an avian lysozyme from (Malcolm *et al.*, 1990).) which is sorted according to the binary string order of each mutant. The resulting product vector can be normalized by the number of total possible interactions, 2^L to derive the matrix of epistatic terms, E.

Applying the WH transform on the F1*U^m neutral network reveals that the strongest epistatic terms are mostly lower-order ones (Figure 4-8a). The mean squared magnitude of epistasis is higher for epistatic terms with order 1 or 2 away from the WT or F1*U^m. The magnitude of epistatic effects in the order 6 to 9 are lower suggesting that on average, higher-order epistatic effects are weaker in this neutral network. However, the deviation of magnitude within most epistatic orders is high suggesting that the neutral network is governed by a few key epistatic terms across all orders. To assess the contribution of lower order epistasis to the topography of the neutral network, I reconstructed the RA values by applying the inverse transform to the epistatic terms, with all terms higher than second order set to zero. The R^2

between the reconstructed RA and the observed RA values is 0.54 (Figure 4-8b), indicating that over half the landscape can be reconstructed from first and second order epistasis alone. Repeating this analysis by successively increasing the order of epistatic terms being included reveals the contribution of higher-order epistasis. Although in theory there could be epistasis up to the 16th order, only epistasis up to the 7th order is needed to achieve almost perfect reconstruction of the neutral network (Figure 4-8c). Measuring epistasis up to the 7th order would be intractable for most experimental set-ups. However, relatively accurate prediction of the neutral network can be achieved by analyzing only the 3rd or 4th order background-averaged epistatic terms. In fact, only the 3rd order terms were required to achieve an accuracy like that of the current MLP model ($R^2 \approx 0.72$). Earlier studies have shown that fitness landscapes can be encoded into sparse background-averaged interaction terms and can be determined from a small fraction of key mutational interactions (Poelwijk, Socolich and Ranganathan, 2019). Similarly, my results showed that the topography of this neutral network is largely encoded within lower-order background-averaged interaction terms with a few higher-order terms possessing stronger magnitude than the rest. Other studies have leveraged this sparsity alongside knowledge from the field of compressed sensing (CS) to better predict fitness values from small sample sizes (Poelwijk, Socolich and Ranganathan, 2019; Aghazadeh *et al.*, 2021). In my algorithm, successive rounds of selection, recombination, and mutation could potentially facilitate this process without the knowledge of the complete landscape. By retaining mutational combinations that were persistently neutral after rounds of diversification, MLP could learn which combinations of mutational effects remained significant in different genetic backgrounds. These and earlier results suggest that learning background average epistatic terms from a sparse dataset could be a promising direction for the field of landscape predictability.

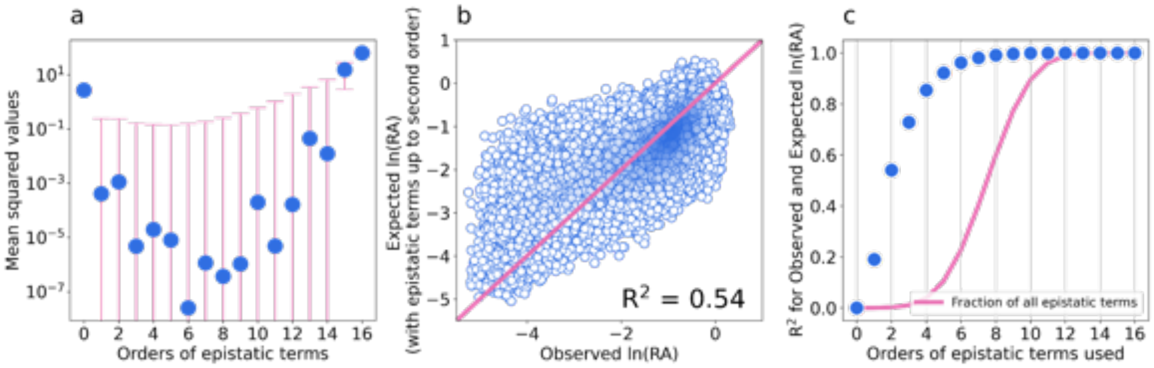


Figure 4-8: Higher order epistasis is prevalent in the WT/Mut library.
a) The magnitude of backgrounds averaged epistasis (could be positive or negative) calculated by the Walsh-Hadamard transform plot as mean squared values as categorized by their epistatic order. The error bars represent standard deviation calculated from terms at each order. b) The ln(RA) as measured by the sequencing assay plotted against the expected ln(RA) that is calculated by inverse Walsh-Hadamard transform using only the first and second-order background averaged epistatic terms while all other terms are set to zero. R^2 indicates the coefficient of determination between the two values. Pink line indicates perfect agreement. c) Coefficients of determination (R^2) values between the observed ln(RA) and the expected ln(RA). Expected ln(RA) was calculated at each step by cumulatively adding background averaged epistatic terms of successively higher order. For each step, R^2 scores were calculated using all the variants in the library. The fraction of all epistatic terms used to calculate the expected ln(RA) at each step is shown by the pink line.

Evolution along the neutral network increased mutational robustness of the F1*U^m P5 stem

Theoretical (van Nimwegen, Crutchfield and Huynen, 1999; LaBar and Adami, 2017) and experimental (Bloom, Lu, *et al.*, 2007; Bershtein, Goldin and Tawfik, 2008) evidence has suggested that evolution along a neutral network will lead to the center of the network, corresponding to regions with increased mutational robustness. I suspected that the F1*U^m variant evolved by an algorithm that are programmed to accumulate neutral mutations would also exhibit increased mutational robustness. A common measure of mutational robustness is to count the number of single neutral mutants of a genotype. With the capacity of my assay, I could go beyond this and measure mutational robustness for higher mutational order. Therefore, I generated all single, double and some triple mutants of the F1*U^m and measured their activities using the sequencing assay. Because this dataset is equivalent to generation 1 dataset for the wildtype F1*U, I can directly compare the mutational robustness between the two variants. A better way to measure the mutational robustness for a dataset with multiple mutations is to fit the data to the directional epistasis model (Wilke and Adami, 2001). In this model, the fraction of neutral mutants is fitted as a function of Hamming distance to the reference sequence (WT or F1*U^m) with the parameter α and β which are mutational robustness and directional epistasis respectively. (See Methods) The fitted α showed that both variants exhibited similar mutational robustness (Figure 4-9a). And the fitted β are greater than 1 for both, indicating an excess of negative epistasis. The excess of negative epistasis is consistent with analysis of previous experimental studies of the fitness landscape (Bendixsen, Østman and Hayden, 2017). However, the α parameters for both ligase ribozymes are lower than for the other natural ribozyme landscape reviewed by Bendixsen et al. A possible explanation was given in the study which suggests that artificial ribozymes such as the ligase ribozyme might not have been extensively evolved to increase mutational robustness which could be the predominant properties selected for by natural evolution. This is supported by the fact that similar α values were obtained in computational studies of predicted RNA secondary structure (Wilke, Lenski and Adami, 2003). The consistency between the structural model and the artificial ligase ribozyme robustness supports this theory.

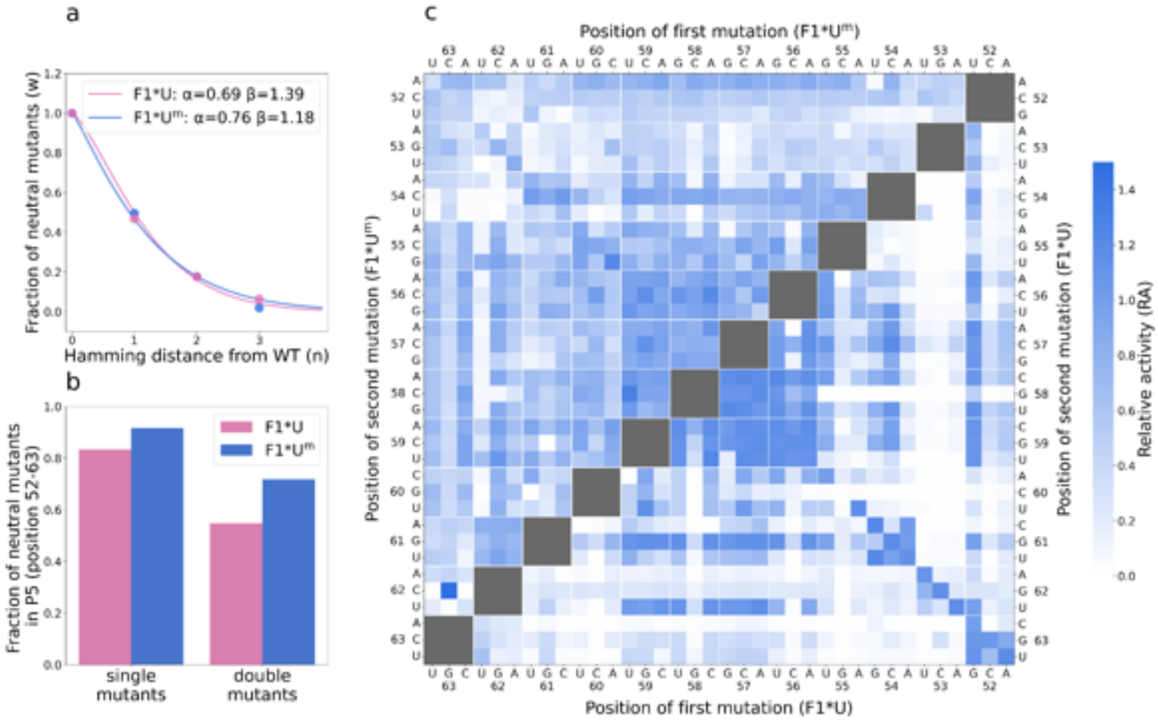


Figure 4-9: The P5 stem of F1*U^{m} has higher mutational robustness than the wildtype ribozyme.

a) The fraction neutral mutants at each Hamming distance from the wildtype calculated from assay data of the local landscape around F1*U and F1*U^{m}. The directional epistasis model ($w(n) = e^{-\alpha n^{\beta}}$) was fitted to the data using the nonlinear least square method. Small decay parameter α indicates higher mutational robustness. Parameter β indicates the strength and direction of epistasis. β greater or less than 1 indicate an excess of negative or positive epistasis respectively. When β is equal to 1, there is a balance of epistasis in both directions or there is no epistasis. b) The fraction of neutral mutants calculated from variants with mutations only in the P5 stem (position 52 to 63 in the ligated ribozyme) for both F1*U and F1*U^{m}. c) Heatmap showing the relative activity of double mutants with mutations in the P5 stem for F1*U (lower triangle) and F1*U^{m} (upper triangle).

However, plotting the double mutant heatmap of the F1*U^{m} shows that the P5 stem regions seem to be more tolerant to single and double mutation than the wildtype (Figure 4-9b & c). Indeed, plotting the fraction neutral mutants within this P5 stem loop shows the fraction is higher for F1*U^{m} than for the wildtype (Figure 4-9b). Most of the mutations acquired by the F1*U^{m} are contiguously located within the P5 region, suggesting that this variant has acquired localized mutational robustness. Looking at the base-pair probabilities of the F1*U^{m}, I can see that the secondary structure of this variant is less stable than the WT (Figure 4-10a). The destabilization of the structure could explain why the F1*U^{m} variant has a lower activity than the WT. The dot plot also shows several possible alternative base-pairing around the P4 and P5 stem (Figure 4-10b). The overall mutational robustness of the P4 and P5 stem region might also be explained by the observation that many of the alternative structures within this region could form without disrupting the P3 and P2 stems. This offers many alternative structural pathways that could maintain the overall catalytic activity of the ribozymes suggesting a possible explanation for the existence of a highly connected neutral paths between the WT and F1*U^{m}. The localized nature of mutational robustness and the high connectivity of this region suggest that the P5 stems could potentially act as a starting region for evolutionary innovation. Previous studies of Azoarcus ribozymes have demonstrated similar effects where mutations are accumulated to localized modules within the ribozyme structure (Hayden, Bendixsen and Wagner, 2015). This study suggests that the structural module can act as a mutational buffer and accumulate cryptic neutral mutations which could prepare the ribozyme for rapid adaptation or innovation upon environmental changes. These results coupled with the lack of large neutral network observe so far for ribozymes with mutation across the whole structure (Bendixsen *et al.*, 2019) suggest that expansion of small local motif rather than sudden global structural changes might be a more effective strategy for evolutionary innovation and adaptation (Popović *et al.*, 2021).

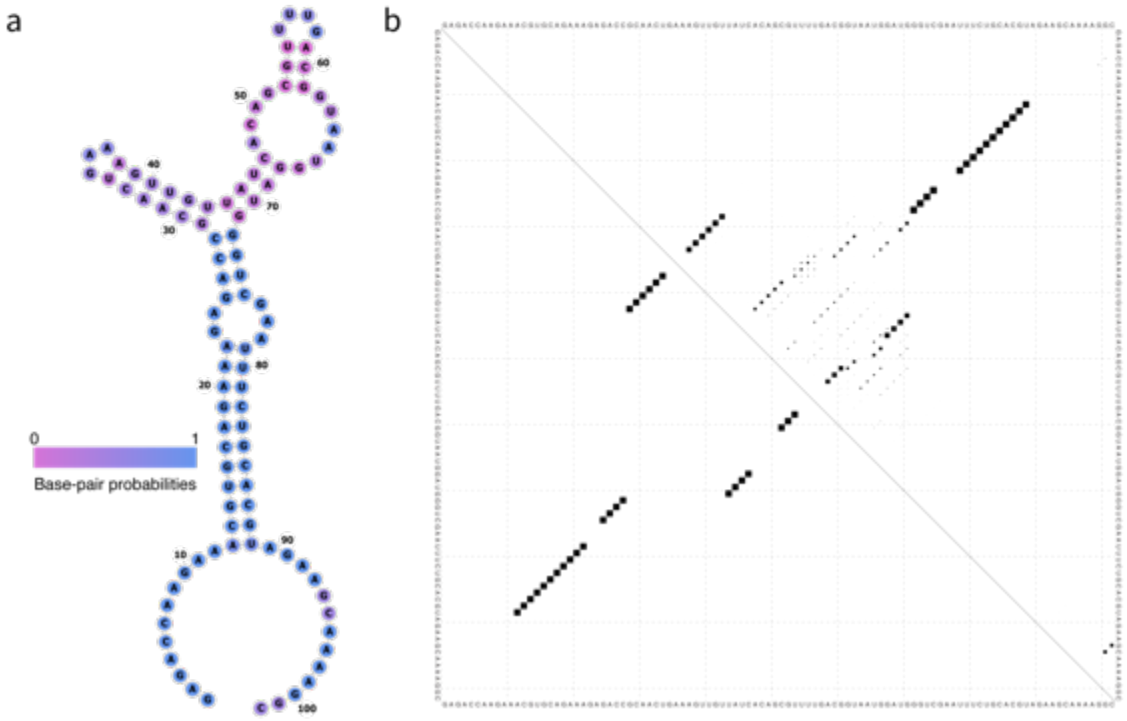


Figure 4-10: Base-pair probabilities of the F1*U^{m} ligase ribozyme.
a) Base-pair probabilities were calculated using RNAfold WebServer and is overlayed as a colormap on the minimum free energy structure. B) Dot plot of the F1*U^{m} ligase ribozyme secondary structure with the lower triangle showing the minimum free energy structure and the upper triangle showing the probability of all possible base pairs with the area of each dot proportional to the pairing probabilities.

Conclusion

The absence of large-scale neutral network in global mapping of fitness landscape combined with the evidence of neutral network in a confined structural region presented in this chapter suggests that global network of direct mutational paths might not be the primary mechanism for adaptation and innovation in RNA evolution. Fitness landscape theory, although an elegant way to visualize evolution, is too simplistic and in the era of high-throughput experimental methods might be too limited for the study of evolution. The biomolecular sequence space is too high dimensional to be thought of as a 3-dimensional mountainous landscape. Greenbury et al. recently showed that the accessibility of functional genotype increases as a function of landscape dimensionality which they defined as the proportion of mutable sites in a given length of sequence (Greenbury, Louis and Ahnert, 2022). They showed that mutational bypass, indirect alternative mutational paths, could enable evolution to navigate around fitness valley enabling fitness peaks to be reached from almost any point within the landscape. However, like earlier works, they have only investigated the predicted secondary structure map of RNA. Mutational bypass has been observed in experimental sequence-function maps (Wu *et al.*, 2016) but little works have been done trying to reconcile experimental evidence with theoretical study. Real evolutionary processes and properties such as dynamic selective environments (Hayden and Wagner, 2012; Steinberg and Ostermeier, 2016; Peri *et al.*, 2022), insertion/deletion (Martin and Ahnert, 2021) and recombination (Klug, Park and Krug, 2019) have all been theoretically implicated in increasing fitness landscape navigability. Furthermore, natural evolution of RNA sequences have been suggested to follow a compensatory evolution model (Kimura, 1985). Because maintaining the base pair structure is important for RNA function, any single substitution on one of the paired residues is thought to be very deleterious. Under this scenario, RNA evolution will primarily act on the level of secondary structure and compensatory substitutions occurred simultaneously or sequentially through a G-U intermediate (Zhang *et al.*, 2020). Subsequently, strong reciprocal sign epistasis, observed when substitutions occur independently in the paired regions, will have a weaker effect on evolutionary trajectories (Chen *et al.*, 1999). Therefore, RNA fitness landscape such as the one explored in this study could appear more neutral with more accessible paths if mutations are restricted to only compensatory substitutions. Although this scheme seems too stringent, the observation that genomic RNA sequence are much more conserved at the secondary structure level than at the primary sequence level suggests that natural evolution also follows this mechanism (Dutheil, Jossinet and Westhof, 2010). The combination of genetic operators with high throughput DNA synthesis, sequencing and experimental assay presented in this study could be used to systematically study all these mechanisms on real sequence-function maps. The approach presented here provides an important starting point towards a shift in the study of fitness landscape from simplistic low dimensional landscape to complex multidimensional landscape that better reflects the dynamic process of natural evolution. Adaptation and innovation on fitness landscape underlies so many important biological phenomena that better studies of these processes could lead to new understanding in a wide range of fields, from viral evolution (Koelle *et al.*, 2006) to drug resistance in cancer (Shaffer *et al.*, 2017).

Methods

Robustness Calculation

Equation $w(n) = e^{-\alpha n^{\beta}}$ was fitted using the nonlinear least-squares curve fitting function in the SciPy Python library. *ω(n)* is the fraction of neutral mutants (RA ≥ 0.2) at Hamming distance *n* from the reference sequence (WT or F1*U^{m}). *α* is the decay parameter, where a lower *α* indicates higher mutational robustness. *β* is the strength of directional epistasis. When *β* > 1, there is an excess of negative epistasis, when *β* < 1 there is an excess of positive epistasis. If *β* is equal to one, there is a balanced mix of positive and negative epistasis, or there is no epistasis.

Estimation of the fraction of reciprocal sign epistasis

For generations 1–8 and the WT/Mut library, I separately identified all unique pairs of mutants that differed by two substitutions. Of these, pairs of mutants in which both intermediate single mutants were present in the dataset were retained. For each pair of sequences, reciprocal sign epistasis was identified if both the RA values were higher or lower than those of the intermediate single mutants. For each set of sequences, the sequence with the lowest Hamming distance to the WT was used as the reference sequence. This was only used to measure reciprocal sign epistasis at each mutational step (Hamming distance) from the WT. The identification of reciprocal sign epistasis is not affected by the choice of the reference sequence. The fraction of reciprocal sign epistasis was calculated by dividing the number of reciprocal sign epistasis by the total number of 2^2 genotype subgraphs identified at each Hamming distance from the WT.

Analyzing epistasis within a combinatorial complete landscape

A combinatorial complete landscape consists of 2^N possible variants with *N* being the number of mutable sites. In the case of the WT/Mut library, this equaled 65,336 possible combinations for 16 mutations (2^{16}). Each variant can be represented as a 16-bit binary with 1 or 0 as each digit, indicating the presence or absence of each mutation. The ln(RA) value of each variant can be sorted according to the binary order to give vector *w*. To calculate the background-averaged epistatic terms, e_{avg} I use the equation $e_{\text{avg}} = VHw$. *H* is the Hadamard matrix which can be defined recursively as:

$$H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix} with\, H_0 = 1 \tag{1}$$

V is a weighting matrix that can be defined recursively as:

$$V_{n+1} = \begin{pmatrix} \frac{1}{2}V_n & 0 \\ 0 & -V_n \end{pmatrix} with\, V_0 = 1$$

(2)

Multiplying w by VH yields the weighted Walsh–Hadamard transform of the ln(RA) values. w can be reconstructed from e_{avg} by multiplying with the inverse of the matrix VH. ($w = (VH)^{-1}e_{avg}$) More detailed explanations of the theory can be found in (Poelwijk, Krishna and Ranganathan, 2016).

Experimental measurements of ribozyme activity

The same experimental protocols were used as described in chapter 2 Methods section. The reproducibility between sequencing repeats and between sequencing and PAGE are shown in Figure S4-1 for the WT/Mut library and in Figure S4-2 for the F1*U$^{\mathrm{m}}$ local landscape library.

Chapter 5: Conclusion and outlook

Fitness landscape was first used to describe the relationship between the phenotype and genotype and how it influences evolutionary processes and trajectories. Molecular biologists adopt this concept to study how sequence of protein, DNA and RNA influences its function. This sequence-function relationship has been used as a roadmap for a wide range of studies such as designing new and better enzymes or predicting gene expression level. Fitness landscape provide an elegant way to reconcile many properties that influences how a biomolecule will behave upon changes in its sequences. The multidimensionality of the genotype space and our own limited understanding in the navigation of this kind of space means that early description of fitness landscape is necessarily simplistic with smooth surface and ridges connecting fitness peaks. However, it has become increasingly evident that this framework is too simplistic to capture the complexity of evolutionary process especially in this new high-throughput era. Throughout this thesis, I used RNA, specifically a ligase ribozyme, to highlight how sparsity, epistasis and high dimensionality of the sequence space can lead to frustrated navigation process during artificial and natural evolution. However, this frustration only holds true if evolution can only proceed in forward direction along smooth ridges within the landscape. But this regime is simply not true in many realistic evolutionary systems. High-dimensional sequence space increases connectivity of mutational paths (Greenbury, Louis and Ahnert, 2022), changing selection and recombination enables fitness valley crossing (Steinberg and Ostermeier, 2016; Klug, Park and Krug, 2019), cryptic mutation enabling rapid adaptation (Hayden, Ferrada and Wagner, 2011), these are some of the many processes natural evolution can utilize to efficiently navigate the seemingly rugged terrain of the fitness landscape. Under this view, ruggedness might not be a natural property but stem from oversimplification of the experimental and theoretical tools used by the field to investigate and understand the fitness landscape. Therefore, I have presented in this thesis a new set of tools that are informed by experimental data and reflect the dynamic quality of real evolutionary processes.

In chapter 2, I showed that computational genetic processes specifically, selection, mutation and recombination can successfully navigate functional regions within the ligase ribozyme sequence space. The customizability of the oligo pools and the information provided by the high-throughput assay can be used to adjust the parameters of these computational processes for adaptive navigation that can be used to investigate evolution under multiple conditions. Furthermore, in chapter 3, I showed that the experimental limitation imposed by the combinatorial explosion of the sequence space can be potentially overcome by data driven computational model such as deep learning. I showed that deep neural network model trained on experimental data can accurately predict the activity of unseen ribozyme sequences. I then showed that this predictive model can be used to guide evolutionary algorithm towards new functional regions. Finally in chapter 4, I showed that the machine learning-guided computational evolution followed neutral mutational pathways towards regions of higher robustness. I experimentally confirmed that the mutational paths explored by this process form a large connecting network of neutral paths. This is the first large scale experimental evidence of such a network. Altogether, I presented a novel hybrid approach combining high-throughput experiments with data-driven computational model and experimentally showed that this approach can lead to discovery of new evidence for an important theoretical property of the fitness landscape.

I believe that this work will serve as inspiration and guide the development of similar approaches that will employ even more advanced computational model coupled with better experimental methods. Future works could develop a machine learning algorithm with better capability by incorporating known biological priors such as sparsity of epistasis (Aghazadeh *et al.*, 2021) or RNA structure (Schmidt and Smolke, 2021). The customizability and adaptability of the experimental method presented here would easily enable design of the model and experiments to be done in a synergistic manner. Training and testing dataset could be designed and collected in a way that ensure good generalization by the model for a given task. When the model is deployed, the experiments could also be adjusted to collect data that will maximize the model performance and correct the model's mistakes in an active learning approach (Borkowski *et al.*, 2020; Hie, Bryson and Berger, 2020; Greenhalgh *et al.*, 2021). These methods could lead to the understanding of many more unanswered questions about the nature of fitness landscapes. For instance, is there a relationship between the size of the sequence space and the emergence of neutral network or how duplication or recombination influence the accessibility of such network from distant regions. I envisioned that systematic evaluation of real fitness landscape using data-informed strategy will be the key to answering these and many more questions about evolution, one of, if not the most important process in the natural world.

References

Aghazadeh, A. *et al.* (2021) “Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions,” *Nature communications*, 12(1), p. 5225. doi: 10.1038/s41467-021-25371-3.

Alipanahi, B. *et al.* (2015) “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nature biotechnology*, 33(8), pp. 831–838. doi: 10.1038/nbt.3300.

Alley, E. C. *et al.* (2019) “Unified rational protein engineering with sequence-based deep representation learning,” *Nature methods*, 16(12), pp. 1315–1322. doi: 10.1038/s41592-019-0598-1.

Ancel, L. W. and Fontana, W. (2000) “Plasticity, evolvability, and modularity in RNA,” *The Journal of experimental zoology*, 288(3), pp. 242–283. doi: 10.1002/1097-010x(20001015)288:3<242::aid-jez5>3.0.co;2-o.

Andreasson, J. O. L. *et al.* (2020) “Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme,” *Nature communications*, 11(1), p. 1663. doi: 10.1038/s41467-020-15540-1.

Angenent-Mari, N. M. *et al.* (2020) “A deep learning approach to programmable RNA switches,” *Nature communications*, 11(1), p. 5057. doi: 10.1038/s41467-020-18677-1.

Angermueller, C. *et al.* (2016) “Deep learning for computational biology,” *Molecular systems biology*, 12(7), p. 878. doi: 10.15252/msb.20156651.

Bank, C. *et al.* (2016) “On the (un)predictability of a large intragenic fitness landscape,” *Proceedings of the National Academy of Sciences of the United States of America*, 113(49), pp. 14085–14090. doi: 10.1073/pnas.1612676113.

Bartel, D. P. and Szostak, J. W. (1993) “Isolation of new ribozymes from a large pool of random sequences [see comment],” *Science*, pp. 1411–1418. doi: 10.1126/science.7690155.

Beck, J. D. *et al.* (2022) “Predicting higher-order mutational effects in an RNA enzyme by machine learning of high-throughput experimental data,” *Frontiers in molecular biosciences*, 9, p. 893864. doi: 10.3389/fmolb.2022.893864.

Bedbrook, C. N. *et al.* (2019) “Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics,” *Nature methods*, 16(11), pp. 1176–1184. doi: 10.1038/s41592-019-0583-8.

Bendixsen, D. P. *et al.* (2019) “Genotype network intersections promote evolutionary innovation,” *PLoS biology*, 17(5), p. e3000300. doi: 10.1371/journal.pbio.3000300.

Bendixsen, D. P., Østman, B. and Hayden, E. J. (2017) “Negative Epistasis in Experimental RNA Fitness Landscapes,” *Journal of molecular evolution*, 85(5–6), pp. 159–168. doi: 10.1007/s00239-017-9817-5.

Bepler, T. and Berger, B. (2021) “Learning the protein language: Evolution, structure, and function,” *Cell systems*, 12(6), pp. 654-669.e3. doi: 10.1016/j.cels.2021.05.017.

Bershtein, S., Goldin, K. and Tawfik, D. S. (2008) “Intense neutral drifts yield robust and evolvable consensus proteins,” *Journal of molecular biology*, 379(5), pp. 1029–1044. doi: 10.1016/j.jmb.2008.04.024.

Biswas, S. *et al.* (2021) “Low-N protein engineering with data-efficient deep learning,” *Nature methods*, 18(4), pp. 389–396. doi: 10.1038/s41592-021-01100-y.

Blanco, C. *et al.* (2019) “Molecular Fitness Landscapes from High-Coverage Sequence Profiling,” *Annual review of biophysics*, 48, pp. 1–18. doi: 10.1146/annurev-biophys-052118-115333.

Bloom, J. D., Lu, Z., *et al.* (2007) “Evolution favors protein mutational robustness in sufficiently large populations,” *BMC biology*, 5, p. 29. doi: 10.1186/1741-7007-5-29.

Bloom, J. D., Romero, P. A., *et al.* (2007) “Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution,” *Biology direct*, 2, p. 17. doi: 10.1186/1745-6150-2-17.

Boone, K. *et al.* (2021) “Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides,” *BMC bioinformatics*, 22(1), p. 239. doi: 10.1186/s12859-021-04156-x.

Borkowski, O. *et al.* (2020) “Large scale active-learning-guided exploration for in vitro protein production optimization,” *Nature communications*, 11(1), p. 1872. doi: 10.1038/s41467-020-15798-5.

Brandes, N. *et al.* (2022) “ProteinBERT: A universal deep-learning model of protein sequence and function,” *Bioinformatics* , 38(8), pp. 2102–2110. doi: 10.1093/bioinformatics/btac020.

Carothers, J. M. *et al.* (2004) “Informational complexity and functional activity of RNA structures,” *Journal of the American Chemical Society*, 126(16), pp. 5130–5137. doi: 10.1021/ja031504a.

Cech, T. R., Zaug, A. J. and Grabowski, P. J. (1981) “In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence,” *Cell*, 27(3 Pt 2), pp. 487–496. doi: 10.1016/0092-8674(81)90390-1.

Chen, Y. *et al.* (1999) “RNA secondary structure and compensatory evolution,” *Genes & genetic systems*, 74(6), pp. 271–286. doi: 10.1266/ggs.74.271.

Chowdhury, R. *et al.* (2022) “Single-sequence protein structure prediction using a language model and deep learning,” *Nature biotechnology*, 40(11), pp. 1617–1623. doi: 10.1038/s41587-022-01432-w.

Dhamodharan, V., Kobori, S. and Yokobayashi, Y. (2017) “Large Scale Mutational and Kinetic Analysis of a Self-Hydrolyzing Deoxyribozyme,” *ACS chemical biology*. American Chemical Society, 12(12), pp. 2940–2945. doi: 10.1021/acschembio.7b00621.

Domingo, J., Baeza-Centurion, P. and Lehner, B. (2019) “The Causes and Consequences of Genetic Interactions (Epistasis),” *Annual review of genomics and human genetics*, 20, pp. 433–460. doi: 10.1146/annurev-genom-083118-014857.

Domingo, J., Diss, G. and Lehner, B. (2018) “Pairwise and higher-order genetic interactions during the evolution of a tRNA,” *Nature*, 558(7708), pp. 117–121. doi: 10.1038/s41586-018-0170-7.

Drummond, D. A. *et al.* (2005) “On the conservative nature of intragenic recombination,” *Proceedings of the National Academy of Sciences of the United States of America*, 102(15), pp. 5380–5385. doi: 10.1073/pnas.0500729102.

Dutheil, J. Y., Jossinet, F. and Westhof, E. (2010) “Base pairing constraints drive structural epistasis in ribosomal RNA sequences,” *Molecular biology and evolution*, 27(8), pp. 1868–1876. doi: 10.1093/molbev/msq069.

Dykstra, P. B., Kaplan, M. and Smolke, C. D. (2022) “Engineering synthetic RNA devices for cell control,” *Nature reviews. Genetics*, 23(4), pp. 215–228. doi: 10.1038/s41576-021-00436-7.

Eigen, M. (1971) “Selforganization of matter and the evolution of biological macromolecules,” *Die Naturwissenschaften*, 58(10), pp. 465–523. doi: 10.1007/BF00623322.

Esteva, A. *et al.* (2017) “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, 542(7639), pp. 115–118. doi: 10.1038/nature21056.

Famulok, M., Hartig, J. S. and Mayer, G. (2007) “Functional aptamers and aptazymes in biotechnology, diagnostics, and therapy,” *Chemical reviews*, 107(9), pp. 3715–3743. doi: 10.1021/cr0306743.

Ferruz, N. and Höcker, B. (2022) “Controllable protein design with language models,” *Nature Machine Intelligence*. Nature Publishing Group, 4(6), pp. 521–532. doi: 10.1038/s42256-022-00499-z.

Ferruz, N., Schmidt, S. and Höcker, B. (2022) “ProtGPT2 is a deep unsupervised language model for protein design,” *Nature communications*, 13(1), p. 4348. doi: 10.1038/s41467-022-32007-7.

Gelman, S. *et al.* (2021) “Neural networks to learn protein sequence-function relationships from deep mutational scanning data,” *Proceedings of the National Academy of Sciences of the United States of America*, 118(48). doi: 10.1073/pnas.2104878118.

Gilbert, W. (1986) *Origin of life: The RNA world*, *Nature Publishing Group UK*. doi: 10.1038/319618a0.

Gonzalez Somermeyer, L. *et al.* (2022) “Heterogeneity of the GFP fitness landscape and data-driven protein design,” *eLife*, 11. doi: 10.7554/eLife.75842.

Greenbury, S. F., Louis, A. A. and Ahnert, S. E. (2022) “The structure of genotype-phenotype maps makes fitness landscapes navigable,” *Nature ecology & evolution*. doi: 10.1038/s41559-022-01867-z.

Greener, J. G. *et al.* (2022) “A guide to machine learning for biologists,” *Nature reviews. Molecular cell biology*, 23(1), pp. 40–55. doi: 10.1038/s41580-021-00407-0.

Greenhalgh, J. C. *et al.* (2021) “Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production," *Nature communications*, 12(1), p. 5825. doi: 10.1038/s41467-021-25831-w.

Groher, A.-C. *et al.* (2019) "Tuning the Performance of Synthetic Riboswitches using Machine Learning," *ACS synthetic biology*, 8(1), pp. 34–44. doi: 10.1021/acssynbio.8b00207.

Hayden, E. J., Bendixsen, D. P. and Wagner, A. (2015) "Intramolecular phenotypic capacitance in a modular RNA molecule," *Proceedings of the National Academy of Sciences of the United States of America*, 112(40), pp. 12444–12449. doi: 10.1073/pnas.1420902112.

Hayden, E. J., Ferrada, E. and Wagner, A. (2011) "Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme," *Nature*, 474(7349), pp. 92–95. doi: 10.1038/nature10083.

Hayden, E. J. and Wagner, A. (2012) "Environmental change exposes beneficial epistatic interactions in a catalytic RNA," *Proceedings. Biological sciences / The Royal Society*, 279(1742), pp. 3418–3425. doi: 10.1098/rspb.2012.0956.

Hie, B., Bryson, B. D. and Berger, B. (2020) "Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design," *Cell systems*, 11(5), pp. 461-477.e9. doi: 10.1016/j.cels.2020.09.007.

Hofacker, I. L. *et al.* (1994) "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie / Chemical Monthly*, 125(2), pp. 167–188. doi: 10.1007/BF00818163.

Hsu, C. *et al.* (2022) "Learning protein fitness models from evolutionary and assay-labeled data," *Nature biotechnology*, 40(7), pp. 1114–1122. doi: 10.1038/s41587-021-01146-5.

Hu, Y.-J. (2002) "Prediction of consensus structural motifs in a family of coregulated RNA sequences," *Nucleic acids research*, 30(17), pp. 3886–3893. doi: 10.1093/nar/gkf485.

Hu, Y.-J. (2003) "GPRM: A genetic programming approach to finding common RNA secondary structure elements," *Nucleic acids research*, 31(13), pp. 3446–3449. doi: 10.1093/nar/gkg521.

Jaganathan, K. *et al.* (2019) "Predicting Splicing from Primary Sequence with Deep Learning," *Cell*, 176(3), pp. 535-548.e24. doi: 10.1016/j.cell.2018.12.015.

Jiménez, J. I. *et al.* (2013) "Comprehensive experimental fitness landscape and evolutionary network for small RNA," *Proceedings of the National Academy of Sciences*, 110(37), pp. 14984–14989. doi: 10.1073/pnas.1307604110.

Johnston, I. G. *et al.* (2022) "Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution," *Proceedings of the National Academy of Sciences*, 119(11), p. e2113883119. doi: 10.1073/pnas.2113883119.

Jumper, J. *et al.* (2021) "Highly accurate protein structure prediction with AlphaFold," *Nature*, 596(7873), pp. 583–589. doi: 10.1038/s41586-021-03819-2.

Kauffman, S. and Levin, S. (1987) "Towards a general theory of adaptive walks on rugged landscapes," *Journal of theoretical biology*, 128(1), pp. 11–45. doi: 10.1016/S0022-5193(87)80029-2.

Kimura, M. (1968) "Evolutionary rate at the molecular level," *Nature*, 217(5129), pp. 624–626. doi: 10.1038/217624a0.

Kimura, M. (1985) "The role of compensatory neutral mutations in molecular evolution," *Journal of genetics*, 64(1), pp. 7–19. doi: 10.1007/BF02923549.

Klug, A., Park, S.-C. and Krug, J. (2019) "Recombination and mutational robustness in neutral fitness landscapes," *PLoS computational biology*, 15(8), p. e1006884. doi: 10.1371/journal.pcbi.1006884.

Kobori, S. *et al.* (2015) "High-throughput assay and engineering of self-cleaving ribozymes by sequencing," *Nucleic acids research*, 43(13), p. e85. doi: 10.1093/nar/gkv265.

Kobori, S., Takahashi, K. and Yokobayashi, Y. (2017) "Deep Sequencing Analysis of Aptazyme Variants Based on a Pistol Ribozyme," *ACS synthetic biology*. American Chemical Society, 6(7), pp. 1283–1288. doi: 10.1021/acssynbio.7b00057.

Kobori, S. and Yokobayashi, Y. (2016) "High-Throughput Mutational Analysis of a Twister Ribozyme," *Angewandte Chemie* , 55(35), pp. 10354–10357. doi: 10.1002/anie.201605470.

Kobori, S. and Yokobayashi, Y. (2018) "Analyzing and Tuning Ribozyme Activity by Deep Sequencing To Modulate Gene Expression Level in Mammalian Cells," *ACS synthetic biology*. American Chemical Society, 7(2), pp. 371–376. doi: 10.1021/acssynbio.7b00367.

Koelle, K. *et al.* (2006) "Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans," *Science*, 314(5807), pp. 1898–1903. doi: 10.1126/science.1132745.

Kruger, K. *et al.* (1982) "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena," *Cell*, 31(1), pp. 147–157. doi: 10.1016/0092-8674(82)90414-7.

Kun, A., Santos, M. and Szathmáry, E. (2005) "Real ribozymes suggest a relaxed error threshold," *Nature genetics*, 37(9), pp. 1008–1011. doi: 10.1038/ng1621.

LaBar, T. and Adami, C. (2017) "Evolution of drift robustness in small populations," *Nature communications*, 8(1), p. 1012. doi: 10.1038/s41467-017-01003-7.

Lam, B. J. and Joyce, G. F. (2009) "Autocatalytic aptazymes enable ligand-dependent exponential amplification of RNA," *Nature biotechnology*, 27(3), pp. 288–292. doi: 10.1038/nbt.1528.

Lauring, A. S., Frydman, J. and Andino, R. (2013) "The role of mutational robustness in RNA virus evolution," *Nature reviews. Microbiology*, 11(5), pp. 327–336. doi: 10.1038/nrmicro3003.

Li, C. *et al.* (2016) "The fitness landscape of a tRNA gene," *Science*, 352(6287), pp. 837–

840. doi: 10.1126/science.aae0568.

Li, C. and Zhang, J. (2018) “Multi-environment fitness landscapes of a tRNA gene,” *Nature Ecology & Evolution*. Nature Publishing Group, 2(6), pp. 1025–1032. doi: 10.1038/s41559-018-0549-8.

Li, F. *et al.* (2022) “Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction,” *Nature Catalysis*. Nature Publishing Group, 5(8), pp. 662–672. doi: 10.1038/s41929-022-00798-z.

Luo, Y. *et al.* (2021) “ECNet is an evolutionary context-integrated deep learning framework for protein engineering,” *Nature communications*, 12(1), p. 5743. doi: 10.1038/s41467-021-25976-8.

Madani, A. *et al.* (2023) “Large language models generate functional protein sequences across diverse families,” *Nature biotechnology*. doi: 10.1038/s41587-022-01618-2.

Malcolm, B. A. *et al.* (1990) “Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing,” *Nature*. Nature Publishing Group, 345(6270), pp. 86–89. doi: 10.1038/345086a0.

Martin, N. S. and Ahnert, S. E. (2021) “Insertions and deletions in the RNA sequence-structure map,” *Journal of the Royal Society, Interface / the Royal Society*, 18(183), p. 20210380. doi: 10.1098/rsif.2021.0380.

Maynard Smith, J. (1970) “Natural Selection and the Concept of a Protein Space,” *Nature*. Nature Publishing Group, 225(5232), pp. 563–564. doi: 10.1038/225563a0.

Meier, J. *et al.* (2021) “Language models enable zero-shot prediction of the effects of mutations on protein function,” *Advances in neural information processing systems*. proceedings.neurips.cc. Available at: https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

Michal, S. *et al.* (2007) “Finding a common motif of RNA sequences using genetic programming: the GeRNAMo system,” *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 4(4), pp. 596–610. doi: 10.1109/tcbb.2007.1045.

Miikkulainen, R. and Forrest, S. (2021) “A biological perspective on evolutionary computation,” *Nature Machine Intelligence*. Nature Publishing Group, 3(1), pp. 9–15. doi: 10.1038/s42256-020-00278-8.

van Nimwegen, E., Crutchfield, J. P. and Huynen, M. (1999) “Neutral evolution of mutational robustness,” *Proceedings of the National Academy of Sciences of the United States of America*, 96(17), pp. 9716–9720. doi: 10.1073/pnas.96.17.9716.

Nomura, Y. and Yokobayashi, Y. (2019) “Systematic minimization of RNA ligase ribozyme through large-scale design-synthesis-sequence cycles,” *Nucleic acids research*. Oxford Academic, 47(17), pp. 8950–8960. doi: 10.1093/nar/gkz729.

Payne, J. L. and Wagner, A. (2019) “The causes of evolvability and their evolution,” *Nature reviews. Genetics*, 20(1), pp. 24–38. doi: 10.1038/s41576-018-0069-z.

Pearce, R., Omenn, G. S. and Zhang, Y. (2022) "De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning," *bioRxiv*. doi: 10.1101/2022.05.15.491755.

Peri, G. *et al.* (2022) "Dynamic RNA Fitness Landscapes of a Group I Ribozyme during Changes to the Experimental Environment," *Molecular biology and evolution*, 39(3). doi: 10.1093/molbev/msab373.

Petrie, K. L. and Joyce, G. F. (2014) "Limits of neutral drift: lessons from the in vitro evolution of two ribozymes," *Journal of molecular evolution*, 79(3–4), pp. 75–90. doi: 10.1007/s00239-014-9642-z.

Pitt, J. N. and Ferré-D'Amaré, A. R. (2010) "Rapid construction of empirical RNA fitness landscapes," *Science*, 330(6002), pp. 376–379. doi: 10.1126/science.1192001.

Poelwijk, F. J. *et al.* (2011) "Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes," *Journal of theoretical biology*, 272(1), pp. 141–144. doi: 10.1016/j.jtbi.2010.12.015.

Poelwijk, F. J., Krishna, V. and Ranganathan, R. (2016) "The Context-Dependence of Mutations: A Linkage of Formalisms," *PLoS computational biology*, 12(6), p. e1004771. doi: 10.1371/journal.pcbi.1004771.

Poelwijk, F. J., Socolich, M. and Ranganathan, R. (2019) "Learning the pattern of epistasis linking genotype and phenotype in a protein," *Nature communications*, 10(1), p. 4213. doi: 10.1038/s41467-019-12130-8.

Popović, M. *et al.* (2021) "In vitro selections with RNAs of variable length converge on a robust catalytic core," *Nucleic acids research*, 49(2), pp. 674–683. doi: 10.1093/nar/gkaa1238.

Pressman, A. D. *et al.* (2019) "Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA," *Journal of the American Chemical Society*. American Chemical Society, 141(15), pp. 6213–6223. doi: 10.1021/jacs.8b13298.

Puchta, O. *et al.* (2016) "Network of epistatic interactions within a yeast snoRNA," *Science*, 352(6287), pp. 840–844. doi: 10.1126/science.aaf0965.

Rao, R. *et al.* (2020) "Transformer protein language models are unsupervised structure learners," *Biorxiv*. biorxiv.org. Available at: https://www.biorxiv.org/content/10.1101/2020.12.15.422761.abstract.

Rives, A. *et al.* (2021) "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences of the United States of America*, 118(15). doi: 10.1073/pnas.2016239118.

Robertson, M. P. and Joyce, G. F. (2014) "Highly efficient self-replicating RNA enzymes," *Chemistry & biology*, 21(2), pp. 238–245. doi: 10.1016/j.chembiol.2013.12.004.

Robertson, M. P. and Scott, W. G. (2007) "The structural basis of ribozyme-catalyzed RNA assembly," *Science*, 315(5818), pp. 1549–1553. doi: 10.1126/science.1136231.

Rogers, J. and Joyce, G. F. (2001) “The effect of cytidine on the structure and function of an RNA ligase ribozyme,” *RNA* , 7(3), pp. 395–404. doi: 10.1017/s135583820100228x.

Romero, P. A., Krause, A. and Arnold, F. H. (2013) “Navigating the protein fitness landscape with Gaussian processes,” *Proceedings of the National Academy of Sciences of the United States of America*, 110(3), pp. E193-201. doi: 10.1073/pnas.1215251110.

Sample, P. J. *et al.* (2019) “Human 5’ UTR design and variant effect prediction from a massively parallel translation assay,” *Nature biotechnology*, 37(7), pp. 803–809. doi: 10.1038/s41587-019-0164-5.

Sánchez-Romero, M. A. and Casadesús, J. (2014) “Contribution of phenotypic heterogeneity to adaptive antibiotic resistance,” *Proceedings of the National Academy of Sciences of the United States of America*, 111(1), pp. 355–360. doi: 10.1073/pnas.1316084111.

Schmidt, C. M. and Smolke, C. D. (2021) “A convolutional neural network for the prediction and forward design of ribozyme-based gene-control elements,” *eLife*, 10. doi: 10.7554/eLife.59697.

Schultes, E. A. and Bartel, D. P. (2000) “One sequence, two ribozymes: implications for the emergence of new ribozyme folds,” *Science*, 289(5478), pp. 448–452. doi: 10.1126/science.289.5478.448.

Schuster, P. *et al.* (1997) “From sequences to shapes and back: a case study in RNA secondary structures,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*. Royal Society, 255(1344), pp. 279–284. doi: 10.1098/rspb.1994.0040.

Shaffer, S. M. *et al.* (2017) “Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance,” *Nature*, 546(7658), pp. 431–435. doi: 10.1038/nature22794.

Shapiro, B. A. (1988) “An algorithm for comparing multiple RNA secondary structures,” *Computer applications in the biosciences: CABIOS*, 4(3), pp. 387–393. doi: 10.1093/bioinformatics/4.3.387.

Shin, J.-E. *et al.* (2021) “Protein design and variant prediction using autoregressive generative models,” *Nature communications*, 12(1), p. 2403. doi: 10.1038/s41467-021-22732-w.

Song, H. *et al.* (2021) “Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning,” *Cell systems*, 12(1), pp. 92-101.e8. doi: 10.1016/j.cels.2020.10.007.

Song, S. and Zhang, J. (2021) “Unbiased inference of the fitness landscape ruggedness from imprecise fitness estimates,” *Evolution; international journal of organic evolution*. academic.oup.com, 75(11), pp. 2658–2671. doi: 10.1111/evo.14363.

Steinberg, B. and Ostermeier, M. (2016) “Environmental changes bridge evolutionary valleys,” *Science advances*, 2(1), p. e1500921. doi: 10.1126/sciadv.1500921.

Szendro, I. G. *et al.* (2013) “Quantitative analyses of empirical fitness landscapes,” *Journal of statistical mechanics* . IOP Publishing, 2013(01), p. P01005. doi: 10.1088/1742-

5468/2013/01/P01005.

Vaishnav, E. D. *et al.* (2022) “The evolution, evolvability and engineering of gene regulatory DNA,” *Nature*, 603(7901), pp. 455–463. doi: 10.1038/s41586-022-04506-6.

Valeri, J. A. *et al.* (2020) “Sequence-to-function deep learning frameworks for engineered riboregulators,” *Nature communications*, 11(1), p. 5058. doi: 10.1038/s41467-020-18676-2.

de Visser, J. A. G. M. and Krug, J. (2014) “Empirical fitness landscapes and the predictability of evolution,” *Nature reviews. Genetics*, 15(7), pp. 480–490. doi: 10.1038/nrg3744.

Wachowius, F., Attwater, J. and Holliger, P. (2017) “Nucleic acids: function and potential for abiogenesis,” *Quarterly reviews of biophysics*, 50, p. e4. doi: 10.1017/S0033583517000038.

Wagner, A. (2008) “Neutralism and selectionism: a network-based reconciliation,” *Nature reviews. Genetics*, 9(12), pp. 965–974. doi: 10.1038/nrg2473.

Weinreich, D. M. *et al.* (2006) “Darwinian evolution can follow only very few mutational paths to fitter proteins,” *Science*, 312(5770), pp. 111–114. doi: 10.1126/science.1123539.

Weinreich, D. M. *et al.* (2013) “Should evolutionary geneticists worry about higher-order epistasis?,” *Current opinion in genetics & development*, 23(6), pp. 700–707. doi: 10.1016/j.gde.2013.10.007.

Weinreich, D. M., Watson, R. A. and Chao, L. (2005) “Perspective: Sign epistasis and genetic constraint on evolutionary trajectories,” *Evolution; international journal of organic evolution*, 59(6), pp. 1165–1174. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16050094.

Whitacre, J. M. (2010) “Degeneracy: a link between evolvability, robustness and complexity in biological systems,” *Theoretical biology & medical modelling*, 7, p. 6. doi: 10.1186/1742-4682-7-6.

Wilke, C. O. and Adami, C. (2001) “Interaction between directional epistasis and average mutational effects,” *Proceedings. Biological sciences / The Royal Society*, 268(1475), pp. 1469–1474. doi: 10.1098/rspb.2001.1690.

Wilke, C. O., Lenski, R. E. and Adami, C. (2003) “Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding,” *BMC evolutionary biology*, 3, p. 3. doi: 10.1186/1471-2148-3-3.

Wright, S. (1932) “The roles of mutation, inbreeding, crossbreeding, and selection in evolution,” *Proceedings of the XI International Congress of Genetics*, 8, pp. 209–222. Available at: http://www.esp.org/books/6th-congress/facsimile/contents/6th-cong-p356-wright.pdf.

Wu, N. C. *et al.* (2016) “Adaptation in protein fitness landscapes is facilitated by indirect paths,” *eLife*, 5. doi: 10.7554/eLife.16965.

Wu, Z. *et al.* (2019) “Machine learning-assisted directed protein evolution with combinatorial libraries," *Proceedings of the National Academy of Sciences of the United States of America*, 116(18), pp. 8852–8858. doi: 10.1073/pnas.1901979116.

Yokobayashi, Y. (2019) "Applications of high-throughput sequencing to analyze and engineer ribozymes," *Methods* , 161, pp. 41–45. doi: 10.1016/j.ymeth.2019.02.001.

Yoshida, M. *et al.* (2018) "Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides," *Chem*, 4(3), pp. 533–543. doi: 10.1016/j.chempr.2018.01.005.

Zhang, X. *et al.* (2020) "Adenine·cytosine substitutions are an alternative pathway of compensatory mutation in angiosperm ITS2," *RNA* , 26(2), pp. 209–217. doi: 10.1261/rna.072660.119.

Zhou, J. *et al.* (2022) "Higher-order epistasis and phenotypic prediction," *Proceedings of the National Academy of Sciences of the United States of America*, 119(39), p. e2204233119. doi: 10.1073/pnas.2204233119.

Zuker, M. and Stiegler, P. (1981) "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic acids research*, 9(1), pp. 133–148. doi: 10.1093/nar/9.1.133.

Appendix

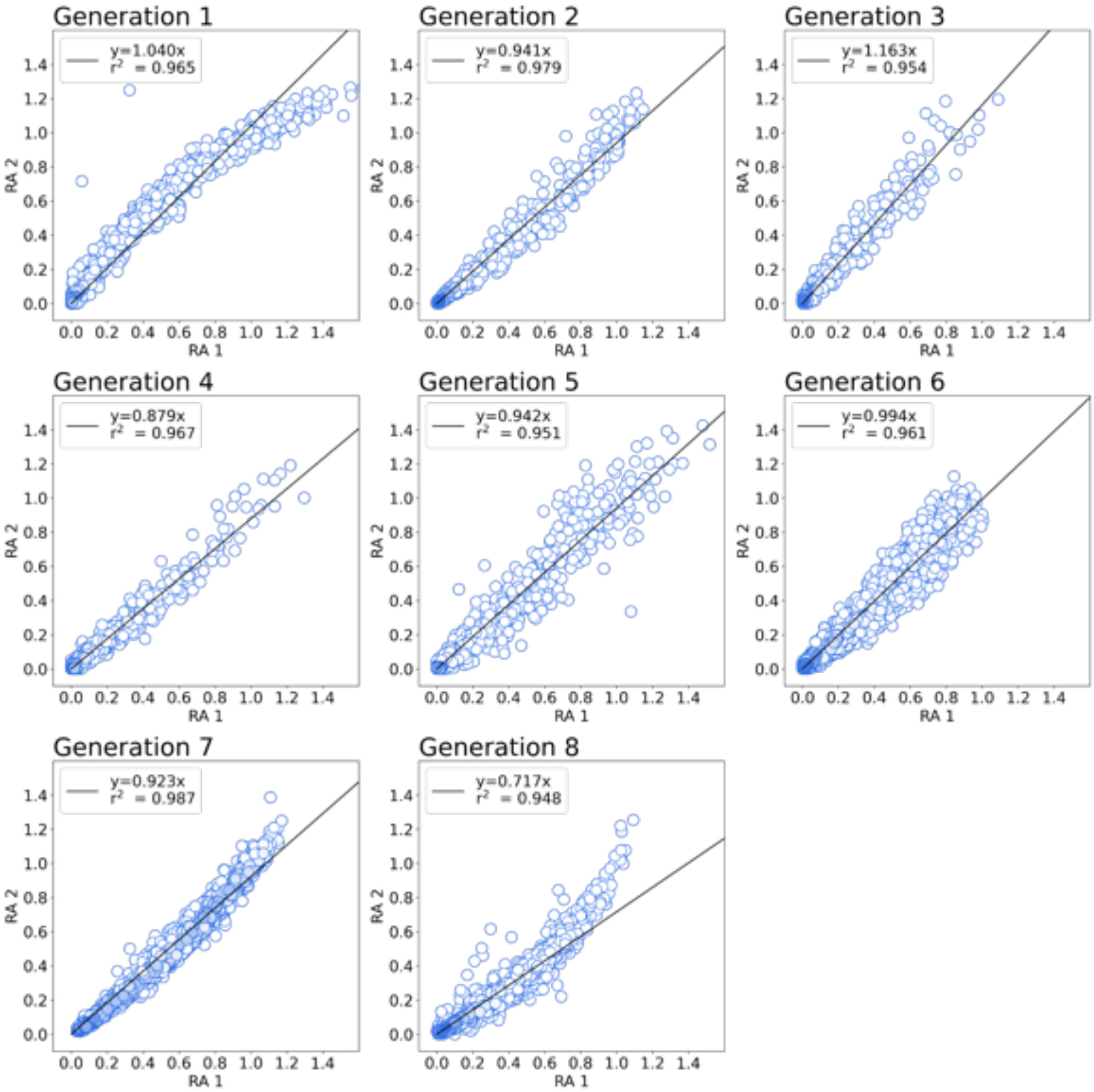


Figure S2-1: Reproducibility of sequencing assay for generation 1 to 8. Two independent experimental assays of ribozyme activities were performed for each generation. Relative activity (RA) values calculated in each repeat are compared by the square of Pearson's correlation coefficient (r^2) values.

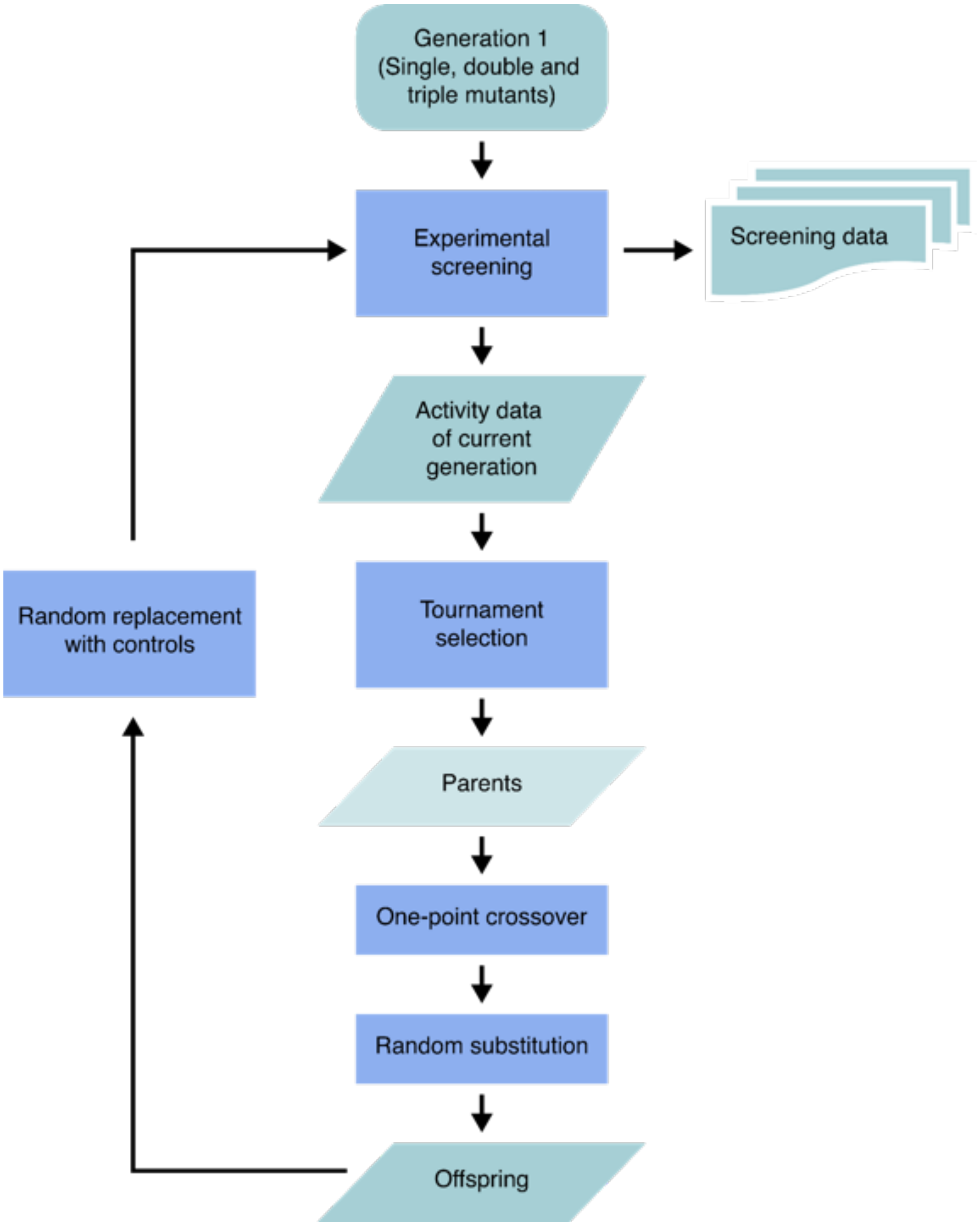


Figure S2-2: Overview of the algorithm used to design generation 2 to 5. In this algorithm, selected parents undergo both crossover and mutation to produce the offspring.

Name	**Sequence (5’ to 3’)**
Ligase-lib-f	CCTAATACGACTCACTATAGAGACCGCA
Ligase-lib-r	GCCTTTTGCTTCTACGTGCAGAA
F1*subA	GAGACCAAGAAACGUGCAGAAA
R1-bc5-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCCTCGCCTTTTGCTTCTACGTGCA
R1-bc6-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAACGGGCCTTTTGCTTCTACGTGCA
R1-bc7-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACGCAGGCCTTTTGCTTCTACGTGCA
R1-bc8-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTATTATGCCTTTTGCTTCTACGTGCA
R1-bc9-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTTACTTTGCCTTTTGCTTCTACGTGCA
R1-bc10-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGGATAGCCTTTTGCTTCTACGTGCA
R1-bc11-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGAAGCCTTTTGCTTCTACGTGCA
R1-bc12-F1-lig	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATATCCGCCTTTTGCTTCTACGTGCA
R1-f2	ACACTCTTTCCCTACACGACGCTCT
R2-F1-lig	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAGACCGCAACTGAAAAGTTG
TruSeq-i7-UDI0004	CAAGCAGAAGACGGCATACGAGATTTGGACTTGTGACTGGAGTTCAGACGTGTG
TruSeq-i5-UDI0004	AATGATACGGCGACCACCGAGATCTACACTATGAGTAACACTCTTTCCCTACACGACGC
TruSeq-i7-UDI0005	CAAGCAGAAGACGGCATACGAGATCAGTGGATGTGACTGGAGTTCAGACGTGTG
TruSeq-i5-UDI0005	AATGATACGGCGACCACCGAGATCTACACAGGTGCGTACACTCTTTCCCTACACGACGC
TruSeq-i7-UDI0006	CAAGCAGAAGACGGCATACGAGATTGACAAGCGTGACTGGAGTTCAGACGTGTG
TruSeq-i5-UDI0006	AATGATACGGCGACCACCGAGATCTACACGAACATACACACTCTTTCCCTACACGACGC

Table S2-1: List of oligonucleotides and primers used in this study.
This table lists the name and sequences of oligonucleotide primers used for experimental procedure in this study including library construction, reverse transcription and sequencing. The role of each primer is described in the Methods section of their corresponding chapter.

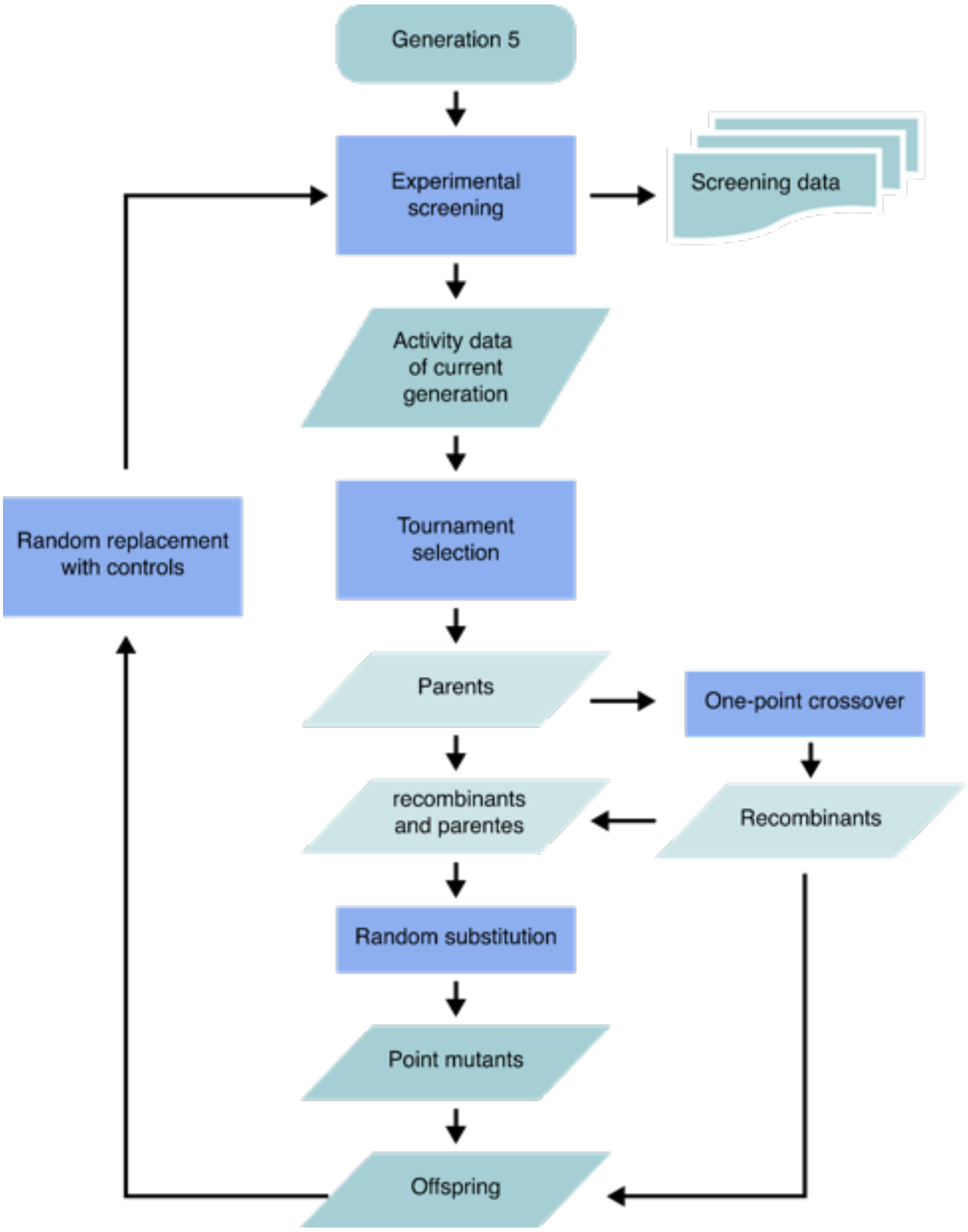


Figure S2-3: Overview of the algorithm used to design generation 6 and 7. In this algorithm, the offspring population is made up of purely recombined variants and recombinants or selected parents that have also undergone point mutation. In generation 7, the offspring population consists only of mutants that are predicted to be functional by a multilayer perceptron (MLP) model that was trained on data from generation 1 to 6.

Name	Sequence (5’ to 3’)
Template	CCTAATACGACTCACTATAGAGACCGCAACTGAAATAG TTG [catalytic core] TTTCTGCACGTAGAAGCAAAAGGC
M01	GATCACTTGTCGTAAGACACTGTGGATGGGTCGAA
M02	GATCACTTGTCGTCAGACACATTGGATGGGTTGAA
M03	GATCACTTGTCGTGAGGGACTTTGGATGGGTTGAA
M04	TATCACTTGACGTCAGACACTTTGGATGGGTCGAA
M05	GATCACTTGTCGCACGACACTCTGGATGGGTTGAA
M06	GATCACTTGTCGTAAAGCACTTTGGATGGGTTGAA
M07	TATCACCTGTCTTAAGACATTTTGGATGGGTTGAA
M08	TATCACTTGCCTTAAGACATTTTGGATGGGTTGAA
M09	GATCACTTATCGGATGATACTTTGGATGGGTTGAA
M10	TATCACTTGCCTTCGGTCACTTTGGATGGGTCGAA
M11	TATCTCTGGTATTATGACACTATGGATGGGTTGAA
M12	GATCACAGGTCGGCAGACTCAATGGATGGGTTGAA
M13	TATCTCAGGTGGTTAGACGCTTTGGATGGGTCGAA
M14	GATCACAGGTCCTACGATACTATGGATGGGTTGAA
M15	TATCACTTGTCTTACGACACTATGGATGGGTTGAA
M16	GATCACTTGTCGTTAGGCACTTTGGATGGGTTGAA
M17	GATCACTTGTCTTCAGACACTTTGGATGGGTTGAA
M18	GATCACTTGTCGTTATACACTTTGGATGGGTTGAA
M19	TATCACTTGTCGTAGAATACAATGGATGGGTTGAA
M20	TATCACTTGCCGTATGACACTGTGGATGAGTAGAA
M21	GATCACTTGTCGTATGACACTTTGAATGGGTTGAA
M22	TATCACTTGTCGAAAGACACTTTGGATGGGTTGAA
M23	TATCACTTGTCATAAGACACTTTGGATGGGTTGAA
M24	GATCACTTGGCGTAGGACACTTTGGATGGGTTGAA
M25	TATCACTTGTCGTACGACACTTTGGATGGGTTGAC
M26	TATCACTTGTCGTCGGGCACTTTGGATGGGTTGAA
M27	TATCACTTGTCGAAGGACACTTTGGATGGGTTGAA
M28	TATCACTAGTCGTAAGACTCTTTGGATGGGTTGAA
M29	GATCACTTGTTGAAAGACACTTTGGATGGGTTGAA
M30	TATCACTTGCCGTAAGACAGTTTGGATGGGTTGAA
WT F1*U	TATCACTTGTCGTAAGACACTTTGGATGGGTTGAA
F1*U^m	TATCACAGCGTTTTGACGGTAATGGATGGGTCGAA

Table S2-2: Sequences of the variants that were individually assayed by PAGE.
Each mutant is synthesized with the catalytic core in template replaced by the listed mutant sequences. Two oligos are synthesized with overlapping region and template is synthesized through anneal and extend PCR.

Generation	Tournament size	Number of parents	Number of pure recombinants	Number of mutants	Population size (Number of offspring)
1	N/A	N/A	N/A	N/A	10000
2	300	200	N/A	N/A	2000
3	300	200	N/A	N/A	2000
4	300	200	N/A	N/A	4000
5	300	200	N/A	N/A	6000
6	300	200	4000	2000	6000
7a	50	1000	8000	2000	10000
7b	50	1000	800	200	1000
7c	Generated by shuffling generation 7b				1000
Computational evolution	32	1000	4800	1200	6000
8	32	1000	9600	2400	12000

Table S2-3: List of parameters used to design the genetic algorithms employed in this study.
The parameters were varied for each generation of sequence design to reflect different stages of fitness landscape exploration and algorithms evaluation. More detailed explanation of the choice of parameters can be found in each chapter.

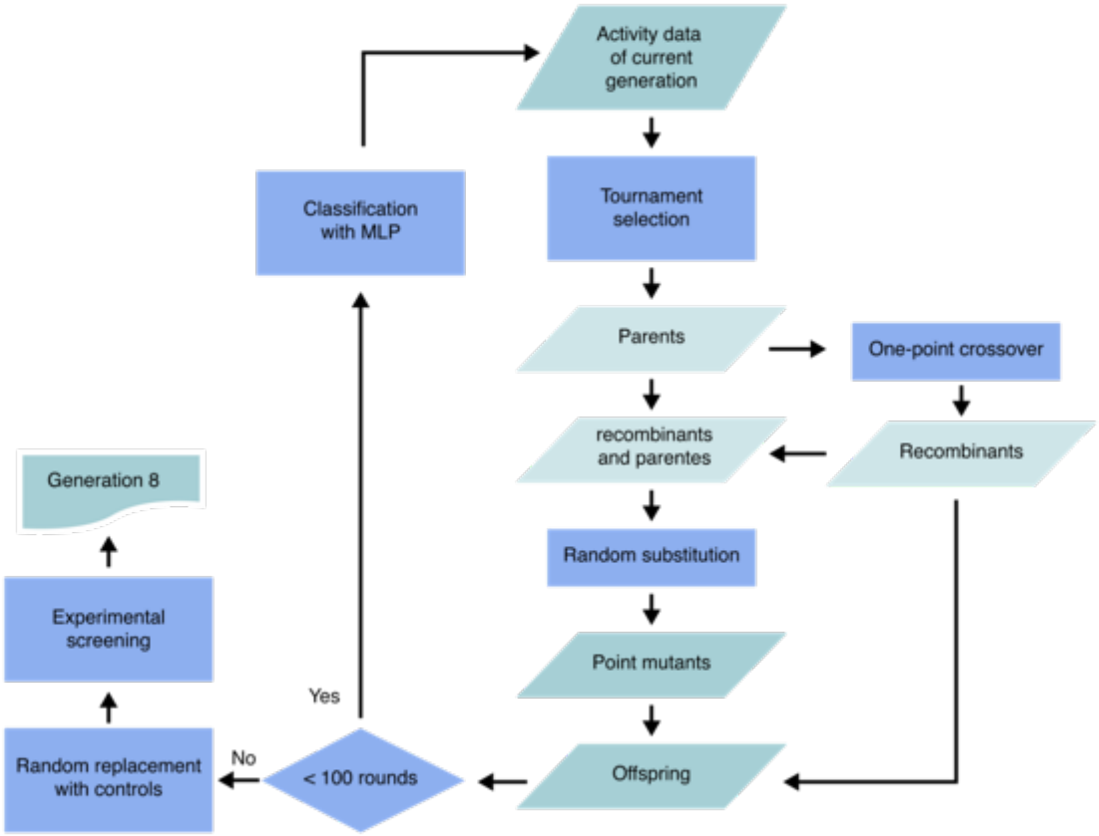


Figure S3-1: Overview of the algorithm used to design generation 8. Generation 8 was designed by 100 rounds of computational selection, mutation, recombination and multilayer perceptron (MLP) model classification. The MLP model was trained by data from generation 1 to 7. After computational evolution, the final population was selected by ensuring all variants are classified as neutral by the MLP.

Logistic/Linear regression	
Hyperparameter	**Range**
Learning rate	0.01,0.001,0.0001
Batch size	32,64,128
Multilayer perceptron	
Hyperparameter	**Range**
Learning rate	0.01,0.001,0.0001
Batch size	32,64,128
Hidden units per layer	32, 64, 128
Number of layers	1, 3, 5
Convolutional neural network	
Hyperparameter	**Range**
Learning rate	0.01,0.001,0.0001
Batch size	32,64,128
Number of filters in 1D convolution layer	32, 128
Number of 1 D convolution layer	1 , 2
Width of convolution kernel	3, 6

Table S3-1: Range of hyperparameters tuned by grid search for neural network.

Logistic/Linear regression, multilayer perceptron (MLP) and convolutional neural network (CNN) were designed using grid search of the hyperparameters listed in this table. More detailed description of the tuning procedure can be found in the Methods section of Chapter 3.

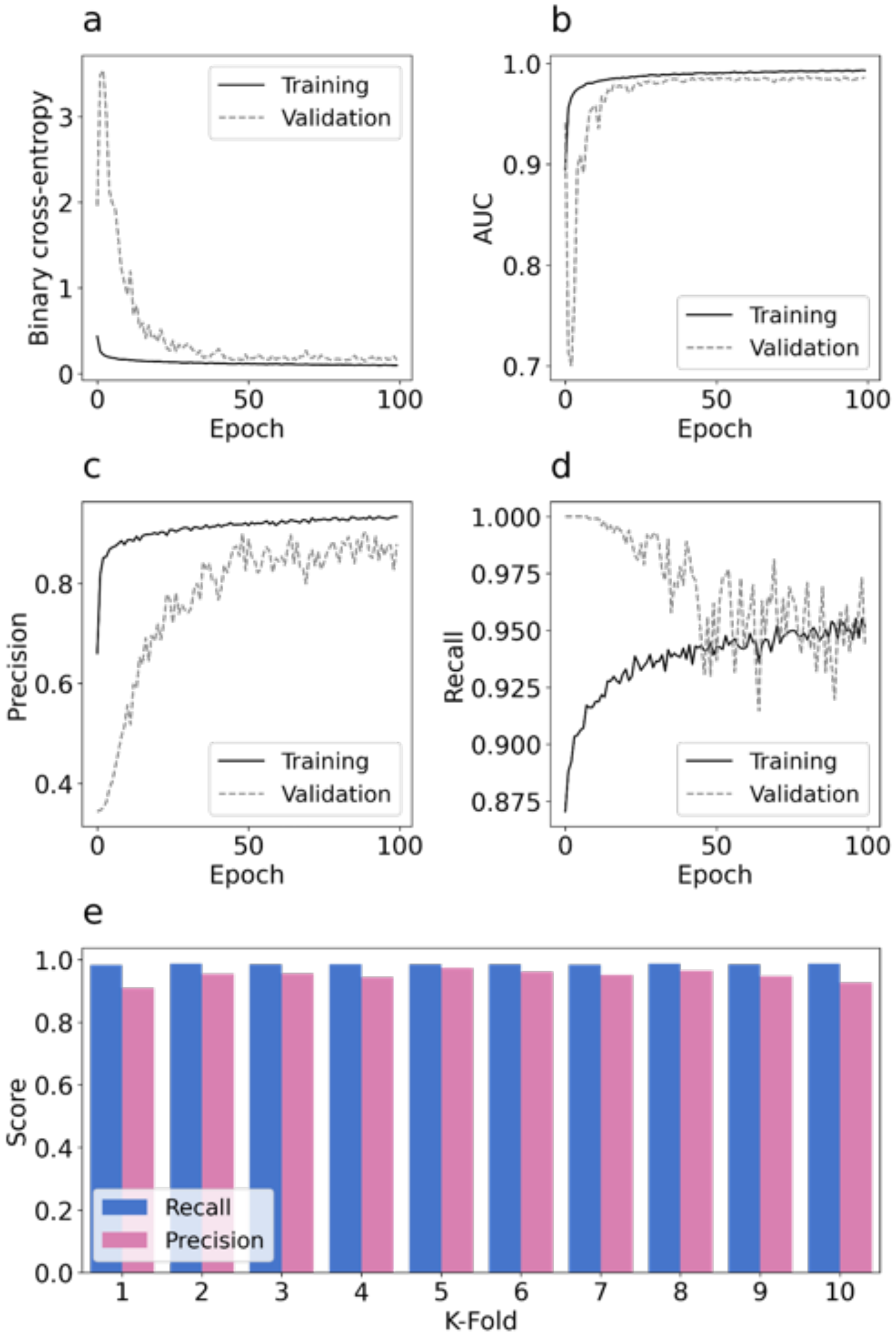


Figure S3-2: Training and evaluation of multilayer perceptron (MLP) model. Model was trained on a total of 26,374 variants and 2,930 variants were used for validation at the end of each epoch. Training was conducted for 100 epochs and tracked by a) binary cross-entropy loss b) area under curve (AUC), c) precision and d) recall. e) Precision and recall of each fold during 10-fold cross validation on a total of 41,863 variants from generations 1–7.

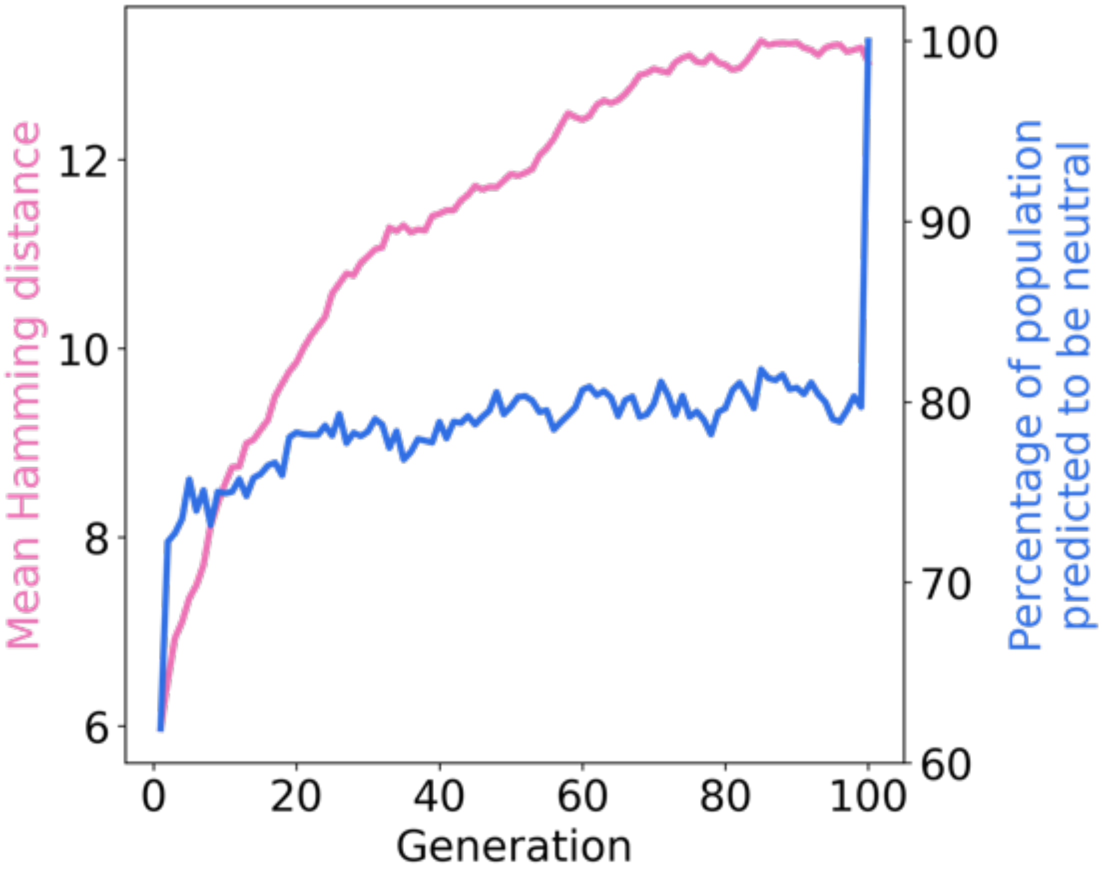


Figure S3-3: Computational evolution of generation 8.
During the MLP-guided computational evolution, the mean Hamming distance and percentage of the population predicted to be neutral by MLP were tracked after each generation. In generation 100, the variants were picked only if predicted to be neutral by the MLP.

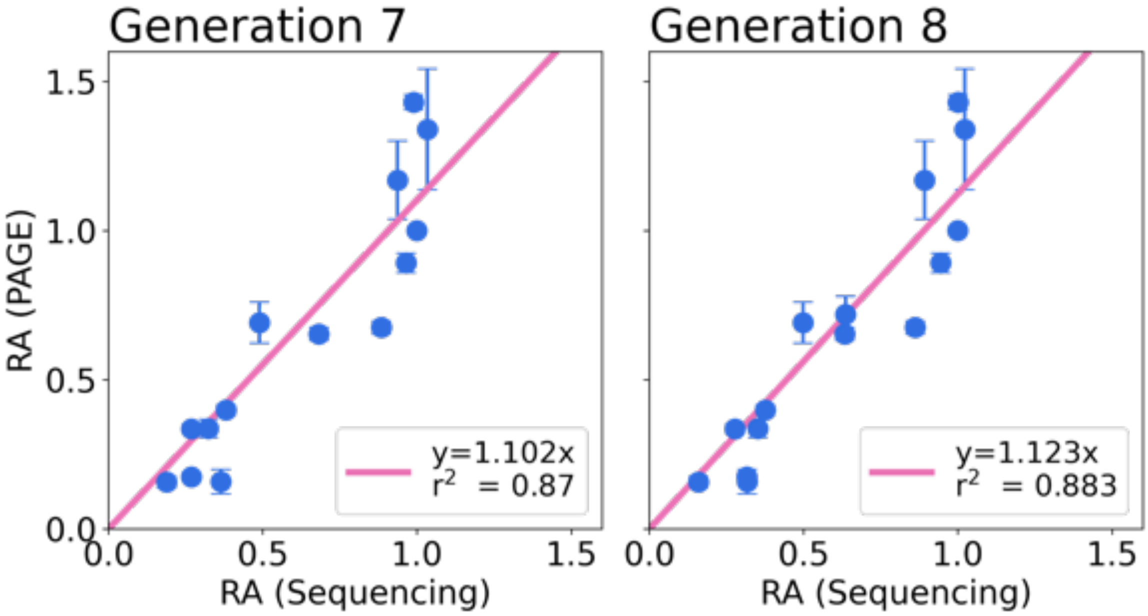


Figure S3-4: Reproducibility between sequencing assay and PAGE for generation 7 and 8.

The PAGE measured RA is compared with the RA measured using the sequencing method. The data points are presented as mean values +/− SD with n = 3 for the PAGE values and n=2 for sequencing values.

Chapter 4

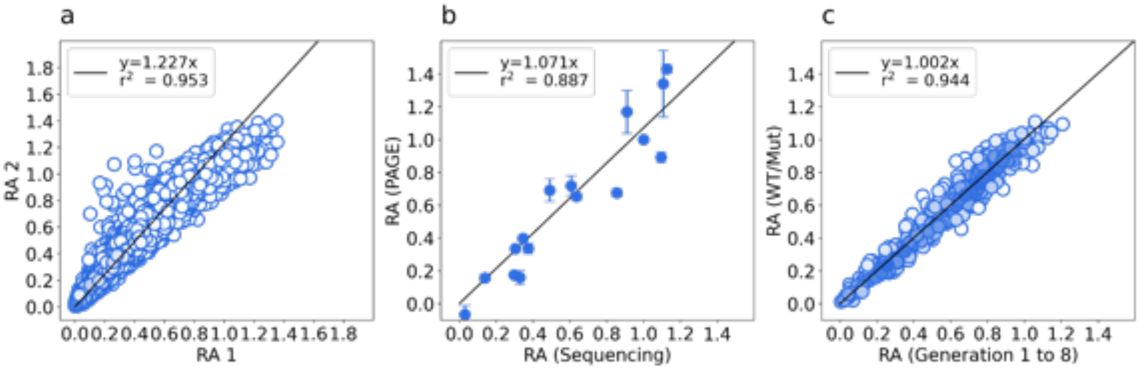


Figure S4-1: Reproducibility of WT/Mut combinatorial library.
a) Reproducibility of RA calculated by two independent sequencing experiments. b) Correlation between RA values from sequencing and PAGE assays. Data are presented as mean values +/− SD with n = 3. c) RA values of 441 variants screened in both WT/Mut library and generations 1 to 8.

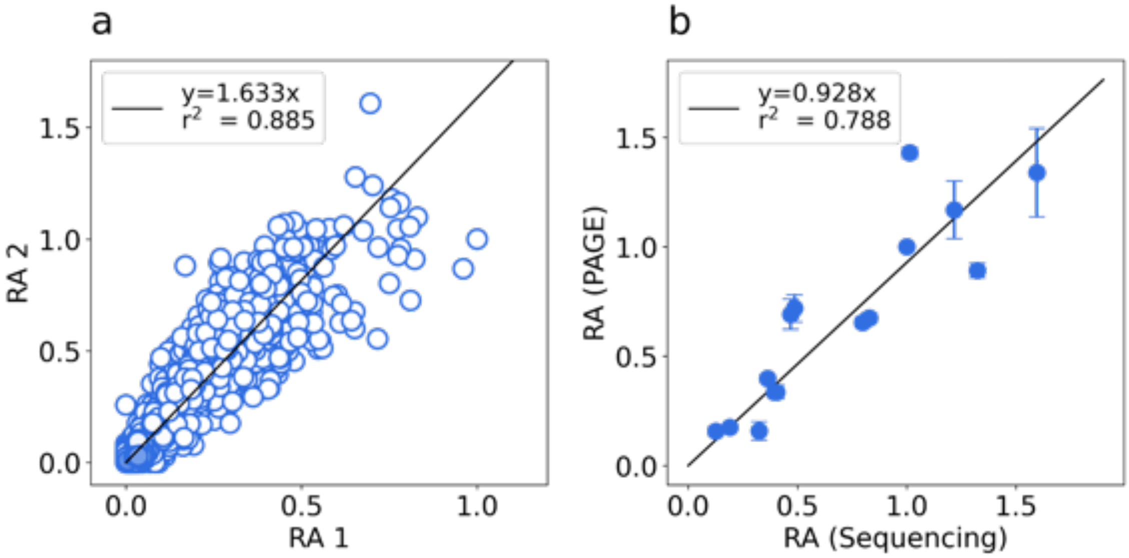


Figure S4-2: F1*U^m Reproducibility of F1*U^m combinatorial library.
a) Reproducibility of RA calculated by two independent sequencing experiments.
b) Correlation between RA values from sequencing and PAGE assays. Data are presented as mean values +/− SD with n = 3.