Title

Discovering pathways through ribozyme fitness landscapes using information theoretic quantification of epistasis.

Journal

RNA, 29(11)

Authors

Charest, Nathaniel

Shen, Yuning

Lai, Yei-Chen

et al.

Publication Date

2023-11-01

DOI

10.1261/rna.079541.122

Discovering pathways through ribozyme fitness landscapes using information theoretic quantification of epistasis

NATHANIEL CHAREST,[1] YUNING SHEN,[1] YEI-CHEN LAI,[2,3] IRENE A. CHEN,[1,3] and JOAN-EMMA SHEA[1]

[1]Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106, USA
[2]Department of Chemistry, National Chung Hsing University, Taichung City 40227, Taiwan
[3]Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, California 90095, USA

ABSTRACT

The identification of catalytic RNAs is typically achieved through primarily experimental means. However, only a small fraction of sequence space can be analyzed even with high-throughput techniques. Methods to extrapolate from a limited data set to predict additional ribozyme sequences, particularly in a human-interpretable fashion, could be useful both for designing new functional RNAs and for generating greater understanding about a ribozyme fitness landscape. Using information theory, we express the effects of epistasis (i.e., deviations from additivity) on a ribozyme. This representation was incorporated into a simple model of the epistatic fitness landscape, which identified potentially exploitable combinations of mutations. We used this model to theoretically predict mutants of high activity for a self-aminoacylating ribozyme, identifying potentially active triple and quadruple mutants beyond the experimental data set of single and double mutants. The predictions were validated experimentally, with nine out of nine sequences being accurately predicted to have high activity. This set of sequences included mutants that form a previously unknown evolutionary "bridge" between two ribozyme families that share a common motif. Individual steps in the method could be examined, understood, and guided by a human, combining interpretability and performance in a simple model to predict ribozyme sequences by extrapolation.

Keywords: ribozyme; fitness landscape; epistasis; mutual information; surprisal

INTRODUCTION

The mapping of genotype to phenotype for functional biopolymers implicitly captures information about structural contacts and mechanism. Fitness landscapes are mathematical maps that relate primary sequence to functional properties, such as catalytic rate enhancement for enzymes or ribozymes. Understanding these landscapes, particularly for RNA, may yield insights into mechanisms as well as the molecular evolution of early life (Athavale et al. 2014; Pressman et al. 2015). The development of quantitative tools and high-throughput experiments for elucidating and analyzing fitness landscapes is, therefore, a major front in research efforts to understand these systems (Kinney and McCandlish 2019). High-throughput collection of data has been used to characterize the fitness landscapes of RNAs (Pitt and Ferré-D'Amaré 2010; Jiménez et al. 2013; Puchta et al. 2016). For example, kinetic measurement using high-throughput sequencing (e.g., *k*-Seq) is able to measure the activities of tens of thousands of ribozyme sequences (Yokobayashi 2020; Shen et al. 2021).

Nevertheless, even with high-throughput experimental techniques, only a small fraction of possible sequence space can be sampled, due to both synthetic and analytical limitations. For example, a fully randomized 30 nt region in a ribozyme sequence would yield 10^{18} different sequences, far exceeding current high-throughput sequencing capacity. Therefore, computational methods are required to predict activities for sequences that were not captured by the empirically available data. Such data presents computational challenges for interpretation, and improved analytical techniques are required to quantitatively characterize fitness landscapes and develop models that advance understanding of the genotype–phenotype relationship.

In this work, we focus on an activity that would have been foundational to the genetic code of protein translation, perhaps the greatest evolutionary invention of an early RNA-based prototypical life (Pressman et al. 2019). A key activity of this process is the covalent attachment of amino acids to

Corresponding authors: shea@chem.ucsb.edu, ireneachen@ucla.edu

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.079541.122. Freely available online through the *RNA* Open Access option.

tRNAs (de Duve 1988), which is catalyzed in contemporary biology by aminoacyl-tRNA protein synthetases. In the preprotein world, however, this activity may have been achieved by self-aminoacylating ribozymes. This hypothesis is supported by the existence of ribozymes that react with aminoacyl adenylates or other activated substrates (Illangasekare and Yarus 1999; Lee et al. 2000; Murakami et al. 2006; Chumachenko et al. 2009). Prior work (Pressman et al. 2019) determined the catalytic activities of thousands of self-aminoacylating ribozymes that react with 5(4*H*)-oxazolones, considered to be prebiotically relevant substrates (Liu et al. 2016). Here, we develop and validate a computational method for extracting additional predictive power from the limited experimental data of the fitness landscape.

Existing methods, such as minimum epistasis interpolations (Zhou and McCandlish 2020) and Gaussian processes (Romero et al. 2013), show promise for interpolating missing data on fitness landscapes and "filling in the map" for regions of sequence space where data is not complete. However, these methods struggle with sparse sampling or require prerequisite knowledge, such as structural data from the Protein Data Bank (Shroff et al. 2020), which are not always accessible for the novel sequences. While modern methods can interpolate fitness landscapes given sufficient sampling, methods for extrapolative predictions looking beyond the boundaries of sampled space are relatively lacking. For example, using information about double mutants of a central sequence to predict activities for triple or quadruple mutants remains an open problem.

A simple extrapolative technique could be based on additivity in the genotype–phenotype map, in which the effects of single mutations on the genotype would be summed to predict the phenotype of the combination. Chemically speaking, additivity corresponds to a separability of chemical moieties that do not interact with one another in the reaction mechanism. For example, a residue that stabilizes the active fold might not interact with a residue that exclusively forms a contact in the transition state. However, additivity is generally not a correct assumption in detail since different residues influence one another through direct contacts or indirect effects (epistasis). Epistatic landscapes feature mutations whose effects are influenced by their genetic context. Attempts to model epistasis include using simple nonlinear functions to capture latent, nonepistatic traits (Kondrashov and Kondrashov 2015; Starr and Thornton 2016; Sailer and Harms 2017; Otwinowski et al. 2018), and machine learning models (Sarkisyan et al. 2016; Yang et al. 2019). The former technique performs well for relatively simple systems in which there is a largely additive landscape subject to random variation, but not for more complex landscapes. Machine learning has strong general capability but requires considerable finesse in parameterization and can pose difficulties with interpretation. In one recent study, in silico evolution was performed on ribozyme variants using empirically determined fitness values, with a deep learning perceptron model applied in the final round. While the approach effectively identified neutral mutants of the ribozyme, the perceptron itself constituted a "black box" (Rotrattanadumrong and Yokobayashi 2022). Therefore, methods that combine performance and interpretability are needed.

One possible approach is to use mathematical language to construct an articulation of epistatic complexity that remains accessible to human insight. The method described here applies information theory to identify regions of sequence space where noninterfering mutations can be exploited to extrapolate beyond the boundaries of the measured space. Instead of fitting a function to the fitness landscape, this method identifies mutations that are likely to yield high activity when combined. Epistasis has previously been analyzed in a probabilistic framework (Ostman et al. 2012). Here, we relate the epistatic quantity to information theory and demonstrate its ability to provide a pairwise decomposition of the information contained in the data set. This pairwise representation can be exploited to create a predictive model without fitting parameters.

Noting that mutual information has seen success improving prediction outcomes when integrated into models of the sequence-activity relationship (Moore et al. 2006; Kinney et al. 2007; Atwal and Kinney 2016), we use surprisal (see Equation 1, below) and mutual information to calculate a quantity termed "epistatic divergence." We demonstrate that epistatic divergence can be used to derive insights from empirical data that extrapolate beyond the explored regions of their fitness landscapes. Using two families of ribozymes for which the activity of all possible double mutants of a central "seed" sequence had been measured, we predicted and validated points in the sequence space of triple and quadruple mutants with a high likelihood of activity. Epistatic divergence identified an evolutionary connection between two "islands" of activity within the fitness landscape, which we validated experimentally. Such extrapolation could be combined with interpolation algorithms to enable greater understanding of fitness landscapes.

This study proposes a representation of interactions in the sequence-activity landscape, in which qualitative properties of a system are articulated mathematically. Representations are an essential part of model development (Bengio and Lecun 2007; Bengio et al. 2013) that affect the fundamental ability to observe patterns in the data. This epistatic divergence representation explicitly captures the degree to which the sequence-ribozyme activity relationship is epistatic, which subsequently enables precise exploitation of noninterfering mutations for extrapolative predictions.

RESULTS

Epistatic divergence as a measure of epistasis

We analyzed a data set of ribozyme variants of a self-aminoacylating RNA sequence (Pressman et al. 2019), S-1B.1-a,

also referred to as a "seed" sequence here. We compared epistatic divergence against a traditional conception of epistasis, namely the difference from additivity of single mutations. For the conventional measure of epistasis, the average difference (μ) of a double mutant's effect on the activity from the sum of the constituent single mutants was calculated for all pairs (m,n) (site pairs) (Fig. 1A). To determine the effect of a single-site mutation, all double mutations applying to that site were included in the averaging, reflecting multiple genetic backgrounds present in the double and single mutant data. The standard deviations (σ) of these values were also calculated, indicating the spread of the differences from additivity (Fig. 1B). These measures reflect the difference from additivity when considering all possible nucleotide combinations across two sites.

For epistatic divergence, low values indicate a site pair (m,n) where the effects of the nucleotide combination were found to be either lacking impact on the activity, or were explainable by considering each site independently, or both. A high value of epistatic divergence indicates a site pair where the combination of nucleotides was found to be important and impactful upon the activity. We calculated the epistatic divergence using two possible values of the classification thresholds θ, namely classified based on activity less than or greater than (or equal to) the seed sequence ($\theta = a_{seed}$; Fig. 1C), or activity above or below the noncatalytic background rate ($\theta = a_{active}$; Fig. 1D). The former threshold (a_{seed}) is quite stringent (three out of 63 possible single mutants and 24 out of 1890 possible double mutants [Shen et al. 2021]) because the seed sequence is a ribozyme that reached high abundance during prior in vitro selection (Pressman et al. 2019), indicating high relative activity. This threshold choice was used to focus on sites that may cause activity enhancements close to or greater than the seed sequence. The second threshold (a_{active}), set at a catalytic rate equal to four times the background (noncatalytic) reaction rate, captures sites that influence whether a sequence is catalytically active at all.

Comparison of the epistatic divergence and the conventional measure of epistasis shows that a site of particular interest is position 38, which exhibits high σ despite low μ, indicative of highly variable epistasis depending on genetic background. Consistent with this, epistatic divergence is high for site 38, suggesting an important relationship

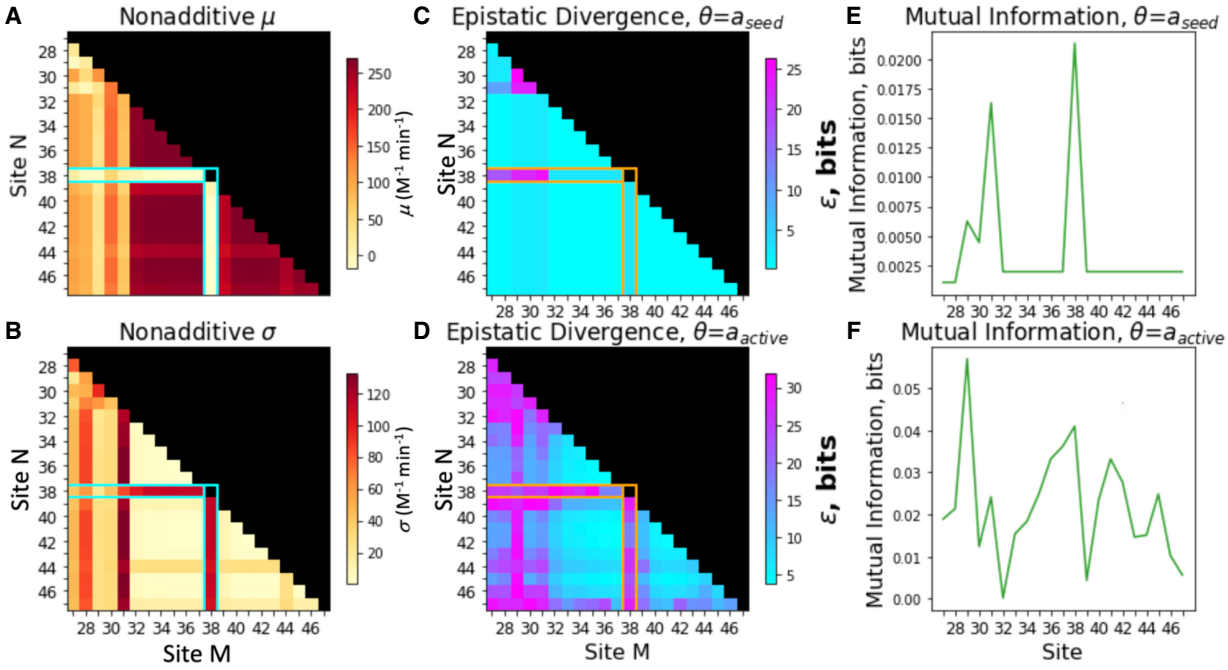

FIGURE 1. Comparison of epistatic divergence and deviation from additivity for ribozyme S-1B.1-a. (*A*) The ensemble average of double mutants' difference from the additive sum of single mutants, a conventional measure of statistical epistasis that shows deviation from additivity. (*B*) The standard deviations of the calculations from part A. (*C*) The epistatic divergence computed using the threshold a_{seed}. The seed sequence defines the sample population; all mutants in the experimental data are a Hamming distance of 1–2 from the seed. The median activity of the seed was taken as a_{seed}, and any sequence whose median activity was found above a_{seed} was marked superior while any whose median was found lower was marked inferior. Note that sequences with activity close to the seed sequence may be incorrectly classified due to experimental noise (Shen et al. 2021). (*D*) The epistatic divergence calculated using the threshold a_{active} based on the background catalytic rate determined in prior work (Janzen et al. 2022). (*E*) The mutual information depicting which sites along the sequence were found to have the greatest relevance to the classification around a_{seed}. (*F*) The mutual information calculated around the classification scheme with a_{active}. The goal of these measures is to detect the site significance to the catalytic activity of the ribozyme.

between activity and the nucleotide identity at this site. Importantly, epistatic divergence also highlights other regions of the sequence that do not appear unusual based on the traditional measure of epistasis, particularly when considering highly active sequences (Fig. 1C vs. 1A,B). Predictions based on the region highlighted by epistatic divergence, but not the traditional measure of epistasis, were tested experimentally (described below). These features demonstrate the ability of epistatic divergence to positively identify regions of interest in the ribozyme.

Conversely, the ability to correctly identify regions that are not of interest for extrapolative combination is also important. An advantage of epistatic divergence in this regard can be seen in the blocks of signal associated with the regions around sites 32–37 along site M and 39–47 on site N (seen as a low signal region in Fig. 1D). The μ values suggest a consistently large deviation from additivity, while the spreads (σ) are quite narrow (Fig. 1A,B). These effects are driven by the fact that these locations are, independently, essential to catalytic function. Mutations in these regions essentially eliminate activity, and so any double mutation of them will lead to high μ values due to a saturation effect. However, such patterns cannot be taken to reflect true interactions (i.e., mechanistic or structural) in the ribozyme. In contrast, when epistatic divergence is used, these sites are appropriately identified as lacking interactions. Thus, the epistatic divergence measure has high specificity in identifying loci with complex epistatic behavior that might be exploitable, particularly for predicting active sequences beyond the boundary of sequence space in the data set.

Identifying hotspots of exploitable complexity

To focus on the potential prediction of high-activity sequences, we used the epistatic divergence measure with $\theta = a_{\text{seed}}$ to develop a predictive model. The epistatic divergence highlighted sites where the data showed a particular dependence on pairwise states when considering the distribution of activity (Fig. 1C). Such pairwise interactions are expected to be important for high activity of the ribozyme. At the same time, mutual information (between a single site and the activity distribution) identifies single sites that are informative for activity. Combining mutual information and epistatic divergence should therefore identify mutations that are likely to interact synergistically in high-activity sequences. We leverage this fact to produce a simple model that maximizes the utility of noninterfering mutations within the landscape.

Specifically, the individual sites 30, 29, 31, and 38 gave the most information about activity, in ascending order (Fig. 1E). Epistatic divergence analysis indicated that the combinations of these loci are synergistic. These observations suggested extrapolative predictions for highly active sequences beyond the experimental data.

Extrapolative prediction of active ribozyme sequences

The epistatic divergence ε is the sum of terms describing each represented state, so the contributions from states can be decomposed into contributions from each specific mutant pair and sorted by activity class. We examined the specific epistatic contributions from states containing a combination of the top four individually informative sites (29, 30, 31, and 38) (Fig. 2), showing regions where pairwise epistatic effects contained more information than constituent sites considered alone, for the high-activity class (using $\theta = a_{\text{seed}}$). Four sites were included in the analysis in order to obtain predictions for quadruple mutants.

In theory, synergistic double mutants might be combined to generate triple and quadruple mutants expected to have high activity. For example, synergistic effects arise from the combinations (38A,31G), (38A,29C), and (31G,29C) (Fig. 2A–F). This observation suggested that the new triple mutant (G38A, A31G, A29C) may yield a high-activity variant. The following process was used to predict new high-activity ribozymes. The epistatic divergence attributed to each site pair (Fig. 2A–F) was examined to select potentially informative combinations among the four most highly informative sites (29, 30, 31, and 38). A strong signal was defined as >4 bits, corresponding to the amount of information needed to completely specify an RNA site pair. This list of strong signals was then searched for compatible combinations that would result in triple or quadruple mutants (Supplemental Table S1). Two pairs having a common mutation were considered compatible with each other if all of the mutation pairs of the resulting triple mutant were strong signals. For example, (29T,38A) is compatible with (38A,31G) because (29T,31G) is also a strong signal, predicting that the triple mutant (29T,38A,31G) should have high activity. Similarly, two pairs of triple mutants, sharing two of three mutations, were deemed compatible with each other if all pairs of the resulting quadruple mutant were strong signals. For example, (29C,30A,31G), a compatible triple mutant, is compatible with (29C,31G,38A), also a compatible triple mutant, because (30A,38A) is also a strong signal. This procedure can be visualized as a network of single mutations, where nodes (single mutations) are connected if the pair constitutes a strong signal. Compatible triple or quadruple mutants are thus found as completely connected triangles or quadrilaterals (i.e., in which every node is connected to every other node in the subgraph; Fig. 2G). This procedure yielded 12 triple mutants and three quadruple mutants that were predicted to have superior activity, assuming that the double mutation information could be combined to produce triple and quadruple mutant predictions. Of these, all of the quadruple mutants were prioritized for experimental testing, since they represent a greater extrapolation compared to triple mutants. Of the triple mutants,

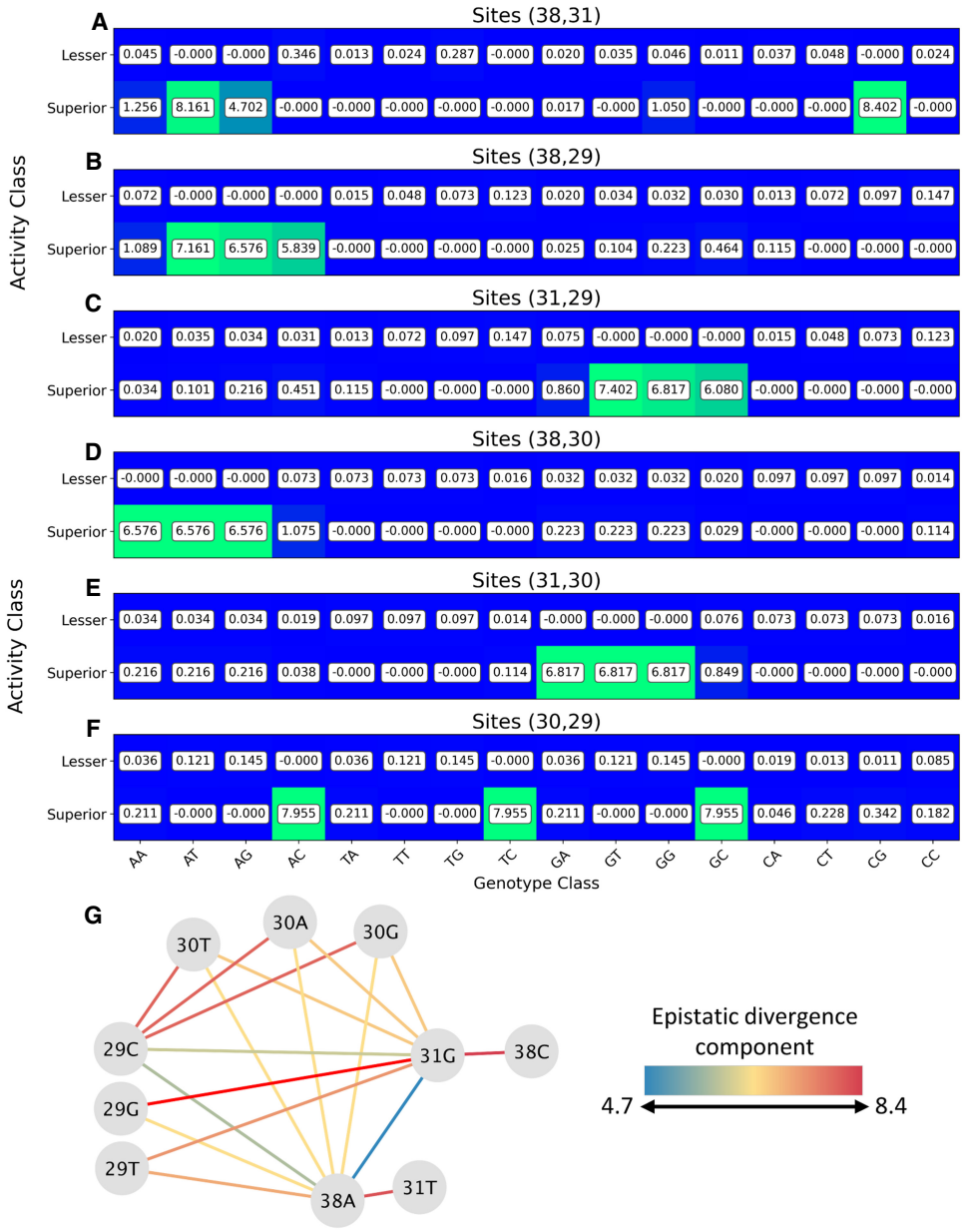

FIGURE 2. Decomposition of epistatic divergence by sites for ribozyme S-1B.1-a. (*A–F*) The matrix components of the epistatic divergence calculations for the indicated site pairs. The decomposition was used to identify potentially compatible mutations, which are genotypes associated with improved function over the seed ribozyme that can relate to other pairs of loci. The combinations result in triple or quadruple mutants that are predicted to be likely to exhibit appreciable activity. These predictions extrapolate beyond the mapped fitness landscape. (*G*) The network of single mutations, in which strong signals are represented by edges. Compatible triple or quadruple mutations are illustrated as completely connected subgraphs. Quantities given are in bits with accompanying heat maps to aid the eye.

half (six) were chosen for experimental testing due to feasibility constraints. The three triple mutants involving the three most informative sites (29, 31, and 38) were all chosen for testing. Of the remainder, triple mutants containing 29C and 38A were prioritized over mutants containing 31G because sites 38 and 29 showed the highest mutual information for $\theta = a_{\text{seed}}$ or $\theta = a_{\text{active}}$, respectively (Fig. 1E,F). The sequences selected for testing are given in Table 1.

Experimental testing of the predicted triple mutant ribozymes

The six triple mutant sequences designed by following the inference process above (Table 1) were used to search the previously obtained high-throughput ribozyme assay data set (Janzen et al. 2022). Although that data set had not been designed to comprehensively cover triple mutants

TABLE 1. Predicted sequences from data on variants of ribozyme S-1B.1-a

Predicted triple mutants	Predicted quadruple mutants
s-1B-29C30A38A	s-1B-29C30T31G38A
s-1B-29C31G38A	s-1B-29C30A31G38A
s-1B-29T31G38A	s-1B-29C30G31G38A
s-1B-29C30T38A	
s-1B-29G31G38A	
s-1B-29C30G38A	

of the seed sequence, some triple mutants had been synthesized by chance along with the variant pool. The distribution of measured activities is shown for analyzable triple mutants from those data (Fig. 3A), with an emphasis on triple mutants found to have high activity (>a_{seed}) (Fig. 3B). The six predicted triple mutants indeed had outperformed seed S-1B.1-a, and ranked in the top 30 out of more than 35,000 triple mutants analyzed. Precisions for these measurements are given in Supplemental Figures S1, S2. Since more active sequences are more likely to have higher relative abundance in the reacted pool (Supplemental Fig. S3), this observation is consistent with the expectation of higher activity level in these triple mutants.

While the epistatic divergence method shows excellent specificity in identifying high-scoring triple mutants, it should be noted that other triple mutants with top-scoring median activities were not detected by the method. The method identifies pairwise contributions to higher activity that are likely compatible, resulting in a set of high-order mutants as candidates for testing. The model relies on an expectation that compatible double mutations would not interfere with each other to cause decreased fitness. In other words, in the case of triple mutants, if mutants **AB**, **BC**, and **AC** all have high activity, then mutant **ABC** is predicted to have high activity. (For quadruple mutants, if **AB**, **BC**, **AC**, **AD**, **BD**, and **CD** all have high activity, then **ABCD** is predicted to have high activity.) This expectation is reasonable if epistatic effects diminish at higher orders beyond pairwise interactions (Zhou et al. 2022). Conversely, higher-

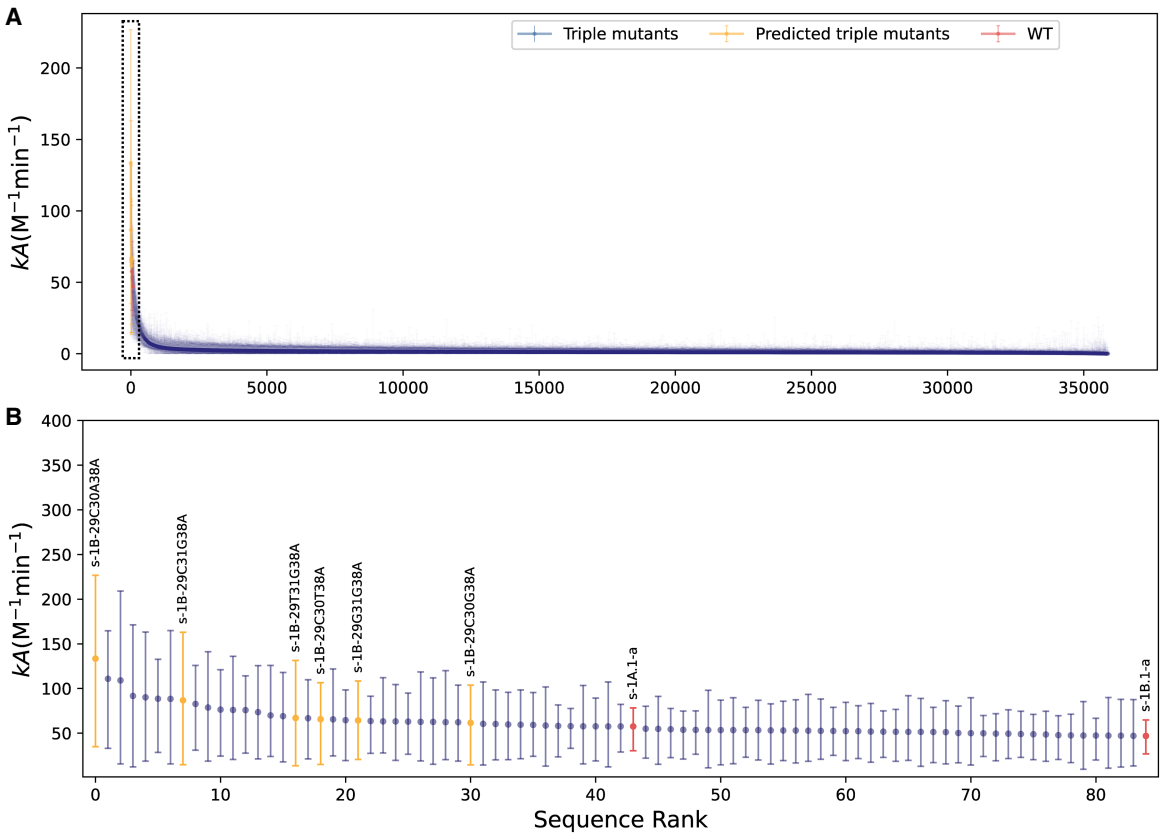

FIGURE 3. Experimental activity measurements (*k*-Seq [Janzen et al. 2022]) for triple mutants of S-1B.1-a. Measurements are shown as median values with 95% confidence intervals. (*A*) All analyzable triple mutants in the pool, ranked in terms of median *kA* values for activity. (*B*) The top 84 sequences, including seed sequence S-1B.1-a and separate seed sequence, S-1A.1-a (red). The predicted triple mutants (orange) were associated with improvements over S-1B.1-a's activity and were generally among the top-ranking activities. See Supplemental Figures S1–S3 for measurement precisions.

order mutants might be missed if higher-order epistasis is significant, for example, if the double mutant subsets of a high-scoring triple combination are not high-scoring, leading to lowered sensitivity of this method for predicting active mutants.

The success of the six predictions suggests this assumption is sometimes appropriate, but violations of this assumption could explain the high-activity triple mutants that were missed in this process.

Prediction of an evolutionary pathway through a quadruple mutant ribozyme

We also analyzed the epistatic divergence and mutual information for a related ribozyme, S-1A.1-a. Sequence S-1A.1-a and S-1B.1-a are related by a shared motif (Fig. 4) offset by two sites, but they contain distinct flanking regions and are separated by a total edit distance of six (Hamming distance = 16). Interestingly, some mutations suggested by the epistatic divergence analysis of variants of S-1B.1-a were noted to decrease the edit distance to sequence S-1A.1-a. Specifically, A29C, C30T, and G38A would reduce the edit distance between these two ribozyme families. Furthermore, the epistatic divergence analysis for variants of S-1A.1-a (Fig. 5) indicated that site 29 is highly informative, and the major signal from epistatic divergence occurs at the (29G, 27A/T/G) pair. Inspection of the sequence alignment indicates that a (29G, 27A) double mutant of S-1A.1-a would reduce the edit distance to sequence S-1B.1-a by two (Fig. 4). These considerations indicate a possible connection between the S-1A.1-a and S-1B.1-a families, suggesting there may be an evolutionary path of active ribozyme variants between them.

Thus, epistatic divergence analysis predicted that high activity would occur with mutation of S-1A.1-a to resemble S1B-29C30T31G38A, and conversely that high activity would occur with mutation of S-1B.1-a to resemble S-1A.1-a. This suggested the presence of a specific, high-activity evolutionary pathway consisting of active mutants

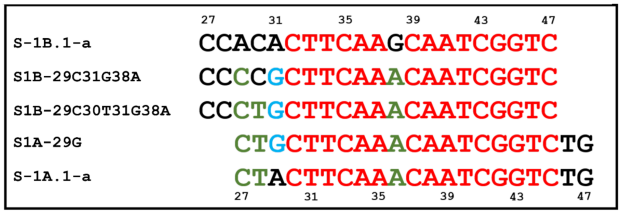

FIGURE 4. Sequence comparison of seed sequences and mutants indicated by epistatic divergence analysis. In red is the shared motif linking S-1A.1-a and S-1B.1-a families. In green are mutations characteristic of S-1A.1-a and indicated by epistatic divergence analysis as improving S-1B.1-a. In blue are shared residues indicated by epistatic divergence analysis for both S-1A.1-a and S-1B.1-a families. These predictions suggest a possible evolutionary pathway connecting S-1A.1-a and S-1B.1-a.

to connect these two ribozyme families. We tested the activity of the intermediate mutants (S1A-29G, S1B-29C30T31G38A, and S1B-29C31G38A) individually experimentally. Reaction with the substrate yields a biotinylated product that can be separated using streptavidin beads and quantified by RT-qPCR. Measurement of reaction product over a concentration series allows determination of the catalyzed rate (Shen et al. 2021). The predicted intermediate mutants indeed exhibited high activity, with some activities being higher than either seed sequence S-1A.1-a or S-1B.1-a, validating the existence of the predicted evolutionary connection (Fig. 6).

DISCUSSION

The epistatic divergence description of pairwise interactions within the primary sequence of self-aminoacylating ribozymes enabled the extrapolation of previously unknown active sequences, at mutational distances beyond the training space. This was accomplished by considering a data set of activities measured via *k*-Seq experiment, for sequences within two point mutations of a central seed sequence, applying information theory to describe the information content of the distribution of activities in terms of pairs of residue identities, and determining where these pairs possess noninterfering or synergistic behavior that can be exploited to predict highly active sequences beyond the training space. Through this process, the most informative (i.e., highest surprisal) observations of pairwise mutants were used. The most informative observations were combined whenever an internally consistent extrapolation was possible. In other words, epistatic divergence identified where a measurement of activity in double mutants most deviated from the expectation of mutual independence between sites. To generate an extrapolative prediction of triple or quadruple mutants (which were not in the training data set), mutations were combined whenever all of their pairwise interactions were positive (Fig. 2G).

Because of the simplicity of this parameter-free approach, the results can be readily interpreted with the language of information theory while simultaneously offering a pragmatic means to identify regions of activity within a fitness landscape, and potential evolutionary pathways, with high specificity.

The epistatic divergence can be mathematically cast as the sum of information contents describing the degree to which a data point contributes to the knowledge of whether a pair of residues predicts an active sample or an inactive sample. Unlike machine learning methods that predict a distribution of activity over the sequences, this method mathematically identifies the parts of the sequence that yield the most information for predicting the activity. This approach systematizes empirical methods that rely on manual curation of significant residues (Miton et al. 2020). Subsequent analysis then allows the construction of predictions

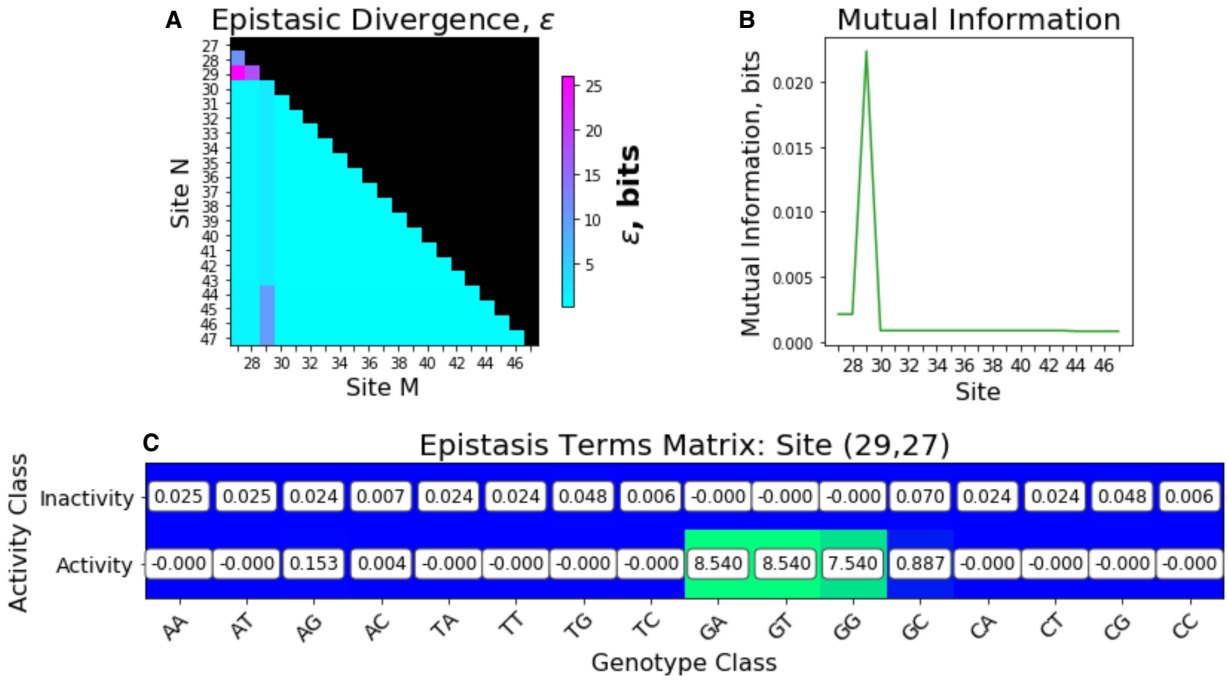

FIGURE 5. Epistatic divergence analysis for variants of ribozyme S-1A.1-a. Epistatic divergence, mutual information and decomposition from data on single and double mutants of S-1A.1-a. These suggest mutations that bring its sequence closer into alignment with a quadruple mutant of S-1B.1-a that was predicted to be active (S1B-29C30T31G38A).

outside the training set, by combining the sequence identities that were associated with the most informative data points within the training set. Mutual information has been previously used for detecting coupled variables in biological contexts. Examples include analyzing combinations of SNPs in genetic studies (Moore et al. 2006; Moore and Hu 2015) and predicting contacts between residues in protein (Gloor et al. 2005; Dunn et al. 2008) or RNA (Chiu and Kolodziejczak 1991; Freyhult et al. 2005) molecules. In this work, we further identify beneficial genotypes using surprisal and combine compatible genotype pairs to detect active higher-order mutants.

Recent work on genotype–phenotype mapping by minimizing epistatic interactions develops a model that allows for maximally locally additive behavior indicated by training observations (Zhou and McCandlish 2020). This assumes an underlying preference for nonepistatic behavior, that is then qualified by the epistasis present in the observations of the training pool. Conversely, the epistatic divergence is not a model in the traditional sense of producing a mathematical construction that can generate predictions. Rather, epistatic divergence is a quantitative representation of how informative a given observation is. In this context, an "observation" is a phenotype class ("active" or "inactive") paired with a genotype class (e.g., "A" in position 38 and "G" in position 29) that is present in the training data set. This defines the support set of the independent variables being used to describe the system. Epistatic divergence quantifies information content relative to the rest of the pool, such that the most informative observations can be used for subsequent prediction building. In our model, we combine observations that were identified as both highly informative as well as pertaining to the active class. Because we explicitly compute over pairs, we explicitly capture the interactions, epistatic or otherwise, described by those pairs. This results in predictions that are combinations of the most informative pieces of information in the data set.

Understanding genotype–phenotype maps, and predicting highly active sequences, are important twin goals in biomolecular engineering. Due to the astronomical size of sequence space, for which the mass of a pool containing every possible protein-coding sequence would readily outstrip the mass of the Earth, computational extrapolation will always be necessary to understand genotype–phenotype maps beyond a few dozen residues. Furthermore, given that the vast majority of sequences are inactive, and that the majority of mutations are deleterious to function, computational methods to make accurate, specific predictions are invaluable for identifying novel functional sequences. In this work, the analysis resulted in prediction of 12 highly active triple mutants (of which six were tested experimentally) and three highly active quadruple mutants (of which all were tested experimentally). It is notable that nine out of nine predicted sequences chosen for testing yielded highly active sequences. In a previous study measuring activities by kinetic

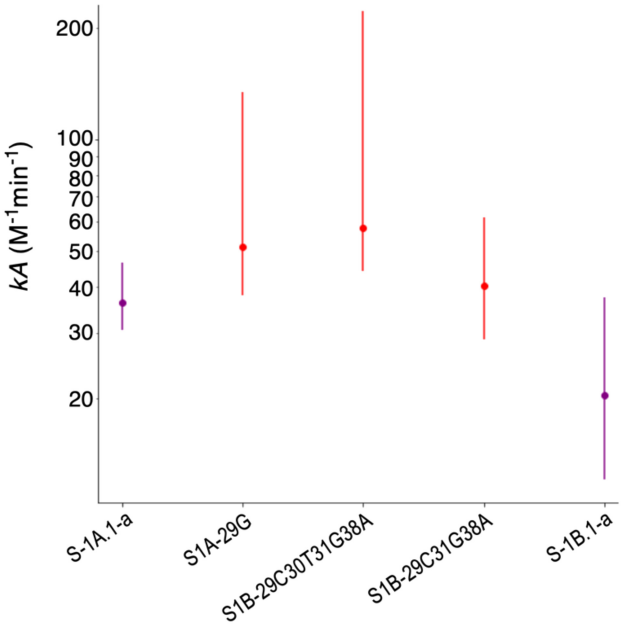

FIGURE 6. Ribozyme activities along a predicted evolutionary pathway (median and 95% confidence interval), measured by qPCR assay. The endpoints on the *x*-axis are the seed sequences (purple), with intermediate mutants as shown (red). Mutant sequences were predicted by extrapolation from the data by analyzing epistatic divergence.

sequencing, ~5% of all single mutants and 1% of all double mutants were found to have high activity. The previous study using a doped library was not designed to measure all possible triple or quadruple mutants, but many were still measured though at low precision due to a small number of sequencing reads. Of these triple and quadruple mutants, 0.2% or less were found to have high activity (Supplemental Table S2; Shen et al. 2021). Therefore, the epistatic divergence method described here compares favorably in identifying active mutants (9/9) compared with the very low frequency of active mutants from an unbiased sample. While the double mutant data, on which this method is based, was comprehensive (i.e., including all possible double mutants), no data on triple or quadruple mutants was used for the predictions. However, the predictive power of this method is likely to decrease for higher-order mutations, since the method assumes that higher-order epistatic interaction is relatively small when predicting mutants. Progress in increasing the throughput of synthetic and analytical techniques would be useful for building larger experimental data sets to validate predictions.

Furthermore, the pattern of these mutants revealed a previously unknown neutral evolutionary pathway of highly active sequences through the fitness landscape, which joined the two ribozyme families centered on S-1A.1-a and S-1B.1-a. In particular, while the experimental data set used for the analysis here described only the local fitness peaks (within a mutational distance of two) around sequences S-1A.1-a and S-1B.1-a, the epistatic divergence specifically illuminated multiple high points outside this region, as well as an evolutionary connection that was previously unknown. An experimental approach to the same goal of discovering new fitness peaks and an evolutionary pathway, while not impossible, would have been significantly more laborious.

Machine learning approaches have been applied to the problem of predicting active sequences by extrapolation from mutational data. For example, a random forest model was applied to predict active mutants of a self-cleaving ribozyme (Breiman 2001; Beck et al. 2022). While often successful in generating predictions, random forest models average over many decision trees and thereby create a difficulty in interpreting the process itself. Deep learning models, such as multilayer perceptrons or Long-Short Term Memory networks (Schmidt and Smolke 2021; Beck et al. 2022; Rotrattanadumrong and Yokobayashi 2022), improve the representation of the data and extract features found to be significant to the endpoint being modeled. However, deep models are complex, requiring many parameters, and interpretability remains an unsolved problem (Kirboga et al. 2023). In this context, an advantage of the analysis presented here is that the statistical quantities are not based on fitted parameters, but are rather calculated directly from the data, and the prediction process follows well-defined steps from the calculation of epistatic divergence components to the assessment of mutant combinations.

Thus, in the epistatic divergence analysis presented here, the steps of the method are directly interpretable in real terms, and the analysis itself is an interactive process with the data, allowing insight into the genotype–phenotype map. The analysis here shows how epistatic divergence can highlight regions significant to the genotype–phenotype model, and provides means to reliably predict their combinatorial nature from simple, meaningful quantities. This expands the capability to discuss these mappings in rigorous terms and complements the application of more sophisticated modeling methods by offering a method to expose the underlying statistical behaviors. Such mixed-approach analyses are crucial for converting large-scale data sets into specific biochemical knowledge.

In this work, we demonstrated a simple quantity that can be calculated from a large but limited bulk of sequence-activity data to produce a probabilistic representation of the ribozyme fitness landscape. This representation explicitly captures the degree to which a given sequence site possesses epistatic interactions with other sites, enabling precise exploitation of these differing forms of interaction.

Contemporary machine learning efforts frequently rely on the application of "shallow learners" (Bengio and Lecun 2007), algorithms applied directly to biochemical data with the hope that the sophistication of the algorithm

is sufficient to overcome the convolutions obscuring the sequence-activity relationship. However, the choice of representation for the input data significantly impacts not just the interpretability of the model but also the performance of the model (Bengio et al. 2013). With this in mind, the epistatic divergence introduced here is a simple transformation of the data, driven by established information theory. The results are used to develop a simple model that maximizes our extrapolation capabilities, such that we could predict and experimentally validate new points in sequence space having high activity. We demonstrated that epistatic divergence is a sufficient representation to create experimentally relevant extrapolative models using a simple analysis workflow. Future integration of epistatic divergence with sophisticated machine learning algorithms (e.g., Shroff et al. 2020) may further improve predictive models of fitness landscapes.

MATERIALS AND METHODS

Construction of epistatic divergence

We construct epistatic divergence in a similar manner to prior work (Ostman et al. 2012) to compare the degree to which a pair of nucleotide identities affects the activity state versus the degree to which an individual constituent site affects the activity state. The motivation is that a more epistatic nucleotide pair requires the knowledge of both nucleotides jointly to describe the activity state likelihood more accurately, while a less epistatic pair would allow for that description from the individual descriptions of each nucleotide identity. We describe the epistatic divergence using information content (I), or surprisal (Shannon 1948). Formally, $I(p(x))$ is the information content of event x with probability $p(x)$, where

$$I(p(x)) = -\log(p(x)). \tag{1}$$

Epistatic divergence is assessed as

$$\Delta I_{A,m,n} = I(p(A|m)) + I(p(A|n)) - I(p(A|m,n)).$$

Here, we denote a random variable representing the activity of the ribozyme with A. Lower case m and n specify a genotype at sites M and N. We term a pair of sites relevant to this calculation as a "site pair." Thus, $p(A|m)$ denotes the probability of observing A conditioned on genotype m, and we have:

$$\begin{aligned} &I(p(A|m)) + I(p(A|n)) - I(p(A|m,n)) \\ &\quad = \log(p(A|m,n)) - \log(p(A|m)p(A|n)) \end{aligned} \tag{2}$$

$$\Delta I_{A,m,n} = \log\left(\frac{p(A|m,n)}{p(A|m)p(A|n)}\right).$$

This expresses epistatic interaction using information content. Because we are concerned with pairs of sites (i.e., "site pair," such as 29 and 38), we average over a probability distribution that describes how the various genotypes predict the phenotypes. Thus we use as our distribution $p(A|m,n)$:

$$\sum_A p(A|m,n)\log\left(\frac{p(A|m,n)}{p(A|m)p(A|n)}\right).$$

To incorporate information about every possible genotype at a site pair, we sum over all genotypes that are represented in the sample (over the support set of the population). Thus we sum over every combination of activity and genotype states (A, m, n) that has at least one representative in the sample population, as follows:

$$\begin{aligned} \epsilon_A(m,n) &= \sum_{m,n}\sum_A p(A|m,n)\log\left(\frac{p(A|m,n)}{p(A|m)p(A|n)}\right) \\ &= \sum_{\substack{\text{occupied states}\\(A,m,n)}} \frac{p(A,m,n)}{p(m,n)}\log\left(\frac{p(A,m,n)p(m)p(n)}{p(A,m)p(A,n)p(m,n)}\right). \end{aligned}$$

Thus,

$$\epsilon_A(m,n) = \sum_{\text{occupied states}} -\left(\frac{p(A,m,n)}{p(m,n)}\right)\log_b\left(\frac{p(A,m)p(A,n)p(m,n)}{p(m)p(n)p(A,m,n)}\right), \tag{3}$$

$$\epsilon_A(m,n) = D_{\mathrm{KL}}(p(A|n,m)||\,p(A|m)p(A|n)), \tag{4}$$

where D_{KL} is the Kullback–Leibler divergence and ϵ_A is the epistatic divergence. The quantity ϵ_A is equal to zero when the interaction between m and n carries no additional information (i.e., no epistasis, such that the activity of the combined genotype can be completely predicted from the activities of the individual genotypes), or when the sites do not affect activity. We set base $b = 2$, so that information is given in units of bits.

It should be noted that the use of the D_{KL} is compact notation and does not relate to the use of D_{KL} to compare two probability distributions, because $p(A|m)p(A|n)$ is not a well-defined probability distribution. Thus, the usage here only results from examining the difference of information contents between informational bodies and weighting it to favor relevance to the empirical distribution, $p(A|m,n)$.

Together, the logarithmic terms quantify the degree to which genotypes are statistically dependent within the context of a given genotype, and the weight factor then adjusts the signal such that its intensity depends on the degree to which the genotype explicitly impacts the phenotype. Detailed discussion of the epistatic divergence quantity is provided in the Supplemental Information, Appendices A–C.

Mutual information to describe the effects of single sites

To assess the single-site effect on polymer activity, we use the mutual information, in bits, $N(A;m)$, as follows:

$$N(A;m) = \sum_A\sum_m p(A,m)\log_2\left(\frac{p(A,m)}{p(A)p(m)}\right). \tag{5}$$

Mutual information is used to interrogate the divergence of the joint state distribution between activity and single-site identity, and the distribution associated with statistical independence, thus quantifying the effect of a given residue on the distribution of the activity classes. In contrast to epistatic divergence, the mutual information assesses single-site contributions rather than interactions between sites with respect to their impact on the activity class.

The activity, A, is a continuous value, which we classify into discrete classes of activity by a schema $\mathbf{A}$ with a set of associated parameters $\{\theta_i\}$, where i indexes the parameters that discretize

the activity space. We used two types of classification. The first, used for extrapolative prediction of active sequences, divides sequences into those with activity less than or greater than the activity of a central reference "seed" sequence. The second, used for visualizing the pairwise epistasis contributing fundamental activity, was found by fitting a Gaussian curve to normalized activity values. This results in a threshold of four times the baseline activity, that statistically defines whether a sequence can be considered catalytically active or not (Janzen et al. 2022). Median values from experimental replicates were used for the activity metric.

We specify the classification scheme **A**, as follows:

$$A(a) = \begin{cases} 0 & \text{if } a < \theta \text{ (inactive class)} \\ 1 & \text{if } a \geq \theta \text{ (active class)} \end{cases}, \tag{6}$$

where a is the continuous value of activity and A is the discrete class. This representation depends only on a single parameter, θ; however, alternative classification schemes could be defined depending on the needs of a given investigation.

Extrapolative prediction of active sequences

To predict regions with high-activity ribozymes, we compute ε, epistatic divergence, using the classification threshold $\theta = a_{\text{seed}}$, the activity of the seed sequence at the center of the 2-Hamming distance radius defining the training space. For other characterization, we set $\theta = a_{\text{active}}$, a value determined by fitting normal distributions to the measured background activities of noncatalytic sequences. Ribozyme activities higher than the threshold value were significantly greater than the background activity (i.e., >4 times the background rate) (Janzen et al. 2022).

The epistatic divergence values were plotted to determine which sites were most associated with improvements upon the base activity of the seed sequence. This process identified pairs of nucleotides that are associated with the most epistatic improvements to the activity. Knowledge of these pairs was then combined with insight from mutual information calculations regarding highly informative sites. This two-step process resulted in prediction of noninterfering epistatic pairs, whose genotypes could be combined to extrapolate activity in unexplored regions of sequence space.

Comparison to established measures of epistasis

We compared the information produced by epistatic divergence with two conventional measures, derived from the additive formulation of epistasis (Phillips 2008). We use two quantities, termed μ and σ, which are related to the activity of sequences as follows:

$$e_{mn} = \Delta_{mn} - \Delta_{m0} - \Delta_{0n}, \tag{7a}$$

$$\mu = \frac{1}{N}\sum_{m,\,n} e_{m,n}, \tag{7b}$$

$$\sigma = \text{Std}(\{e_{m,n}\}). \tag{7c}$$

In μ, the difference between the change in molecular activity, associated with a double mutant (Δ_{mn}) and the change described by summing the individual single mutations (Δ_{m0} and Δ_{0n}) are averaged over N genotypic backgrounds. σ is the standard deviation (Std) associated with that set, and describes the variance of effects across different genetic backgrounds. Note that μ and σ (not ε) are used for denotation of the measures described in Equations 7a–7c.

Experimental data set of ribozyme activities

The data consists of high-throughput (*k*-Seq) activity measurements on two families of ribozymes originally discovered by in vitro selection starting from a 21-site variable region. Ribozymes self-aminoacylate by reaction with a tyrosine analog substrate, biotinyl-Tyr(Me)-oxazolone (BYO). Two families, 1B.1 and 1A.1, were chosen for this analysis due to a shared motif. Activity measurements were performed for all sequences within Hamming distance of two from the core seed sequences, providing detailed mapping of the localized fitness landscapes. Although no evolutionary pathway had been discovered between these families, the shared motif suggested the possibility of a connection between the two active families. Data from single- and double-mutant sets of these two families were used to produce predictions of high activity within the triple- and quadruple-mutant range of interest, allowing us to explore predictive power beyond the Hamming range of the training set. Data were normalized by background activity, consistent with other work on this data set (Janzen et al. 2022).

Experimental measurement of ribozyme activity by RT-qPCR

The activities of selected ribozymes were determined by reverse transcription-qPCR assay (RT-qPCR) as previously described (Lai et al. 2021). DNA sequences were chemically synthesized and polyacrylamide gel electrophoresis (PAGE)-purified by Integrated DNA Technologies. The synthesized DNA sequences were 5′-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-N21-TTCACTGCAGACTTGACGAAGCTG-3′, where the nucleotides upstream of the transcription start site for T7 RNA polymerase are underlined and N21 denotes 21 consecutive nucleotides, which are varied for different ribozyme sequences. The sequences of the N21 region of tested ribozymes are: CCACACTTCAAGCAATCGGTC (S-1B.1-a), CCCCGCTTCAAACAATCGGTC (S1B-29C31G38A), CCCTGCTTCAAACAATCGGTC (S1B-29C30T31G38A), CTGCTTCAAACAATCGGTCTG (S1A-29G), and CTACTTCAAACAATCGGTCTG (S-1A.1-a). RNAs were transcribed using HiScribe T7 polymerase (New England Biolabs) and purified by denaturing PAGE (National Diagnostics). 0.1 µM of RNA samples in the aminoacylation buffer (100 mM HEPES [pH = 7], 100 mM NaCl, 100 mM KCl, 5 mM $MgCl_2$, and 5 mM $CaCl_2$) were incubated for 90 min with various BYO substrate concentrations (10, 50, 100, 250, 500, and 1000 µM) in the total volume of 100 µL for each sample. The reactions were stopped by removing unreacted substrate using Bio-Spin P-30 Tris desalting columns (Bio-Rad). The RNA concentration of each sample was quantified by Qubit 3.0 Fluorometer (Thermo Fisher Scientific). To isolate the reacted RNA, streptavidin MagneSphere paramagnetic beads (Promega) were added to all reacted RNA samples (20 ng RNA for each sample from the dissolved reacted RNA stock solutions) with a volume ratio of 1:1. Samples were incubated for 10 min at room temperature with end-over-end tumbling, followed by three washing steps. The aminoacylated RNAs were eluted with UltraPure DEPC-Treated Water (Invitrogen)

incubation at 70°C for 1 min. The amounts of aminoacylated RNAs were quantified using iTaq SYBR green mix (#1725150, Bio-Rad) using the Bio-Rad CFX96 Touch system. The samples were prepared following the manufacturer's protocol. An amount of 2 µL sample was mixed in the total 10 µL RT-qPCR reaction volume with 500 nM of both forward and reverse primers. The forward and reverse primers sequence were 5′-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGA-3′ and 5′-CAGCTTCGTCAAGTCTGCAGTGAA-3′, respectively. A calibration standard curve was measured for each RT-qPCR measurement batch to reduce measurement error. The standard RNA sequence was 5′-GGGAAUGGAUCCACAUCUACGAAUUCAAAAACAAAAACAAAAACAAANUUCACUGCAGACUUGACGAAGCUG-3′ which has the same length (i.e., 71 bp) and primer-complementary regions as the ribozymes used in this study. The standard curve was determined by adding 2 µL standard RNA samples with the concentrations of 1000, 100, 10, 1, and 0.1 pg/µL. Triplicates were performed for each sample. Results were fit to the pseudo-first-order rate equation

$$F = A(1 - e^{-k[\mathrm{BYO}]t}),$$

where F is the reacted fraction, A is the maximum reacted fraction, t is the incubation time of 90 min, and k is the effective rate constant of the aminoacylation reaction. The two fitting parameters A and k are poorly estimated individually for low-activity sequences (ca., $k < 0.5\,\mathrm{min}^{-1}\,\mathrm{M}^{-1}$), but due to the inverse correlation between estimated A and k during curve fitting, the product of the estimated k and estimated A is more accurate (Shen et al. 2021). Therefore, the product of the two estimated parameters, kA, from the pseudo-first-order curve fitting, was used to represent the catalytic activity of ribozymes in the present study.

DATA DEPOSITION

The Python code used in the calculations is available at https://github.com/ncharest/epistatic-divergence. The ribozyme data set is publicly available at the Dryad Digital Repository under DOI 10.25349/D92C9C (https://doi.org/10.25349/D92C9C).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

Nathaniel Charest now works as a Federal Postdoctoral Chemist for the United States Environmental Protection Agency. All work for this manuscript was completed before working for the U.S. EPA. The views expressed in this work are those of the author and do not reflect U.S. EPA opinions or policy. The authors thank B. Schindlinger for assistance with the graphical abstract. This work was supported by the Simons Collaboration on the Origin of Life (290356FY18), NASA (80NSSC21K0595), and the National Science Foundation (NSF grants 1935372, 1935087). Support from the National Science Foundation (NSF grant MCB-1716956) and the Center for Scientific Computing at the California Nanosystems Institute (NSF grant CNS-1725797) is also acknowledged. Y.-C.L. acknowledges partial research and stipend support from the National Science and Technology Council of Taiwan (grant no. 111-2113-M-005-008) and the "Innovative Center on Sustainable Negative-Carbon Resources" under The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Received December 2, 2022; accepted July 29, 2023.

REFERENCES

Athavale SS, Spicer B, Chen IA. 2014. Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Curr Opin Chem Biol* **22:** 35–39. doi:10.1016/j.cbpa.2014.09.008

Atwal GS, Kinney JB. 2016. Learning quantitative sequence–function relationships from massively parallel experiments. *J Stat Phys* **162:** 1203–1243. doi:10.1007/s10955-015-1398-3

Beck JD, Roberts JM, Kitzhaber JM, Trapp A, Serra E, Spezzano F, Hayden EJ. 2022. Predicting higher-order mutational effects in an RNA enzyme by machine learning of high-throughput experimental data. *Front Mol Biosci* **9:** 893864. doi:10.3389/fmolb.2022.893864

Bengio Y, Lecun Y. 2007. Scaling learning algorithms towards AI. In *Large-scale kernel machines* (ed. Bottou L, et al.). MIT Press.

Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* **35:** 1798–1828. doi:10.1109/TPAMI.2013.50

Breiman L. 2001. Random forests. *Mach Learn* **45:** 5–32. doi:10.1023/A:1010933404324

Chiu DK, Kolodziejczak T. 1991. Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci* **7:** 347–352. doi:10.1093/bioinformatics/7.3.347

Chumachenko NV, Novikov Y, Yarus M. 2009. Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J Am Chem Soc* **131:** 5257–5263. doi:10.1021/ja809419f

de Duve C. 1988. The second genetic code. *Nature* **333:** 117–118. doi:10.1038/333117a0

Dunn SD, Wahl LM, Gloor GB. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24:** 333–340. doi:10.1093/bioinformatics/btm604

Freyhult E, Moulton V, Gardner P. 2005. Predicting RNA structure using mutual information. *Appl Bioinform* **4:** 53–59. doi:10.2165/00822942-200504010-00006

Gloor GB, Martin LC, Wahl LM, Dunn SD. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44:** 7156–7165. doi:10.1021/bi050293e

Illangasekare M, Yarus M. 1999. Specific, rapid synthesis of Phe-RNA by RNA. *Proc Natl Acad Sci* **96:** 5470–5475. doi:10.1073/pnas.96.10.5470

Janzen E, Shen Y, Vazquez-Salazar A, Liu Z, Blanco C, Kenchel J, Chen IA. 2022. Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes. *Nat Commun* **13:** 3631. doi:10.1038/s41467-022-31387-0

Jiménez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA. 2013. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc Natl Acad Sci* **110:** 14984–14989. doi:10.1073/pnas.1307604110

Kinney JB, McCandlish DM. 2019. Massively parallel assays and quantitative sequence–function relationships. *Annu Rev Genom Hum Genet* **20:** 99–127. doi:10.1146/annurev-genom-083118-014845

Kinney JB, Tkačik G, Callan CG. 2007. Precise physical models of protein–DNA interaction from high-throughput data. *Proc Natl Acad Sci* **104:** 501–506. doi:10.1073/pnas.0609908104

Kirboga KK, Abbasi S, Kucuksille EU. 2023. Explainability and white box in drug discovery. *Chem Biol Drug Des* **102:** 217–233. doi:10.1111/cbdd.14262

Kondrashov DA, Kondrashov FA. 2015. Topological features of rugged fitness landscapes in sequence space. *Trends Genet* **31:** 24–33. doi:10.1016/j.tig.2014.09.009

Lai YC, Liu Z, Chen IA. 2021. Encapsulation of ribozymes inside model protocells leads to faster evolutionary adaptation. *Proc Natl Acad Sci* **118:** e2025054118. doi:10.1073/pnas.2025054118

Lee N, Bessho Y, Wei K, Szostak JW, Suga H. 2000. Ribozyme-catalyzed tRNA aminoacylation. *Nat Struct Biol* **7:** 28–33. doi:10.1038/71225

Liu Z, Rigger L, Rossi JC, Sutherland JD, Pascal R. 2016. Mixed anhydride intermediates in the reaction of 5(4H)-oxazolones with phosphate esters and nucleotides. *Chemistry (Easton)* **22:** 14940–14949. doi:10.1002/chem.201602697

Miton CM, Chen JZ, Ost K, Anderson DW, Tokuriki N. 2020. Statistical analysis of mutational epistasis to reveal intramolecular interaction networks in proteins. *Methods Enzymol* **643:** 243–280. doi:10.1016/bs.mie.2020.07.012

Moore JH, Hu T. 2015. Epistasis analysis using information theory. *Methods Mol Biol* **1253:** 257–268. doi:10.1007/978-1-4939-2155-3_13

Moore JH, Gilbert JC, Tsai C-T, Chiang F-T, Holden T, Barney N, White BC. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* **241:** 252–261. doi:10.1016/j.jtbi.2005.11.036

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

Ostman B, Hintze A, Adami C. 2012. Impact of epistasis and pleiotropy on evolutionary adaptation. *Proc Biol Sci* **279:** 247–256.

Otwinowski J, McCandlish DM, Plotkin JB. 2018. Inferring the shape of global epistasis. *Proc Natl Acad Sci* **115:** E7550–E7558. doi:10.1073/pnas.1804015115

Phillips PC. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9:** 855–867. doi:10.1038/nrg2452

Pitt JN, Ferré-D'Amaré AR. 2010. Rapid construction of empirical RNA fitness landscapes. *Science* **330:** 376–379. doi:10.1126/science.1192001

Pressman A, Blanco C, Chen Irene A. 2015. The RNA world as a model system to study the origin of life. *Curr Biol* **25:** R953–R963. doi:10.1016/j.cub.2015.06.016

Pressman AD, Liu Z, Janzen E, Blanco C, Müller UF, Joyce GF, Pascal R, Chen IA. 2019. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J Am Chem Soc* **141:** 6213–6223. doi:10.1021/jacs.8b13298

Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G. 2016. Network of epistatic interactions within a yeast snoRNA. *Science* **352:** 840–844. doi:10.1126/science.aaf0965

Romero PA, Krause A, Arnold FH. 2013. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci* **110:** E193–E201. doi:10.1073/pnas.1215251110

Rotrattanadumrong R, Yokobayashi Y. 2022. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. *Nat Commun* **13:** 4847. doi:10.1038/s41467-022-32538-z

Sailer ZR, Harms MJ. 2017. Detecting high-order epistasis in nonlinear genotype–phenotype maps. *Genetics* **205:** 1079–1088. doi:10.1534/genetics.116.195214

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature* **533:** 397–401. doi:10.1038/nature17995

Schmidt CM, Smolke CD. 2021. A convolutional neural network for the prediction and forward design of ribozyme-based gene-control elements. *Elife* **10:** e59697. doi:10.7554/eLife.59697

Shannon CE. 1948. A mathematical theory of communication. *Bell Syst Tech J* **27:** 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shen Y, Pressman A, Janzen E, Chen IA. 2021. Kinetic sequencing (k-seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Res* **49:** e67. doi:10.1093/nar/gkab199

Shroff R, Cole AW, Diaz DJ, Morrow BR, Donnell I, Annapareddy A, Gollihar J, Ellington AD, Thyer R. 2020. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth Biol* **9:** 2927–2935. doi:10.1021/acssynbio.0c00345

Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci* **25:** 1204–1218. doi:10.1002/pro.2897

Yang KK, Wu Z, Arnold FH. 2019. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16:** 687–694. doi:10.1038/s41592-019-0496-6

Yokobayashi Y. 2020. High-throughput analysis and engineering of ribozymes and deoxyribozymes by sequencing. *Acc Chem Res* **53:** 2903–2912. doi:10.1021/acs.accounts.0c00546

Zhou J, McCandlish DM. 2020. Minimum epistasis interpolation for sequence–function relationships. *Nat Commun* **11:** 1782. doi:10.1038/s41467-020-15512-5

Zhou J, Wong MS, Chen WC, Krainer AR, Kinney JB, McCandlish DM. 2022. Higher-order epistasis and phenotypic prediction. *Proc Natl Acad Sci* **119:** e2204233119. doi:10.1073/pnas.2204233119

See the following page for ***Meet the First Author***

MEET THE FIRST AUTHOR

Nathaniel Charest

***Meet the First Author(s)* is an editorial feature within *RNA*, in which the first author(s) of research-based papers in each issue have the opportunity to introduce themselves and their work to readers of *RNA* and the RNA research community. Nathaniel Charest is the first author of this paper, "Discovering pathways through ribozyme fitness landscapes using information theoretic quantification of epistasis." At the time of this work, he was a postdoctoral scholar with Joan-Emma Shea at the University of California Santa Barbara. The main focus of his research is informatics and using fundamental theories of machine learning to extract and interpret patterns within chemical and biochemical data. The emphasis of this particular research is to demonstrate that first-principles theory has direct implications on analyzing modern data problems without relying on black box regression models or elaborate interpretation schemes.**

What are the major results described in your paper and how do they impact this branch of the field?

The major results of this paper show that simple probability theory can derive exploitable patterns within RNA sequence data and pragmatically predict activity within unexplored sequence space. The impact on this branch of the field is to assist in the characterization of RNA sequence landscapes in a manner that does not obscure the data behind intricate optimizations or stochastic regression models, and to provide first-principles tools for theoretically capturing the relationships between primary sequence and biochemical activities.

What led you to study RNA or this aspect of RNA science?

RNA is a natural candidate for informatics tools and approaches. The relative combinatoric straightforwardness of the primary sequence results in an enormous but foundationally simple sequence space that is amenable to the sort of first-principles analysis that drives modern machine-learning approaches. My coauthors in Dr. Chen's lab provided excellent data for work-up, and so collaboration on the subject of RNA sequence landscapes was natural.

During the course of these experiments, were there any surprising results or particular difficulties that altered your thinking and subsequent focus?

The benefit of first-principles approaches is that they are remarkably reliable in terms of prediction and result, and "debugging" where things become unexpected is a relatively transparent process. I was pleasantly surprised that some equations writable on a single sheet of paper resulted in robust predictions of heightened activity within the sequence space. Ultimately, my thinking was guided by a desire to test the probability concepts laid out in Claude Shannon's seminal writings on information theory as a proof-of-concept that these older approaches to informatics might still yield valuable insights in an era dominated by elaborate regression algorithms.

What are some of the landmark moments that provoked your interest in science or your development as a scientist?

This answer could easily extend back to my mother first purchasing a 10-yr-old me a cheap microscope to look at mold cells; however, my current interests have been primarily guided by encountering the writings of Geoffrey Hinton, Yoshua Bengio, and Yann Lecun. They demystified algorithms I had been taught were opaque by applying elegant and comprehensive theory, resulting in my fascination with reconciling the old first-principles paradigm of informatics with modern technologies.

If you were able to give one piece of advice to your younger self, what would that be?

I would urge myself to respect the value of taking scientific theory into praxis and to remember that theory is not particularly "real" to a vast majority of stakeholders. Learning communication skills to bridge that gap earlier would have been expedient to my skill as a professional scientist.

Are there specific individuals or groups who have influenced your philosophy or approach to science?

My major high-profile influences are Hinton, Leo Breiman, Claude Shannon, and Bengio. They shaped my philosophy of striving to maintain some human understanding of our data approaches and algorithms. I have also found Hinton's recent words on artificial intelligence to be highly impactful on my general balancing of technological optimism and caution.

What are your subsequent near- or long-term career plans?

My short-term goals are to continue developing my expertise and the practical skills relevant to proficiency with modern software practices and computational research. My interests are broad enough that I can foresee myself happy with a number of trajectories, but all long-term plans involve remaining close with informatics technology and research.