Structural Dynamics ARTICLE pubs.aip.org/aip/sdy

Aminoacyl-tRNA synthetase urzymes optimized by deep learning behave as a quasispecies

Cite as: Struct. Dyn. **12**, 024701 (2025); doi: 10.1063/4.0000294
Submitted: 29 January 2025 · Accepted: 19 March 2025 ·
Published Online: 25 April 2025

Sourav Kumar Patra,[1] Nicholas Randolph,[1] Brian Kuhlman,[1,2] Henry Dieckhaus,[3,4] Laurie Betts,[1] Jordan Douglas,[5] Peter R. Wills,[5] and Charles W. Carter, Jr.[1,2,a)]

AFFILIATIONS

[1]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7260, USA
[2]Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, USA
[3]Division of Chemical Biology and Medicinal Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7355, USA
[4]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27517, USA
[5]Department of Physics, University of Auckland, Auckland, New Zealand

Note: Paper published as part of the special topic on Artificial Intelligence and Structural Science.
[a)]**Author to whom correspondence should be addressed:** carter@med.unc.edu

ABSTRACT

Protein design plays a key role in our efforts to work out how genetic coding began. That effort entails urzymes. Urzymes are small, conserved excerpts from full-length aminoacyl-tRNA synthetases that remain active. Urzymes require design to connect disjoint pieces and repair naked nonpolar patches created by removing large domains. Rosetta allowed us to create the first urzymes, but those urzymes were only sparingly soluble. We could measure activity, but it was hard to concentrate those samples to levels required for structural biology. Here, we used the deep learning algorithms ProteinMPNN and AlphaFold2 to redesign a set of optimized LeuAC urzymes derived from leucyl-tRNA synthetase. We select a balanced, representative subset of eight variants for testing using principal component analysis. Most tested variants are much more soluble than the original LeuAC. They also span a range of catalytic proficiency and amino acid specificity. The data enable detailed statistical analyses of the sources of both solubility and specificity. In that way, we show how to begin to unwrap the elements of protein chemistry that were hidden within the neural networks. Deep learning networks have thus helped us surmount several vexing obstacles to further investigations into the nature of ancestral proteins. Finally, we discuss how the eight variants might resemble a sample drawn from a population similar to one subject to natural selection.

INTRODUCTION

Deep learning platforms[1–4] were recognized by Nobel Prizes in 2024 for their contributions to physics[5] and chemistry.[6] Both have made major contributions to our effort to construct experimental models for ancestral aminoacyl-tRNA synthetases (AARS). These are the enzymes that translate the genetic code.[7] They act as AND gates, selecting amino acid and tRNA substrates from the cellular milieu, and make a covalent bond between them if, and only if, both satisfy the relationships set out in the genetic coding table.

An extensive literature describes the nested hierarchy of AARS constructs we made as model systems for ancestral AARS. "Protozymes"[8,9] are ∼50-residue peptides containing the ATP binding sites. "Urzymes"[10–15] have 120–130 residues. Recently, we described the *in vivo* genesis of both urzymes and protozymes from a full-length double mutant plasmid. That work[16] confirmed a significant literature on both kinds of analytical constructs. They have nearly intact catalytic sites and generally retain ∼60% of the transition state stabilization free energy of the mature AARS. The constructs thus may represent active models for very early stages of AARS evolution.[17–21] Their primitive enzymology likely played a central role in enabling nature to discover an optimal code and then teach the earliest AARS to interpret the blueprints in their own genes.[22,23]

However, urzymes are among the hardest biological objects to study. Biochemistry is difficult because urzymes have limited

stability and their activities are scarcely above background. Structural biology is next to impossible because of their low solubility and disorder.

We turned to deep learning algorithms with several goals in mind. First, we hoped to find sequences that had higher solubility when expressed. Second, we hoped to increase the stability and perhaps the 3D order of AARS urzymes. Third, we expected to test drive procedures for efficient biochemical assays of AARS urzymes with closely related sequences. Finally, we expected that by balanced sampling we could exploit the implicit factorial nature of the subset to answer questions such as these. How do the sequence changes affect solubility of the related variants? How do they affect catalysis and substrate specificity? And, eventually, how do they affect the thermodynamic stability?

Moreover, ancient proteins coded before translation was as accurate as it is today were doubtless related proteins without a single defined sequence, but which all had similar biochemical and structural properties. They are ensembles in a similar sense to that used in statistical mechanics. Eigen called them quasispecies (QS).[24] We will also refer to the variants as representing such an ensemble.

Eigen defined the term quasispecies to describe the impact of the high error rates expected from primitive replicators.[25] The term identifies a (stationary) distribution of nucleic acid sequences that replicates (breeds true) at selection equilibrium. Such a "species" is not genetically singular, so a *quasi*-species. Eigen derived rate equations to describe the time evolution of the population distribution (from any arbitrary initial condition). The solution of those equations is a set of population eigenvectors. The selected "quasispecies" (QS) distribution is the eigenvector with the highest eigenvalue. We are in the process of creating the necessary algorithms to reconstruct the likely sequence distributions of ancestral AARS from the time before the genetic coding table was complete.[17,22,26,27] That work will then require new kinds of biochemical analyses. For these reasons, any work that makes working with urzymes easier and/or provides a platform for studying quasispecies-like ensembles is valuable.

The accompanying papers in this collection focus mainly on what artificial intelligence contributes to structure solution. The experiences described here emphasize how deep learning also brings important new resources to downstream experimental work that is based on structural studies. Our purpose is to describe how deep learning helped us to address issues raised in the foregoing paragraphs.

METHODS AND MATERIALS

Cloning, overexpression, and purification of redesigned LeuAC variants

Plasmids pMAL-c2X containing the redesigned maltose binding protein (MBP)-LeuAC genes were synthesized by Twist Bioscience and expressed in BL21 (DE3)pLysS competent cells (Promega). Cells were grown at 37 °C and early log phase cells induced with 300 μM IPTG overnight. Harvested cells were resuspended in a buffer containing 20 mM Tris, pH 7.4, 1 mM EDTA, 5 mM β-ME, 17.5% glycerol, 0.1% NP40, 33 mM $(NH_4)_2SO_4$, 1.25% glycine, 300 mM guanidine hydrochloride plus cOmplete protease inhibitor (Roche). The cell suspension was lysed using a glass homogenizer followed by sonication (with eight pulses of 10 s with 70% amplitude sonic vibrations) keeping 20 s of pause time, ensuring the tube remained in ice during sonication. LeuAC crude extract was then pelleted at 4 °C with centrifugation at 15 K rpm for 30 min to remove insoluble material. The extract supernatant was then diluted 1:4 with lysis buffer and loaded onto equilibrated Amylose FF resin (Cytiva). The resin was washed with five column volumes of buffer and the protein was eluted with 30 mM maltose in optimal buffer. The purified fraction was then dialyzed overnight with 50 mM HEPES buffer containing 1 mM EDTA, 5 mM β-mercaptoethanol, 17.5% glycerol. After dialysis, fractions containing protein were pooled together, concentrated, and mixed with 50% glycerol and stored at 80 °C. Protein concentrations were determined using the Pierce™ Detergent-Compatible Bradford Assay Kit (Thermo Scientific).[11] The purification of all redesigned LeuAC variants was done using these same methods on separate days ensuring proper wash of FPLC amylose resin columns and systems to avoid cross contamination of proteins.

Single turnover active-site titration assay of all redesigned urzymes

To identify the active fraction of purified LeuAC proteins, active-site titration (AST) assays were performed using the same principle as described in Refs. 28 and 29 with slight modifications. Briefly, 5 μM of LeuAC protein was added to a reaction mix containing 50 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 50 mM leucine, 1 mM DTT, 9 μM adenosine triphosphate (ATP), 0.005 units of inorganic pyrophosphatase, and ~5000 cpm α-labeled [^{32}P] ATP to start the reaction. A volume of 2 μl for all representative timepoints was added from the ongoing reaction mix into separate tubes containing 4 μl quenching buffer containing 0.4 M sodium acetate and 0.1% sodium dodecyl sulfate (SDS) and kept on ice until all timepoints had been collected. Around 2 μl of quenched samples were spotted on pre-run (PEI) thin-layer chromatography (TLC) plates. These were developed in TLC running buffer containing 850 mM Tris pH 8.0. The plate was then dried and exposed for varying amounts of time to a phosphor image screen and visualized with a Typhoon Scanner (Cytiva). The intensities of each nucleotide were quantified by densitometry scanning using measure functions of ImageJ.[30] The time dependence of loss (ATP) or de novo appearance (ADP, AMP) of the three adenine nucleotides phosphates was fitted using the nonlinear regression module of JMP™ Pro to the following equation:

$$\text{Product(calc)} = \text{A} \times \exp - \text{k}_{\text{chem}} \times \text{seconds} - \text{k}_3 \times \text{seconds} + \text{C}, \quad (1)$$

where k_{chem} is the first-order rate constant, k_3 is the turnover rate, A is the amplitude of the first-order process, and C is an offset.

Michaelis–Menten kinetics of amino acid activation by various LeuAC

Michaelis–Menten kinetics of amino acid activation assays for individual amino acids were done according to Ref. 31 with some slight modifications. In brief, the assay reaction mixture contained 50 mM HEPES of pH 7.0, 20 mM $MgCl_2$, 50 mM KCl, 0.2 mM ATP, 0.1 unit of inorganic pyrophosphatase [NEB], 200 nM of different redesigned LeuAC urzymes, and varied concentrations of a particular amino acid in separate tubes in a total reaction volume of 100 μl. Reactions were carried out at 37 °C for less than 1 h along with an enzyme blank in a separate tube. After this, 400 μl of MG-ammonium

molybdate solution was added to the reaction tubes and kept for 5 min after mixing properly to develop the phosphomolybdate complex. Around 40 μl of sodium citrate (w/v) was then added to the tubes, then solutions were allowed to stand for 20 min and the optical absorbance at 620 nm was measured with DU800 spectrophotometer (Beckman Coulter Inc., Brea, CA, USA). The MG-ammonium molybdate solution was prepared as described.[9] The phosphoric acid concentration was calculated from a calibration curve prepared by using 0–250 μM K_2HPO_4, to determine the relationship between phosphoric acid concentration and absorbance. The specific activities of the LeuACs for individual amino acid concentrations were calculated and plotted in a specialized nonlinear fit using the modified Michaelis–Menten equation introduced by Johnson[32] [Eq. (2)] and implemented in JMP^{TM} Pro software. The K_M and k_{cat} of LeuACs for individual amino acids were determined, with standard deviations, from the maximum likelihood fit,

$$V_0 = K_{sp}[S]/\left(1 + K_{sp}/k_{cat}\right), \quad (2)$$

where $K_{sp} = k_{cat}/K_{M0}$.

Study of relative solubility

To study the relative solubility of the purified redesigned LeuACs (MBP-LeuACs), the pure protein lysates were treated with TEV proteases and compared on reducing SDS-PAGE using two simultaneous sets of prepared TEV protease treated samples, one of which was centrifuged and the other was non-centrifuged. In brief, purified MBP-LeuAC protein(s) was mixed with more than adequate amount of TEV protease (NEB) and TEV buffer into a microcentrifuge tube in such a way that the final concentration of proteins (LeuACs) and TEV protease became 4–5 μg/μl and 0.4 U/μl, respectively. This mixture was then incubated at 30 °C for 2 h or overnight at 4 °C to ensure complete cleavage of MBP-LeuACs fusion protein.[33] After that, samples were distributed equally into two different tubes. One was subjected to centrifugation at 14 K rpm, 4 °C for 20 min. The other sample was kept uncentrifuged in ice. The supernatant was separated from the pellet after centrifugation. This step ensured the removal of insoluble cleaved LeuACs, keeping the uncentrifuged fraction as control for this experiment. Then, three protein samples for SDS-PAGE were prepared from the uncentrifuged (soluble and insoluble content), centrifuged (soluble fraction), and pellet fraction using reducing protein loading dyes having SDS followed by SDS-PAGE analysis using 4%–20% Mini-PROTEAN® TGX^{TM} Precast Protein Gels of BioRAD.[34] After electrophoresis, the gel was subjected to Coomassie staining followed by destaining to visualize protein bands. The soluble fraction was then determined by doing densitometric measurements of cleaved LeuAC bands and comparing the centrifuged set with the uncentrifuged ones of corresponding LeuAC protein wells.

Computational analysis of LeuAC variant stability

The predicted folding stability (ΔG) for LeuAC variants in the apo state was calculated using ESM-IF, a deep neural network trained for inverse folding[35] using a previously reported procedure[36] to convert from conditional likelihoods to ΔG units of kcal/mol. All calculations used modeled structures from AlphaFold2.[4] Higher ΔG values indicate greater predicted folding stability. To evaluate the effect of ligand binding, we used all-atom sequence design model LigandMPNN[37] to score LeuAC variants with and without ligand present using autoregressive decoding. Higher LigandMPNN scores indicate more favorable sequences, i.e., sequences predicted to be more likely to adopt the modeled fold. All reported values represent mean and standard deviation values for triplicate runs. For ESM-IF, 0.05 Å backbone noise was added to estimate model variance between replicates. LeuAC was analyzed by *in silico* site-saturation mutagenesis using the Colab server for structure-based graph neural network ThermoMPNN,[38] which predicts a Δ(ΔG) (in kcal/mol) for each possible single mutation to the wild type protein. More positive Δ(ΔG) indicates reduced stability relative to the wild type, while more negative Δ(ΔG) indicates improved stability.

Data processing and statistical analysis by multiple regression methods

Densitometry of phosphor imaging screens of TLC plates was done using ImageJ.[30] Data were transferred to JMP16PRO^{TM} Pro 16 via Microsoft Excel (version 16.49), after intermediate calculations. We fitted active-site titration curves to Eq. (1) using the JMP16PRO^{TM} nonlinear fitting module. R^2 values were all >0.97 and most were >0.99.

Rate data from Michaelis–Menten experiments were fitted to Eq. (2) using the nonlinear module of JMP16Pro^{TM}. Factorial matrices with dependent and independent variables were processed using the Fit Model multiple regression analysis module of JMP16PRO^{TM} Pro, using an appropriate form of the following equation:[39]

$$Y_{obs} = \beta_0 + \Sigma\beta_i{}^*P_i + \Sigma\beta_{ij}{}^*P_i{}^*P_j + \varepsilon, \quad (3)$$

where Y_{obs} is a dependent variable, usually an experimental observation, β_0 is a constant derived from the average value of Y_{obs}, β_i and β_{ij} are coefficients to be fitted, $P_{i,j}$ are independent predictor variables from the factorial matrix, and ε is a residual to be minimized. All rates were converted to free energies of activation, $\Delta G^{\ddagger} = -RT\ln(k)$, before regression analysis because free energies are additive, whereas rates are multiplicative. For example, the activation free energy for the first-order decay rate in single turnover experiments is $\Delta G^{\ddagger}(k_{chem})$.

Multiple regression analyses of factorial experimental data exploit the replication inherent in the full collection of experiments to estimate experimental variances on the basis of t-test P-values, in contrast to the presenting error bars showing the variance of individual datapoints. Multiple regression analyses reported here also entail triplet experimental replicates, which enhance the associated analysis of variance.

RESULTS

There are two distinct AARS Classes: I and II.[40,41] They have entirely unrelated primary, secondary, and tertiary structures.[42] Class I AARS consist of four distinct modules: denoted A, B, C, and D. Class I urzymes retain modules A and C from the full-length enzymes and are missing connecting peptide 1 (CP1), which is module B, and the anticodon-binding domain, module D. We add the letters AC to the AARS abbreviation to denote such urzymes. The first such construct in 2007, TrpAC, relied heavily on Rosetta.[43] The Class I active sites are interrupted by the long, variable insertion element B. That insertion had to be deleted and the resulting gap sealed. In addition, both the insertion (B) and the C-terminal anticodon-binding domain (D) provide nonpolar surfaces that combine with the external surfaces of the

catalytic cores to form the hydrophobic core. Thus, full-length leucyl-tRNA synthetase (LeuRS) evolved to fold using many of the residues that end up on the urzyme surface as part of its core. Rosetta offered workable solutions to both problems, as it also did for our second Class I urzyme, LeuAC.[11]

Although TrpAC and LeuAC served as models for the ancient enzymology, structural questions remained difficult to answer. TrpAC was sufficiently soluble that we could record an NMR (nuclear magnetic resonance) HSQC (heteronuclear single quantum coherence) spectrum.[44] The dispersion of proton peaks in an HSQC spectrum is broad if a polypeptide has a nearly unique conformation, but is quite narrow if it is a molten globule. These spectra are thus a useful test for proper folding of a polypeptide. The TrpRS HSQC spectrum has a quite narrow dispersion. Together with related biophysical measurements, that suggested that it is molten-globular catalyst. LeuAC gave similar results. However, LeuAC was sufficiently soluble only in dimethyl sulfoxide (DMSO). To overcome the low solubility, we turned to the enhanced design program ProteinMPNN.[2] Our goal was to overcome the solubility issues that likely result from cutting the urzyme core out of a larger protein.

A second goal was to create protocols for characterizing ensembles. Characterization of a sample of the redesigned variants opened a deeper investigation into how these primordial enzymes might have operated. Ensembles are of special significance to us. AARS urzymes are highly promiscuous in their amino acid specificity. The primordial coding tables would therefore have been enforced with unknown but low fidelity. The early AARS genes thus made populations of urzymes that likely did not have unique amino acid sequences. They certainly had a range of specificities. Such populations have been described as quasispecies.[45] We imagine that the properties of ancestral AARS ensembles were thus highly heterogeneous in both how proficient they were and their substrate specificities. This preliminary report compares the properties—soluble fractions, single turnover kinetics, and amino acid activation with different amino acids—representative of such a population..

We describe procedures for selecting a representative sample of eight from a set of thirty MPNN designs, their purification, and their physical properties. Our most important conclusions are these: (i) The new designs are all active. (ii) Most are much more soluble than LeuAC. (iii) They have diverse catalytic and specificity properties. That inherent biochemical diversity reinforces the need to study urzymes as quasispecies populations, rather than individuals.

We purified the maltose binding protein (MBP) fusion proteins by affinity chromatography on amylose. A typical gel of the elution profile [Fig. 1(a)] shows that the eluted fractions are ∼75% pure and that the most prominent contaminating band represents ∼4% of the total protein. Burst size measurements described below indicate that all or most of the molecules in the principal band are active aminoacyl-tRNA synthetases.

The low solubility of the original LeuAC posed a serious barrier to structural studies. TEV cleavage of that purified protein left LeuAC only sparingly soluble in aqueous solutions. Figure 1(b) shows this measurement for the LeuAC20 variant, which is the third least soluble of the nine variants with a soluble fraction of ∼0.5.

The original LeuAC was very soluble in DMSO.[46] So, we recorded an HSQC spectrum of 15N/13C-enrinched LeuAC in DMSO. We could not measure either amino acid activation of tRNA aminoacylation in DMSO, so although the HSQC spectrum resembled that of a molten globule, we could not conclude that was the active species.

Further attempts to use Rosetta to enhance solubility met with limited success. At that time, ProteinMPNN was under development.[2,37] That offered a substantially enhanced search of sequence space. So we began to consider new LeuAC variants based on the new deep learning paradigm.

Selecting a representative sample from a larger number of designs

Protein chemistry entails a substantial investment of resources. We wished to ensure as diverse a characterization of the sample of variants as possible. To that end, we took the thirty initial designs and reduced these to a set of eight in two steps. First, we rejected outright three variants whose AlphaFold2 structures differed radically from the *P. horikoshii* crystal structure.[47] We assembled a set of parameters that could be computed from the sequences and their 3D structures predicted by AlphaFold2 [Fig. 2(a)]. The estimated solubility is the mean transfer free energy, ΔGw>c, from water to cyclohexane of all residues in the sequence. Structural parameters drawn from the AlphaFold2 (AF) predictions of 3D conformation include the AF confidence level, its standard deviation, the root-mean-squared deviation from the coordinates of the *P. horikoshii* LeuRS crystal structure, and the mean Simplicial

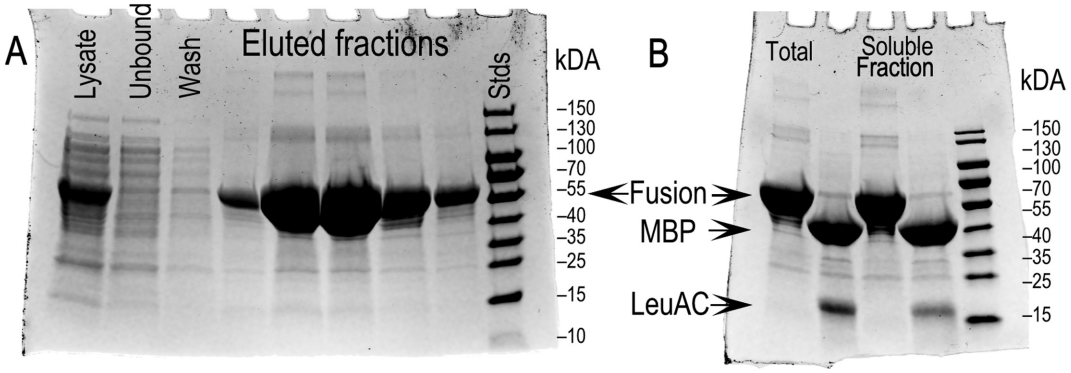
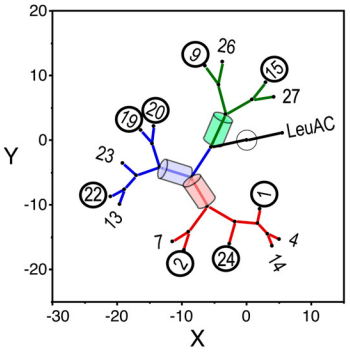


FIG. 1. Polyacrylamide gels of the LeuAC20 variant elution profile and soluble fraction. (a) Purification on amylose resin. Densitometry indicates that the eluted fractions are ∼75% pure. The most prominent contaminating band represents ∼4% of the total protein. Burst sizes in Fig. 4 are fractions of the corresponding relative purity, hence are active fractions. (b) Measurement of the soluble fraction is the ratio of the densities of the LeuAC band in the fourth lane divided by the total of lanes 2 + 4.

A

Design	Rel hydrophilicity	Stability	<SNAPP>	RMSD	Score	Seq_recov	<Afconf>	Stdev(Afconf)
1	**-0.09**	**-5.63**	**0.269**	**2.50**	**1.14**	**0.31**	**81.04**	**11.52**
2	**2.81**	**-4.84**	**0.181**	**3.20**	**1.32**	**0.30**	**72.52**	**13.88**
4	0.06	-5.39	0.209	2.68	1.30	0.26	78.75	10.53
7	0.33	-5.25	0.144	3.10	1.22	0.24	76.41	14.44
9	**1.39**	**-3.92**	**0.105**	**12.78**	**1.47**	**0.28**	**52.76**	**13.78**
13	4.36	-3.88	0.184	4.43	1.03	0.34	73.85	11.24
14	0.28	-5.29	0.217	2.98	1.07	0.30	80.59	11.51
15	**0.00**	**-5.07**	**0.110**	**10.66**	**1.01**	**0.29**	**62.91**	**11.55**
16	0.51	-4.76	0.165	4.71	1.05	0.31	71.85	15.39
19	1.34	-4.48	0.088	2.66	1.07	0.33	79.47	10.98
20	**2.08**	**-4.96**	**0.159**	**2.76**	**1.14**	**0.33**	**77.42**	**11.91**
22	**4.60**	**-3.76**	**0.169**	**2.38**	**1.21**	**0.33**	**80.65**	**10.98**
23	8.36	-4.16	0.180	6.78	1.11	0.33	65.11	10.74
24	**-2.45**	**-4.21**	**0.157**	**2.29**	**1.18**	**0.28**	**83.05**	**10.84**
26	4.37	-2.80	0.319	8.98	1.12	0.25	52.77	13.57
27	4.18	-3.52	0.149	11.26	1.14	0.31	46.30	13.20
LeuAC	**0.00**	**-3.12**	**0.1260**	**19.46**	**2.30**	**1.00**	**90.52**	**7.09**

B



C

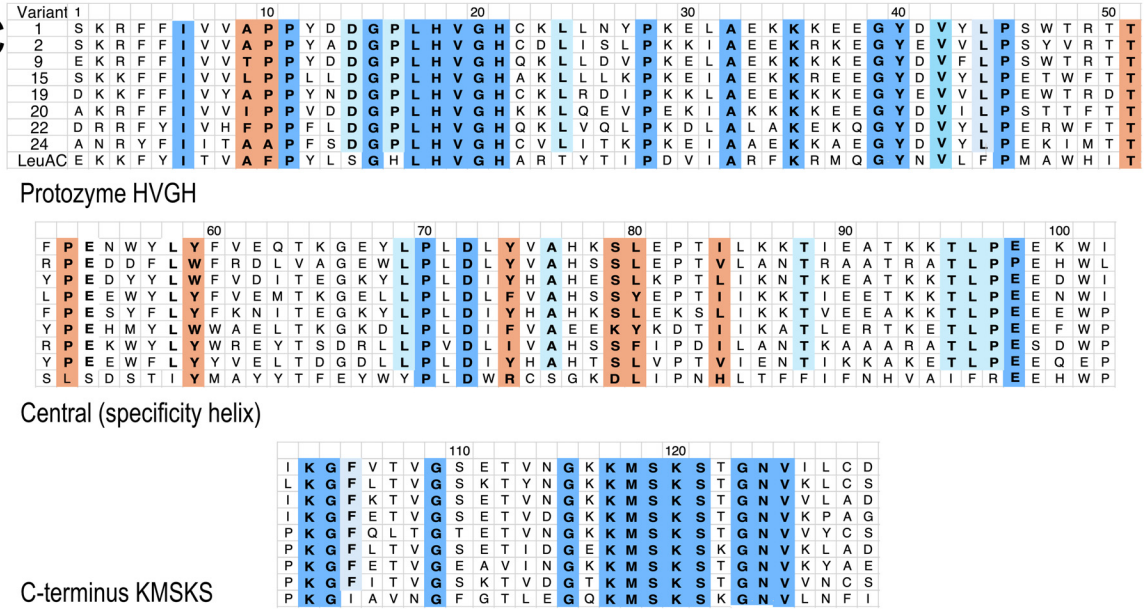


FIG. 2. Use of principal components to select a representative subset for expression and detailed characterization. (a) Matrix of computational scores for the 27 of the 30 original variants for which AlphaFold2 predictions matched the LeuAC structure. (b) Constellation plot of the hierarchical clustering of the same variants using the top five principal components derived for the eight columns of the matrix in A. Circles denote the sequences selected for expression. Their amino acid sequences are aligned in (c). Blue, red, and green segments of the plot allow for informal tests of the randomness of those samples. (c) Sequences of the eight LeuAC variants described herein. Bold face residues are invariant among the eight variants and cannot be related to differences in their properties. Blue backgrounds denote residues identical to that of the original LeuAC. Brown backgrounds denote residues close to the amino acid substrate in 3D structures predicted by AlphaFold2.[4]

Neighborhood Analysis of Protein Packing (SNAPP) value. SNAPP is a statistical potential derived from the likelihood of observing Delaunay simplices with identical compositions in the AlphaFold2 structures in the database of known structures[51] and metrics based on the sequences themselves. The SNAPP value thus estimates the contribution of a packing motif to overall stability.

We performed principal component analysis on the covariations between those eight columns, and selected a random but balanced set of eight sequences [Fig. 2(b)]. We then purified and characterized the eight variants highlighted in Figs. 3(a) and 3(b). Sequences of these variants are shown in Fig. 2(c). In the figure, invariant residues throughout the sampled set are highlighted in blue. Side chains that impinge on the active site in AlphaFold2 predicted structures are highlighted in brown.

As noted further in the discussion, this process resulted in the broadest possible sampling of the original set of variants. The results described in the following sections suggest that the sampled variants represent the variations possible with sequences derived from current deep learning algorithms.

The new variants span a wide range of solubilities

Our original goal in redesigning LeuAC was to enhance solubility for the purpose of structural studies. Solubility is a difficult property to

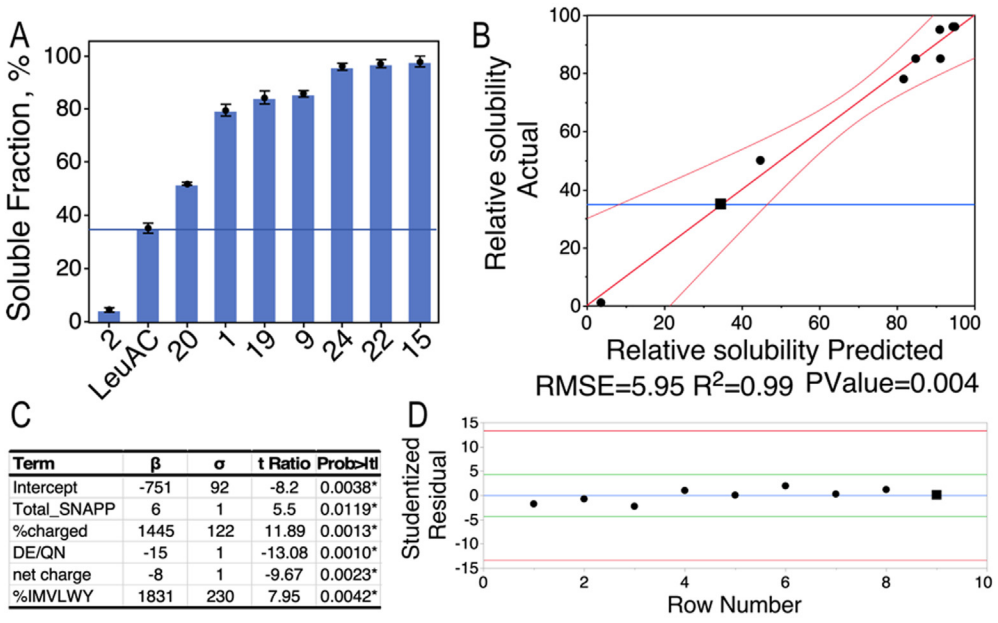


FIG. 3. Enhanced solubility of LeuAC variants. (a) Soluble fractions following Tobacco Etch Virus protease (TEV) cleavage of the MBP solubility tag. For reference, the horizontal blue line indicates the value for the original LeuAC. Bar plot is sorted in increasing value. (b)–(d) Regression model relating the observed soluble fractions for the nine variants to independent compositional parameters in Table I. (b) Shows the linearity of observed vs calculated values. The large square denotes the original LeuAC. (c) Contains the model coefficients. The % charged, net charge, DE/QN ratio [(ASP + Glu)/(GLN + ASN)], and %IMVLWY are compositional parameters derived directly from the sequence. The "Total SNAPP" score is a likelihood potential derived from the composition of all Delaunay simplices in the convex hull.[49] **β** and **σ** are the coefficients and their standard deviations. (d) Shows the studentized residual values. Data points outside the red lines would be considered outliers.

measure accurately.[48] To do so, one must follow the asymptotic approach to equilibrium with some solid state from both below and above the equilibrium soluble concentration. Rather than investing resources in such studies, we compared the soluble fraction of the same concentration of each variant before and after cleavage of the maltose binding protein (MBP) tag. The factorial matrix in Table I shows that the LeuAC variants have a wide range of solubilities by that metric [Fig. 3(a)]. LeuCA2 is only sparingly soluble without the MBP tag. All other variants, including the original LeuAC, are more soluble. Six of the variants have soluble fractions approaching 1. Soluble fractions depend on the initial concentrations of the purified variants. Thus, values close to 1 may underestimate the true solubility. That being the case, most of the more soluble variants could facilitate future NMR structural studies.

The coherence of the nine observations affords the opportunity to investigate relationships between composition and solubility. Accurate solubility prediction from compositional parameters is an ongoing challenge. For reference, expected solubilities according to one of the popular algorithms, DeepSoluE,[50] are essentially uncorrelated with the observed values ($R^2 = 0.19$, $P = 0.23$). Our model, on the other hand [Figs. 3(b) and 3(c)], provides excellent correlation ($R^2 = 0.99$, $P = 0.004$) with three degrees of freedom for estimation of errors. We compare predictions of solubility in more detail in the

TABLE I. Design matrix for soluble fraction. The dependent variable column is shaded light gray. Other columns are potential predictors.

Variant	Sol. fract	SNAPP_tot	⟨SNAPP⟩	% IVWREL	% Class I	% charged	% hydroxylic	DE/QN	net chg	% IMVLWY	% RE	DeepsoluE	% < 0
LeuAC1	78	−4.12	−0.27	0.39	0.44	0.29	0.13	10.00	−2	0.31	0.13	0.30	0.18
LeuAC2	1	−7.11	−0.25	0.36	0.43	0.25	0.14	4.67	4	0.29	0.12	0.24	0.31
LeuAC9	85	−11.80	−0.29	0.33	0.40	0.31	0.13	5.00	0	0.29	0.10	0.74	0.87
LeuAC15	95	−10.56	−0.26	0.35	0.44	0.28	0.13	3.40	2	0.31	0.11	0.42	0.74
LeuAC19	85	−6.75	−0.22	0.33	0.42	0.32	0.12	3.80	3	0.27	0.12	0.76	0.33
LeuAC20	50	−8.80	−0.24	0.36	0.42	0.33	0.12	11.00	−1	0.29	0.12	0.70	0.43
LeuAC22	96	−5.66	−0.19	0.36	0.44	0.26	0.10	2.67	2	0.30	0.12	0.25	0.18
LeuAC24	96	−6.55	−0.20	0.33	0.43	0.24	0.15	3.60	−5	0.30	0.10	0.65	0.32
LeuAC	35	−4.03	−0.16	0.31	0.43	0.17	0.12	1.13	4	0.33	0.08	0.04	0.25

section titled Discussion. We now describe the enzymology of the eight variants in comparison with the initial LeuAC.

Single turnover kinetics reveal precise differences in catalytic properties between variants

Our previous work with AARS urzymes made substantial use of active-site titration.[28,29] Single turnover assays build the experimental signal by using enough enzyme that one round of catalysis can be measured. We first used the burst sizes to show that essentially all protein molecules in solution contribute to the observed changes in adenine nucleotides. Subsequently, we showed that first-order rates, k_{chem}, were useful in interpreting variation in the catalytic properties of active-site mutations.[51,52] Very recently, we recognized that the ratio of AMP to ADP burst sizes in single turnover experiments measures the efficiency of using the free energy of ATP hydrolysis.[31] For these reasons, we first carried out single turnover assays on all variants. Figure 4 summarizes the various metrics derived from those measurements.

Overall, the redesigned variants have improved catalytic properties. We note at the outset that triplicate measurements are highly reproducible. That lends credibility to our interpretations of Fig. 4. Consistent with our earlier conclusion that the sampling protocol is representative, the experimental properties of the variants include some that are less favorable and many that are more favorable than the initial LeuAC. The dynamic range is roughly a kcal/mol in transition state stabilization free energy.

The first-order rate, k_{chem}, is the rate at which the first molecule of ATP is hydrolyzed. LeuAC19 and the original LeuAC have about the same rate; all other variants are faster. Since we began to use α-labeled ATP, it became clear that that rate is a composite rate for two reactions. One is the synthesis of aminoacyl-5′AMP, the normal product. The other reaction produces ADP instead. That reaction is not fully understood. However, we have argued for two reasons that it also represents productive use of ATP. First, the presumptive active site appears to open space for binding various ATP conformations, including two that predominate in solution.[11] These are in position to phosphorylate first the α-phosphate and then the resulting β-phosphate of the bound adenylate. Subsequently, we supported that interpretation by showing that the AMP/ADP ratio in single turnover assays depends both on the urzyme construct and on the presence of cognate tRNA.[31]

LeuAC24 is unusually free of nonproductive side reactions that produce ADP [Fig. 4(f)]. Its AMP/ADP ratio, 0.21, is 3.7 standard deviations higher than the mean of all the others, including the original LeuAC. LeuAC24 also is one of the most soluble of the redesigned variants, as is LeuAC22, which has a typical AMP/ADP ratio of 0.07. The redesigned variants may enable us to pursue this curious difference with structural studies.

Steady-state kinetic measurements show significant differences in amino acid specificity

AARS serve *in vivo* to assure that the correct amino acids match properly with successive codons in genetic messages. For that reason, it is of interest to know how the different designs affect the relative

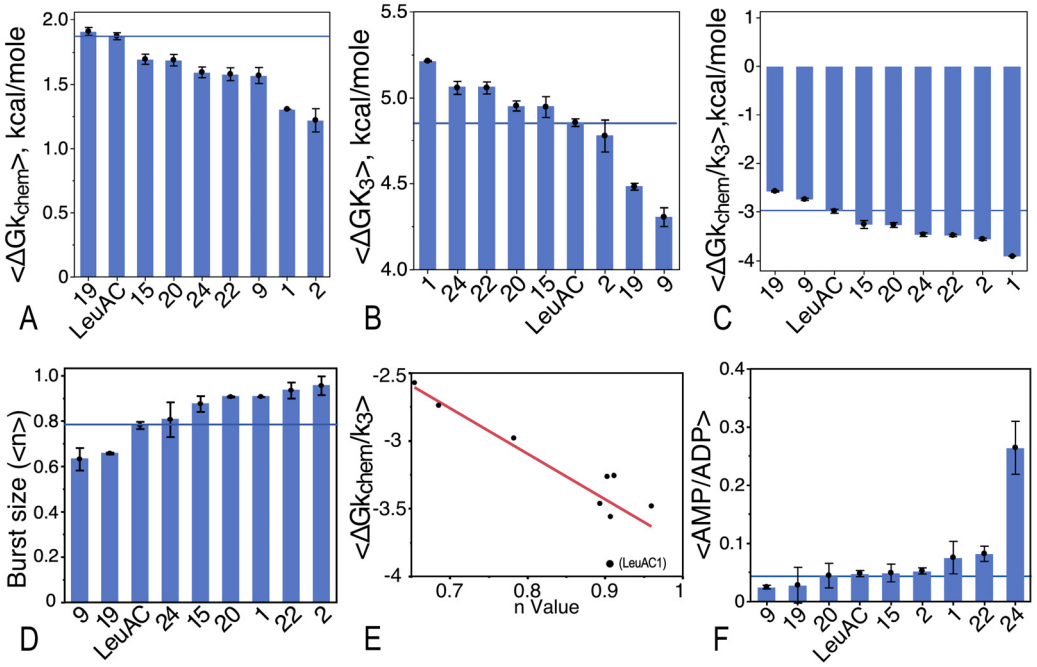


FIG. 4. Parameters derived from single turnover kinetic measurements with leucine. All bar plots are sorted, with more favorable values to the right. Rates in (a)–(c) are expressed as free energies. The blue horizontal lines denote the original LeuAC. (a) First-order rates, ΔGk_{chem}, for the first round of catalysis. (b) Turnover rates, ΔGk_3. Turnover is several orders of magnitude slower than k_{chem} for all variants [see (c)]. (c) The ratio k_{chem}/k_3 measures the binding affinity of the activated aminoacyl-5′AMP intermediate. More negative values (to the right of LeuAC) bind the aminoacyl intermediate more tightly. (d) The burst size, n, is the fraction of macromolecules in the purified catalyst that contribute to ATP consumption. Six of the eight variants have higher active fractions than LeuAC. (e) The burst size, n, is proportional to the apparent affinity for the activated amino acid ($R^2 = 0.77$; $P = 0.02$). (f) The ratio of AMP to ADP produced in the first-order phase of the reaction measures the efficiency with which the urzyme uses ATP.

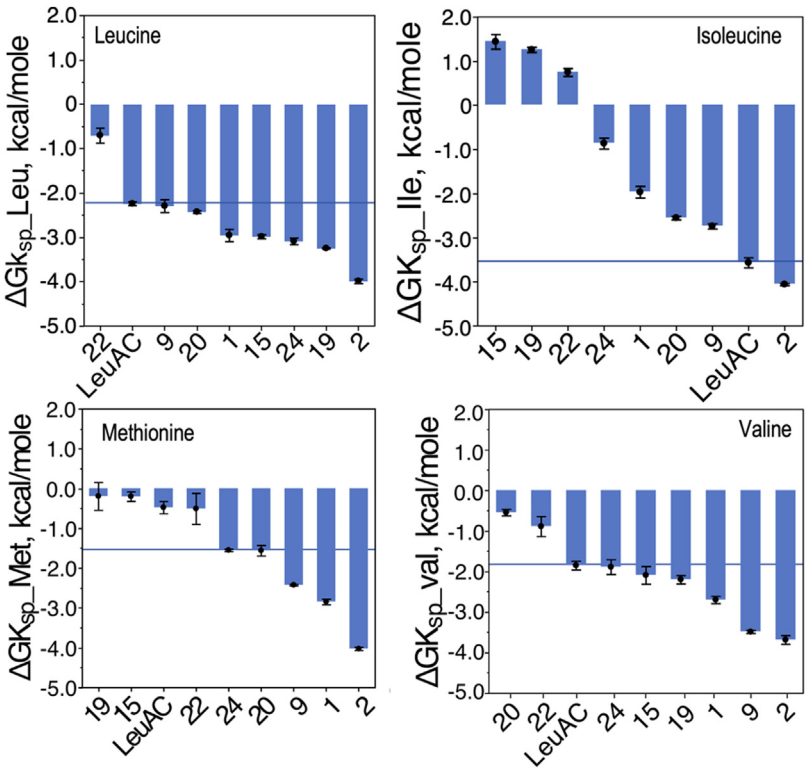


FIG. 5. Specificity constants for activation of four related Class IA amino acids by the nine LeuAC variants. Following the conventions used in Fig. 3, values are sorted in order of increasing catalytic proficiency and the horizontal blue line denotes the original LeuAC variant. This arrangement highlights the variation in activity between variants.

specificities of the variants for different, related amino acids. The second-order rate constant, k_{cat}/K_M, defines the rate at which free enzyme and free substrate combine to form product. It is frequently referred to as the specificity constant, K_{sp}, because the ratio K_{sp_A}/K_{sp_B} defines the relative throughput of competing substrates, A and B.[53] The Malachite Green assay enabled us to measure the spectrum of different specificities of each variant for the four amino acids most closely related to the proper specificity of the contemporary LeuRS. Thus, we measured the steady-state dependence on the concentrations of leucine, isoleucine, valine, and methionine. Figure 5 shows these spectra as bar graphs of the free energy derived from the specificity constant, $K_{sp} = k_{cat}/K_M$. We previously had noted that the original LeuAC urzyme has a preference for isoleucine rather than leucine. The data in Fig. 6 confirm that result.

LeuAC and the eight variants exhibit far more variation in their activities with different amino acids (i.e., their specificity) than they do for their overall rate enhancement for a given amino acid, as indicated by the first-order or turnover in single turnover experiments (Fig. 4). The free energies of the rate constants differ by less than a kcal/mol. For each amino acid, the original LeuAC falls within the range of the eight variants, falling roughly in the middle for all but isoleucine (Fig. 5). Specificities for different variants differ by nearly 5 kcal/mol. For that reason, we delve more deeplyin the next section into how the ProteinMPNN design process affected the relative specificities.

Regression modeling relates specificity differences to mutational changes

It is widely understood that caching the actual physics into the hidden deep learning networks means that they are black boxes. In this work, we have been able to exploit the coherence of the design set, together with extensive experimental characterization, toward illuminating what is inside the black boxes. The new LeuAC variants provide a coherent window on structure/function relationships important to substrate recognition. Specificities of the nine LeuAC variants for four, related amino acids are highlighted in Fig. 6. Six of the mutant residues highlighted in brown in Fig. 2 line the amino acid binding site. These affect the activation reaction significantly. The selected variants include between two and five different amino acid residues at these positions. There thus are a total number of 22 possible independent variables in the experimental matrix. Because one column of each set of residues at the same position depends linearly on the others, Table II shows only 17 of these.

Although the amino acid substitutions can be considered randomly distributed between the nine variants, the sample is not balanced. There is incomplete coverage of two-way interactions between active-site mutations. That lack of balance limits us to consideration of main effects of individual side chain mutations in Fig. 2(c).

The number of truly independent rows in Tables I and II is 9, the number of variants. For that reason, no regression model can use more than eight of the independent variables. With a total of nine variants, we may be able to estimate 5–6 such contributions of the free energy for the specificity constant, $\Delta G(K_{sp})$, without overfitting the data. The JMP^{PRO} stepwise regression module provides a means of interactive selection of models. The experimental matrix (Table II) includes 17 independent variables that might alter the specificities for the four related Class IA amino acids. In our case, the mean relative error in $\Delta G(K_{sp})$ is ~0.15. As noted earlier in Ref. 54, one sign of overfitting is an R^2 value in excess of $1-(\sigma)^2 \sim 0.98$. We have taken care to assure that the models summarized

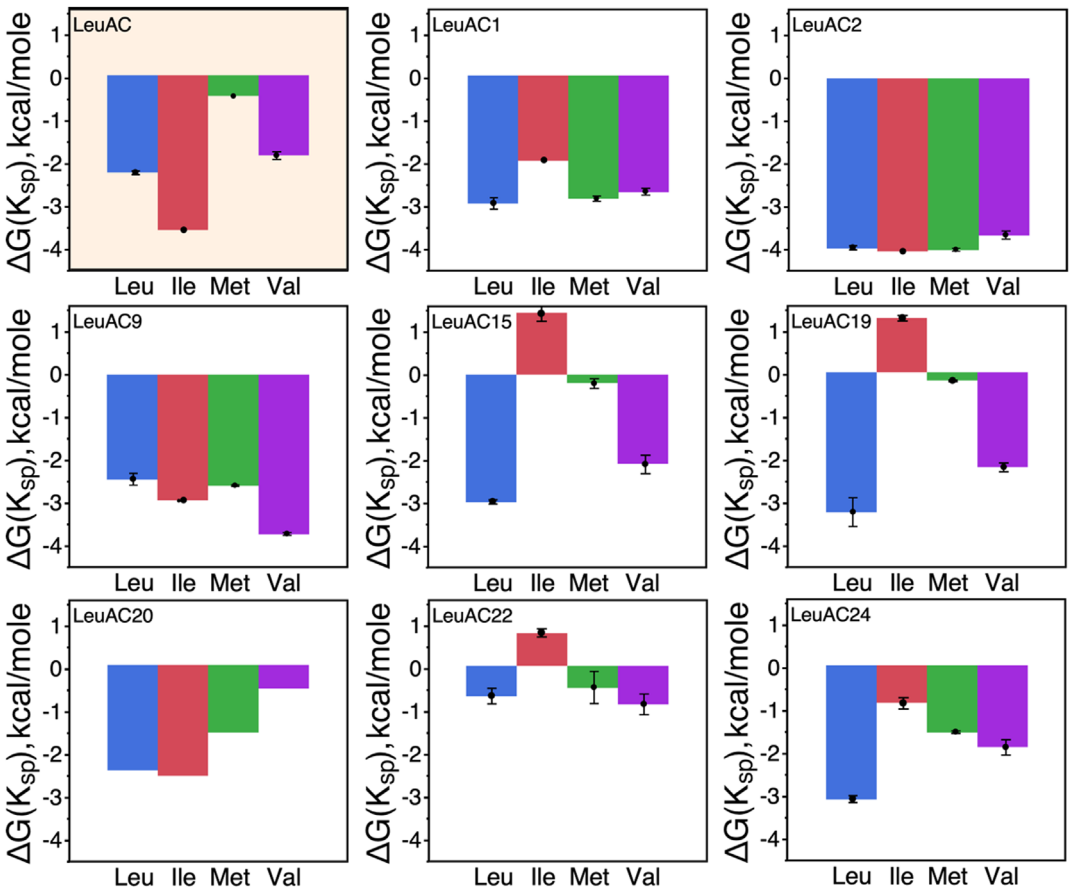


FIG. 6. Comparison of how specificity constants for each amino acid vary with LeuAC variant. The initial LeuAC variant is highlighted in the upper left hand corner by a colored background. Note in particular that because of the conversion to free energies, the error bars are very small for most variants.

in Fig. 7 for each amino acid all have $0.96 < R^2 \leq 0.99$. The models in Fig. 7 are all formally overdetermined by measurements for at least two extra variants. Moreover, the stepwise algorithm shows that all four models represent subsets of predictors whose P-values are orders of magnitude smaller than those of the remaining possible predictors.

Other than the consensus signatures, which were constrained to be wild type histidine and lysine, we did not constrain other residues lining the active site in the design algorithm. Thus, optimization of the various metrics used by ProteinMPNN to sift through the different variants found samples in which the capacity to stabilize tertiary structure also identified some substitutions that affected the amino acid specificity. These residues are highlighted with a brown background in Fig. 1(c). Second, the error bars in Figs. 5 and 6 show very high reproducibility of the Malachite Green assay for the specificity constants, $K_{sp} = k_{cat}/K_M$, for amino acid activation.

Thus, our data predict how unique combinations of the appropriate mutations should enhance the specificity for each of the four amino acids. Those predictions can be tested experimentally. For that reason, this dataset affords a substantial new window on relationships between the presence or absence of particular amino acid side chains and binding specificity.

Detailed structural comparisons between variants support the regression models

We used the coordinates provided by AlphaFold2 to construct from the respective structure predictions what the configuration might be for the side chain binding environments most consistent with optimizing selectivity for each of the four Class IA amino acids, when grafted onto the LeuAC framework (Fig. 8). Each amino acid substrate is surrounded by side chains consistent with the specificity of the aminoacyl-5′ AMP. The results all tend to validate the indications of the respective regression models. None of the arrangements illustrated here are represented in the selected subset of variants.

The strength of these modeling results suggests that if and when we compile experimental melting free energies for the eight variants, we can perform similarly effective analyses of the sources of stability. For that reason, future work using molecular tweezer measurements on single molecules of LeuAC and the eight variants is under way.

However, the models are not all unique. Overlapping models also appear whose predicted kinetic values agree well with the observed values, but which have mutually inconsistent sets of predictors. This result echoes concerns voiced by Ref. 54 about the need for data from additional mutants in order to build confidence in the fitting.

TABLE II. Design matrix for regression analysis of amino acid specificity. The dependent variable columns are shaded light gray. Other columns are potential predictors.

Variant	ΔGKsp L	ΔGKsp I	ΔGKsp M	ΔGKsp V	9T	9L	9I	9F	10A	10P	59W	74Y	74F	74I	79S	79K	80Y	80F	84I	84L	84V
1	−2.95	−1.88	−2.90	−2.77	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
1	−2.82	−2.12	−2.77	−2.72	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
1	−3.10	−1.90	−2.87	−2.59	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
2	−3.93	−4.04	−4.01	−3.81	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1
2	−4.02	−4.09	−3.99	−3.61	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1
2	−4.01	−4.06	−4.07	−3.64	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1
9	−2.13	−2.75	−2.40	−3.45	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0
9	−2.35	−2.80	−2.44	−3.48	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0
9	−2.40	−2.68	−2.44	−3.52	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0
15	−3.00	1.56	−0.12	−1.86	0	1	0	0	0	1	0	1	0	0	1	0	1	0	1	0	0
15	−2.93	1.51	−0.34	−2.11	0	1	0	0	0	1	0	1	0	0	1	0	1	0	1	0	0
15	−3.02	1.25	−0.14	−2.30	0	1	0	0	0	1	0	1	0	0	1	0	1	0	1	0	0
19	−3.23	1.20	−0.41	−2.30	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0
19	−3.25	1.32	−0.37	−2.09	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0
19	−3.26	1.26	0.21	−2.20	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0
20	−2.44	−2.60	−1.64	−0.63	0	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	0
20	−2.45	−2.52	−1.41	−0.47	0	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	0
20	−2.39	−2.53	−1.63	−0.52	0	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	0
22	−0.68	0.84	−0.94	−0.66	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	0
22	−0.88	0.74	−0.18	−1.15	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	0
22	−0.54	0.66	−0.39	−0.84	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	0
24	−3.17	−1.01	−1.58	−2.02	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
24	−3.06	−0.80	−1.53	−1.96	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
24	−3.03	−0.79	−1.53	−1.68	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
LeuAC	−2.21	−3.61	−0.47	−1.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LeuAC	−2.29	−3.44	−0.63	−1.83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LeuAC	−2.23	−3.65	−0.33	−1.96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The methods also should be useful in balancing factorial matrices aimed at restricting the substrate specificity of other urzymes. That possibility is especially relevant in light of our goal ultimately to characterize the optimal specificity possible with such short catalysts, as discussed below.

DISCUSSION

This report combines the use of the two complementary deep learning algorithms ProteinMPNN[2,37] and AlphaFold[4] as well as the DeepSoluE[50] neural network to predict solubility. It illustrates their extended benefits while complementing some of their weaknesses. With few reservations, adapting deep learning algorithms to the challenge of protein design and redesign greatly enhanced our window on how the earliest fledgling enzymes acquired the capacity to translate the genetic code. The chief practical result is that we now have model AARS with sufficient solubility to enable structural studies.

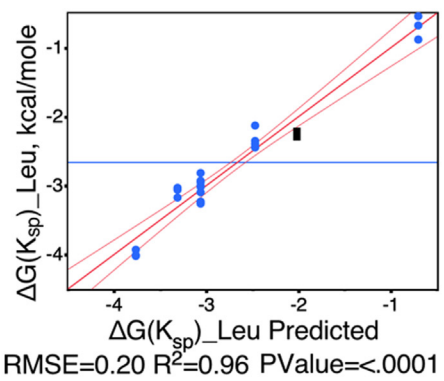
Neural networks are able to use hidden relationships between the elements contributing to desired properties implicitly to enhance them in redesigned sequences. As is true for any neural network, the scope of both deep learning algorithms also depends almost entirely on the training sets. That is especially true for AlphaFold. Structure solution, for example, reveals quite distinct limitations in the use of AlfaFold as a source for molecular replacement.[55] Thus, although AlphaFold2 offers a decisive shortcut to solving structures, it cannot yet substitute for structural data with respect to high resolution details.

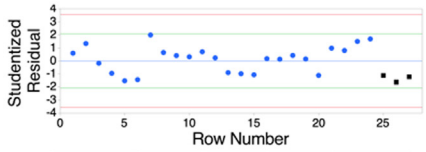
ProteinMPNN, the inverse folding complement of AlphaFold, actually does widen the scope of design applications. In our case, that inversion brings out novel and highly useful results: it substantially increased the LeuAC solubility and it gives us experimental access to the catalytic properties that approximate a virtual protein quasispecies. Until this time, only RNA viral quasispecies[56] had been amenable to experimental study.

We describe here several conclusions from what we were able to do even a short time ago. First, we outline a general protocol with various procedures for analyzing combinatorial libraries of ancestral AARS. The effort to enhance LeuAC solubility led us to a new, potentially useful perspective on how to predict solubility from sequence and structure as predicted by AlphaFold. The sampled ensemble of sequences allowed us to begin to unpack some of the networks that remain hidden in the deep learning algorithms. Finally, our data seem to be surprisingly consistent with properties expected of a protein quasispecies.

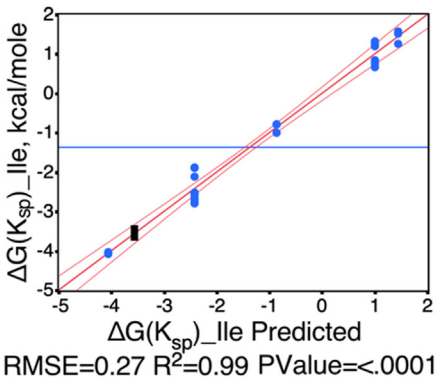
There may be no alternative to studying individual samples of combinatorial libraries

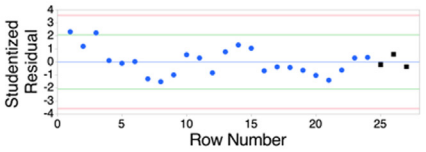
In other work, we have extended the algorithms for phylogenetic ancestral gene reconstruction to accommodate the transitions from an



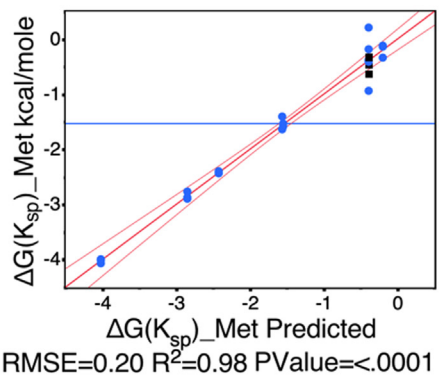


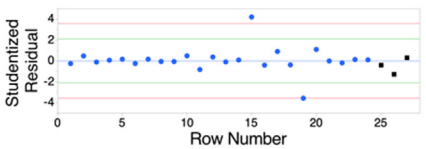
Term	β	σ	t Ratio	Prob>|t|
Intercept	-2.02	0.10	-21	<.0001*
9F	2.36	0.13	18	<.0001*
10P	-1.04	0.12	-9	<.0001*
59W	0.59	0.10	6	<.0001*
84V	-1.29	0.11	-12	<.0001*



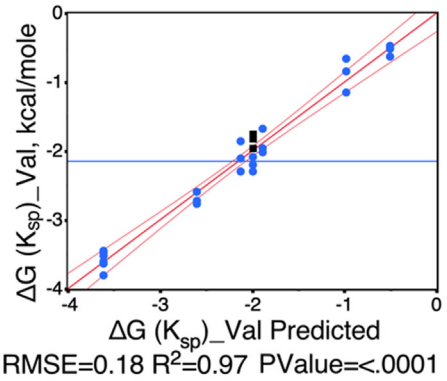


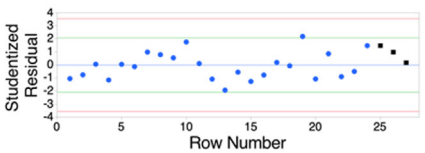
Term	β	σ	t Ratio	Prob>|t|
Intercept	-3.57	0.16	-23	<.0001*
9L	3.87	0.18	21	<.0001*
10A	4.34	0.29	15	<.0001*
10P	1.15	0.18	6	<.0001*
74F	3.42	0.14	24	<.0001*
84V	-1.64	0.18	-9	<.0001*





Term	β	σ	t Ratio	Prob>|t|
Intercept	-3.57	0.16	-23	<.0001*
9L	3.87	0.18	21	<.0001*
10A	4.34	0.29	15	<.0001*
10P	1.15	0.18	6	<.0001*
74F	3.42	0.14	24	<.0001*
84V	-1.64	0.18	-9	<.0001*





Term	β	σ	t Ratio	Prob>|t|
Intercept	-0.39	0.06	-6	<.0001*
9T	1.60	0.15	11	<.0001*
9L	2.65	0.15	18	<.0001*
59W	-1.17	0.10	-11	<.0001*
74Y	-2.46	0.10	-24	<.0001*
80F	1.30	0.15	9	<.0001*

FIG. 7. Regression models implicating specific side chain differences in the active site with differences in amino acid specificity. Blue circles in plots of actual vs predicted values are redesigned variants. Larger black squares in each plot are the original LeuAC. Studentized residuals are dimensionless numbers obtained by dividing the difference between the actual and fitted values and an estimate for the uncertainty of that value, both expressed in the same units as the dependent variable. Such plots reveal outliers outside the red lines above and below 0. Only the regression for ΔG(Ksp)Met has outliers. Since these offset one another, we chose to keep all data when estimating the regression coefficients, β, which are in kcal/mol. Negative β values imply that the residue in question enhances activation of the respective amino acid. Student t-test values are all significant at the level of $P \ll 0.0001$. As noted in the text, that alone does not assure that the models are correct.

n-letter code to an (n-1) letter code. That work[27] together with new techniques for weighing the effects of punctuated equilibria[26] and the impact of insertions and deletions on evolutionary trees[17] should enhance our ability to define likely sequence distributions for successions of ancestral AARS. Those sequence distributions will require new experimental protocols to deal with combinatorial libraries of related proteins. As noted in the Introduction, such libraries can be considered to be protein quasispecies.

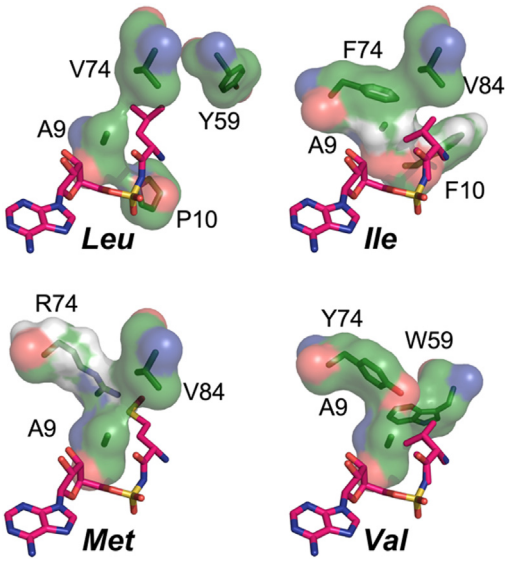


FIG. 8. Structural support for the regression models in Fig. 7 for synthetase specificity. Side chains were selected according to the signs of their regression coefficients. At the same time, we replaced the Leu-5′ sulfoamyl AMP ligand with that corresponding to each amino acid (hot pink), either using coordinates from a corresponding crystal structure (Met, Val, Leu) or by grafting the corresponding amino acid onto the leucyl-5′ sulfoamyl AMP (Ile). This figure both supports the models and stands as a prediction that a variant with the illustrated constellation of side chains should improve selectivity for the amino acid in bold face.

Principal component analysis assured a representative sampling. Doing so enabled us to balance sequence variation with a variety of computational metrics that do not arise from primary structure. Thus, we preferred this approach to alternatives, such as sequence clustering.[57] The resulting sample of sequences was both random and balanced. Extensive searches for both computational and experimental parameters associated with each sequence failed to identify significant correlations between any of the columns of the extended design matrix [see the table in Fig. 2(a)]. Thus, the sample should represent the properties of the fuller set of redesigned sequences.

The sampling was also diverse. The only dependences we found were related to the amino acid specificities of samples on the three major trunks of the constellation plot in Fig. 2(b). The red branch showed a weak but statistically significant preference for methionine relative to leucine. The blue branch showed a similar association with the rejection of valine relative to leucine, isoleucine, and methionine. The green branch showed approximately the same affinity for all four amino acids in Subclass IA. Although these correlations are statistically significant, we cannot attribute any interpretation to them alone.

Our hope was that studies of mixed samples from combinatorial libraries might give a preliminary indication of the diversity of the catalysts in the mix. At the outset, we co-expressed all eight variants and measured the specificity constants of the mixture for the four amino acids compared in Figs. 6 and 7. The results were disappointing in that respect. The $<\Delta GK_{sp}>$ for all four amino acids for the eight variants was about fivefold lower than those for the unfractionated mix. Moreover, the individual values (Fig. 9) did not correlate at all with those measured for the mixture ($R^2 = 0.48$; $P = 0.3$). For these reasons, effective sampling proved to be key to characterizing the variants. At this point, we have not found alternatives to purifying and assaying individual variants.

Regression models for solubility and specificity appear to be predictive

ProteinMPNN found several variants with substantially higher solubility. We can now envision preparing LeuAC variants at high enough concentrations to carry out NMR structure determinations in aqueous media, beginning with HSQC characterization[44,46] of as many as six of the variants we have characterized. It is hard to overstate the potential importance of achieving even that modest goal. It suggests that emergent catalytic functionality is much more broadly distributed in structure space than previously thought possible. We return to this question below in discussing quasispecies.

Predicting various properties, notably solubility and stability, of designed proteins is an active pursuit. While modern sequence design methods have enabled investigators to reliably design proteins with good solubility and stability, it is still difficult to quantitatively predict such properties *a priori* from structural or sequence information alone.

The two additional predictors (% < 0, %IMVLWY) act strongly and synergistically to enhance the significance of the compositional predictors. Both supplemental predictors are more closely associated with tertiary structure than with solubility. In particular, the Delaunay simplices with negative compositional likelihood are exclusively located at the solvent interface. Their contribution may be related to the overall contribution of stability to strengthening the solvent interface. The solubility measurements in Fig. 3 correlate far better with our regression model than with estimates made using neural networks. The large number of predictors (6 of 9 possible) raises the possibility of overfitting. None of the individual predictors has any predictive value. However, the compositional parameters show stronger correlations when three are used together ($R^2 \sim 0.7$, $P \sim 0.05$). In principle, complex dependency of solubility suggests that deep learning methods such as DeepSoluE may be required to predict it.

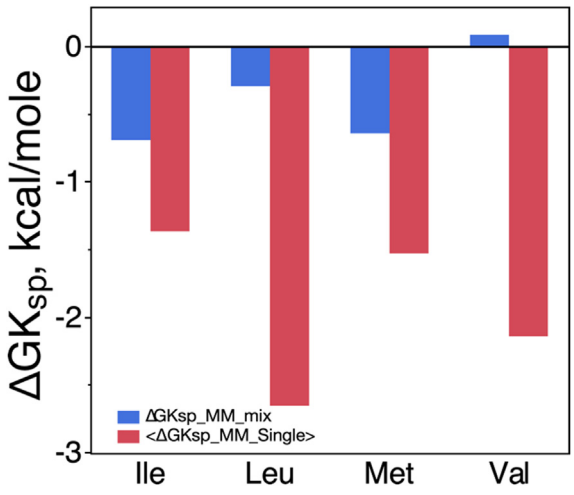


FIG. 9. Comparison of catalytic proficiencies of the eight variant LeuACs and an unfractionated mix.

Like other attempts to predict solubility,[58] DeepSoluE[50] uses a deep learning neural network, which implies the contribution of multiple factors that impact the dependent variable in unknown ways. The regression model for solubility in Fig. 3 is meant to interrogate very specific observed data and look for interpretable patterns to be further explored. Neural networks like DeepSoluE are meant to predict *a priori* how new data will look without access to assay data on very similar variants.

Irrespective of how good the neural networks themselves might be, their usefulness is only as good as the datasets on which they are trained. The training and test sets used for DeepSoluE have little value. They are drawn from anecdotal reports of success or failure to solubilize the respective proteins. The scores used to train and test the neural networks thus are either 0 or 1. For that reason, DeepSoluE cannot be used in efforts to increase the soluble fraction, which is what matters to structural biologists. Thus, it is not surprising that their predictions have so little utility. Nor can the corresponding datasets be used to test the validity of our model.

Instead, we tested the predictive power of our algorithm by performing exhaustive leave out cross-validation. We fitted each of the 81 possible collections of 6 variants, leaving 3 for cross-validation. The median R^2 for the entire set of 81 3-variant test sets was 0.8. However, these were asymmetrically distributed as the top quartile of predictions of the three omitted variants was a sharp peak with $R^2 = 0.97$. Moreover, the regression coefficients were also distributed sharply around the values in the table in Fig. 3. Using the median values for the distributions of the intercept and each of the five coefficients gave an $R^2 = 0.99$. Thus, the model can predict values on which it was trained quite well. In any case, the regression module in Fig. 3(b) appears to be a useful prototype if and when better training and test data become available. In any case, higher solubilities enable downstream structural studies.

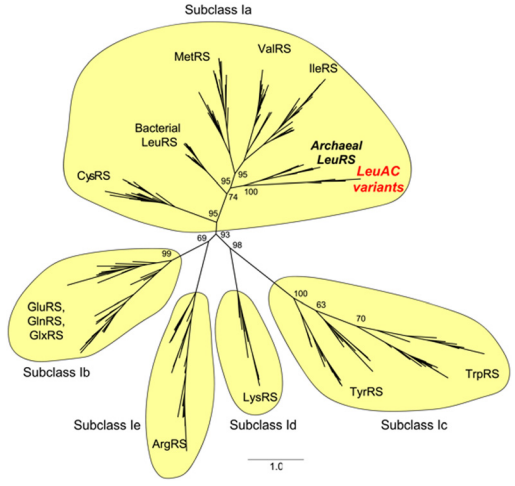


FIG. 10. Maximum likelihood tree showing the relationship between our eight designed LeuAC variants and wild type Class I aaRS sequences. We used IQ-tree v1.6[68] to assess the relationship between our eight designed LeuAC sequences and other wild type Class I sequences. The Class I aaRS alignment contained members from all of the other Class I functional families, and was extracted from aars.online.[67] IQ-tree performed 1000 bootstrap replicates and selected the Blosum62+R5 site model. Bootstrap supports on a scale from 0 to 100 are indicated on backbone internal nodes. Branch lengths are in units of amino acid substitutions per site.

The eight sampled variants exhibit many of the properties expected from a quasispecies

A final important result arises from the various ways in which the LeuAC variant dataset may resemble a protein "quasispecies." Co-evolving molecular sequences cannot be conceived as being in direct competition with one another. Copying of one sequence produces a whole distribution of products, owing to the significant replication error rate. Eigen's concept had immediate application to virology, first in relation to the phage replication,[59] and later with regard to HIV replication in single patients.[60] The distribution of viral genes in a cell of whole organism was important to track how the virus was evolving in patients.

Translational fidelity was doubtless limiting during early evolution. Woese voiced this concern by describing the earliest proteins as "statistical proteins,"[61,62] but without a quantitative framework. The protein "species" required for the cooperative process of translation likely were stable statistical distributions of peptide sequences that were self-generating in the sense that they are responsible for the variations that produce the sequence distribution.[63,64] These would have been genuine protein quasi-species. Hoffmann provided the earliest quantitative model for the impact of translational errors on the self-assembly of genetic coding.[65] Our conceptualization of the original and evolution of genetic coding relies heavily on protein QS narrowing, as fidelity "q" increases, and protein QS bifurcation, as the effective amino acid alphabet size "n" increases.[45] We developed a quantitative dependence of the effective coding alphabet size and on the translational fidelity, q, and n, the actual size of the coding alphabet (see Fig. 3 in Ref. 66).

A key challenge is to be able to characterize the fidelity and coding repertoire of ancestral AARS quasispecies. To that end, we are developing phylogenetic methods to extend ancestral sequence reconstruction algorithms beyond the threshold of the ancestral nodes of the 20 canonical AARS families.[17,26,67] In that regard, we note that the fidelity data in Figs. 5 and 6 provide us with benchmark specificities from a distribution of sequences that bears strong similarities to the archaeal LeuRS clade (Fig. 10). We suggest on that basis that the variants sample an ensemble similar to that of the ancestral LeuRS. To what extent does ProteinMPNN produce sequences that actually resemble those selected by evolution? It will be interesting to follow the sequence comparisons once we have access to ancestral reconstructions using the new, better adapted phylogenetic algorithms.

The sample of related sequences with similar activities described here does not in itself justify thinking of them as a quasispecies. Early QS cannot have had access to all 20 canonical amino acids. However, it is probably fair to assert that they will resemble potential members of any historical QS with comparable catalytic activity that may have existed at some time. Stability is widely thought to be a driver of protein evolution.[69] In that sense, natural selection draws from criteria similar to those embedded in ProteinMPNN. In any case, we have described a protocol for generating and studying closer approximations to the ancestral QS once we improve our estimates of the ancestral coding alphabets.

ACKNOWLEDGMENTS

This work was supported by the Matter-to-Life program of the Alfred P. Sloan Foundation, Grant G-2021-16944, by the National Institutes of Health [Grant R35GM131923 (B.K.), and NSF

fellowship DGE-2040435 (N.R., H.D.)]. H.D. acknowledges receiving support from a Pre-doctoral Fellowship from the American Foundation for Pharmaceutical Education.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Sourav Kumar Patra: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Nicholas Randolph:** Conceptualization (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal); Writing – review & editing (equal). **Brian Kuhlman:** Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Writing – review & editing (equal). **Henry Dieckhaus:** Conceptualization (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Laurie Betts:** Investigation (equal); Methodology (equal); Resources (equal); Writing – review & editing (equal). **Jordan Douglas:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Peter R. Wills:** Conceptualization (equal); Investigation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Charles W. Carter:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1]J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," Nature **630**, 493 (2024).

[2]J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning-based protein sequence design using ProteinMPNN," Science **378**, 49–56 (2022).

[3]B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," Science **302**, 1364–1368 (2003).

[4]J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. T. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583–592 (2021).

[5]E. Gibney and D. Castelvecchi, "Physics Nobel scooped by machine learning pioneers," Nature **634**, 523–524 (2024).

[6]E. Callaway, "Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures," Nature **634**(8034), 525–526 (2024).

[7]M. A. R. Gomez and M. Ibba, "Aminoacyl-tRNA synthetases," RNA **26**, 910–936 (2020).

[8]L. Martinez-Rodriguez, M. Jimenez-Rodriguez, K. Gonzalez-Rivera, T. Williams, L. Li, V. Weinreb, S. N. Chandrasekaran, M. Collier, X. Ambroggio, B. Kuhlman, O. Erdogan, and C. W. J. Carter, "Functional class I and II amino acid activating enzymes can be coded by opposite strands of the same gene," J. Biol. Chem. **290**(32), 19710–19725 (2015).

[9]K. Onodera, N. Suganuma, H. Takano, Y. Sugita, T. Shoji, A. Minobe, N. Yamaki, R. Otsuka, H. Mutsuro-Aoki, T. Umehara, and K. Tamura, "Amino acid activation analysis of primitive aminoacyl-tRNA synthetases encoded by both strands of a single gene using the malachite green assay," BioSystems **208**, 104481 (2021).

[10]Y. Pham, L. Li, A. Kim, O. Erdogan, V. Weinreb, G. Butterfoss, B. Kuhlman, and C. W. Carter, Jr., "A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases," Mol. Cell **25**, 851–862 (2007).

[11]J. J. Hobson, Z. Li, and C. W. Carter, Jr., "A leucyl-tRNA synthetase urzyme: Authenticity of tRNA synthetase urzyme catalytic activities and production of a non-canonical product," Int. J. Mol. Sci. **23**, 4229 (2022).

[12]C. W. Carter, Jr., "Urzymology: Experimental access to a key transition in the appearance of enzymes," J. Biol. Chem. **289**(44), 30213–30220 (2014).

[13]L. Li, C. Francklyn, and C. W. Carter, Jr., "Aminoacylating urzymes challenge the RNA world hypothesis," J. Biol. Chem. **288**, 26856–26863 (2013).

[14]L. Li, V. Weinreb, C. Francklyn, and C. W. Carter, Jr., "Histidyl-tRNA synthetase urzymes: Class I and II aminoacyl-tRNA synthetase urzymes have comparable catalytic activities for cognate amino acid activation," J. Biol. Chem. **286**, 10387–10395 (2011).

[15]Y. Pham, B. Kuhlman, G. L. Butterfoss, H. Hu, V. Weinreb, and C. W. Carter, Jr., "Tryptophanyl-tRNA synthetase Urzyme: A model to recapitulate molecular evolution and investigate intramolecular complementation," J. Biol. Chem. **285**, 38590–38601 (2010).

[16]G. Q. Tang and C. W. Carter, Jr., "*Escherichia coli* transforms the leucyl-tRNA synthetase gene in vivo into primordial genes," bioRxiv (2025).

[17]J. Douglas, R. Bouckaert, C. W. J. Carter, and P. Wills, "Enzymic recognition of amino acids drove the evolution of primordial genetic codes," Nucleic Acids Res. **52**, 558–571 (2024).

[18]G. Q. Tang, J. J. H. Elder, J. Douglas, and C. W. Carter, Jr., "Domain acquisition by class I aminoacyl-tRNA synthetase urzymes coordinated the catalytic functions of HVGH and KMSKS motifs," Nucleic Acids Res. **51**(15), 8070–8084 (2023).

[19]C. W. Carter, Jr., A. Popinga, R. Bouckaert, and P. R. Wills, "Multidimensional phylogenetic metrics identify class I aminoacyl-tRNA synthetase evolutionary mosaicity and inter-modular coupling," Int. J. Mol. Sci. **23**, 1520 (2022).

[20]V. Weinreb, L. Li, S. N. Chandrasekaran, P. Koehl, M. Delarue, and C. W. Carter, Jr., "Enhanced amino acid selection in fully-evolved tryptophanyl-tRNA synthetase, relative to its urzyme, requires domain movement sensed by the D1 switch, a remote, dynamic packing motif," J. Biol. Chem. **289**, 4367–4376 (2014).

[21]L. Li and C. W. Carter, Jr., "Full implementation of the genetic code by tryptophanyl-tRNA synthetase requires intermodular coupling," J. Biol. Chem. **288**, 34736–34745 (2013).

[22]C. W. Carter, Jr., G. Q. Tang, S. K. Patra, L. Betts, H. Dieckhaus, B. Kuhlman, J. Douglas, P. R. Wills, R. Bouckaert, M. Popovic, and M. Ditzler, "Structural enzymology, phylogenetics, differentiation, and symbolic reflexivity at the dawn of biology, genome biology and evolution," bioRxiv (2025).

[23]C. W. Carter, Jr. and P. R. Wills, "The roots of genetic coding in aminoacyl-tRNA synthetase duality," Annu. Rev. Biochem. **90**, 349–373 (2021).
[24]M. Eigen, J. S. McCaskill, and P. Schuster, "Molecular quasi-species," J. Phys. Chem. **92**, 6881–6891 (1988).
[25]M. Eigen, "Molecular self-organisation and the early stages of evolution," Q. Rev. Biophys. **4**, 149–212 (1971).
[26]J. Douglas, R. Bouckaert, S. Harris, C. W. Carter, Jr., and P. R. Wills, "Evolution is coupled with branching across many granularities of life," Philos. Trans. R. Soc. (in press) (2025).
[27]J. Douglas, R. Bouckaert, C. W. Carter, Jr., and P. R. D. Wills, "Reduced amino acid substitution matrices find traces of ancient coding alphabets in modern day proteins," bioRxiv (2025).
[28]C. S. Francklyn, E. A. First, J. J. Perona, and Y.-M. Hou, "Methods for kinetic and thermodynamic analysis of aminoacyl-tRNA synthetases," Methods **44**, 100–118 (2008).
[29]A. R. Fersht, J. S. Ashford, C. J. Bruton, R. Jakes, G. L. E. Koch, and B. S. Hartley, "Active site titration and aminoacyl adenylate binding stoichiometry of amionacyl-tRNA synthetases," Biochemistry **14**(1), 1–4 (1975). [Database].
[30]C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," Nat. Methods **9**, 671–675 (2012).
[31]S. K. Patra, L. Betts, G. Q. Tang, J. Douglas, P. R. Wills, R. Bouckeart, and C. W. Carter, Jr., "A genomic database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class II urzymes," Nucleic Acids Res. **52**, 13305–13324 (2024).
[32]K. A. Johnson, "New standards for collecting and fitting steady state kinetic data," Beilstein J. Org. Chem. **15**, 16–29 (2019).
[33]S. Raran-Kurussi, S. Cherry, D. Zhang, and D. S. Waugh, "Removal of affinity tags with TEV protease," Methods Mol. Biol. **1586**, 221–230 (2017).
[34]S. K. Patra, N. Sinha, F. Molla, A. Sengupta, S. Chakraborty, S. Roy, and S. Ghosh, "In-vivo protein nitration facilitates *Vibrio cholerae* cell survival under anaerobic, nutrient deprived conditions," Arch. Biochem. Biophys. **728**, 109358 (2022).
[35]C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, in Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, 17-23 July 2022 (MLResearchPress, 2022), Vol. 162, pp. 8946-8970, https://proceedings.mlr.press/v162/hsu22a.html.
[36]M. Cagiada, S. Ovchinnikov, and K. Lindorff-Larsen, "Predicting absolute protein folding stability using generative models," Protein Sci. **34**, e5233 (2025).
[37]J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, and D. Baker, "Atomic context-conditioned protein sequence design using LigandMPNN," Nat. Methods **22**, 717–723 (2025).
[38]H. Dieckhaus, M. F. Brocidiaconob, N. Z. Randolpha, and B. Kuhlman, "Transfer learning to leverage larger datasets for improved prediction of protein stability changes," Proc. Natl. Acad. Sci. U. S. A. **121**(6), e2314853121 (2024).
[39]G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters* (Wiley Interscience, New York, 1978).
[40]G. Eriani, M. Delarue, O. Poch, J. Gangloff, and D. Moras, "Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs," Nature **347**, 203–206 (1990).
[41]S. Cusack, C. Berthet-Colominas, M. Härtlein, N. Nassar, and R. Leberman, "A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 Å," Nature **347**(6290), 249–255 (1990).
[42]C. W. Carter, Jr., "Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases," Annu. Rev. Biochem. **62**, 715–748 (1993).
[43]Y. Liu and B. Kuhlman, "RosettaDesign server for protein design," Nucleic Acids Res. **34**, W235–238 (2006).
[44]P. J. Sapienza, L. Li, T. Williams, A. L. Lee, and C. W. Carter, Jr., "An ancestral tryptophanyl-tRNA synthetase precursor achieves high catalytic rate enhancement without ordered ground-state tertiary structures," ACS Chem. Biol. **11**, 1661–1668 (2016).
[45]P. R. Wills, in *Artificial Life IX*, edited by J. Pollack, M. Bedau, P. Husbands, T. Ikegami, and R. A. Watson (MIT Press, Cambridge, 2004), pp. 51–56.
[46]Z. Li and C. W. Carter, Jr., presented at the American Crystallographic Association Annual Meeting, Lexington, KY (unpublished) (2019).
[47]R. Fukunaga and S. Yokoyama, "Crystal structure of leucyl-tRNA synthetase from the archaeon *Pyrococcus horikoshii* reveals a novel editing domain orientation," J. Mol. Biol. **346**, 57–71 (2005).
[48]M. Riès-Kautt and A. Ducruix, "Inferences drawn from physicochemical studies of crystallogenesis and precrystalline state," Methods Enzymol. **276**, 23–59 (1997).
[49]A. Tropsha, C. W. J. Carter, S. Cammer, and I. I. Vaisman, "Simplicial neighborhood analysis of protein packing (SNAPP): A computational geometry approach to studying proteins," Methods Enzymol. **374**, 509–544 (2003).
[50]C. Wang and Q. Zou, "Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE," BMC Biol. **21**, 12 (2023).
[51]S. N. Chandrasekaran and C. W. Carter, Jr., "Adding torsional interaction terms to the Anisotropic Network Model improves the PATH performance, enabling detailed comparison with experimental rate data," Struct. Dyn. **4**, 032103 (2017).
[52]C. W. Carter, Jr., "High-dimensional mutant and modular thermodynamic cycles, molecular switching, and free energy transduction," Annu. Rev. Biophys. **46**, 433–453 (2017).
[53]A. R. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Company, New York, 2017).
[54]F. Yi, D. A. Sims, G. Pielak, and M. H. Edgell, "Testing hypotheses about determinants of protein stability with high-precision, high-throughput stability measurements and statistical modeling," Biochemisty **42**, 7594–7603 (2003).
[55]T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, C. J. Schlicksup, C. Millán, J. S. Richardson, R. J. Read, and P. D. Adams, "AlphaFold changes everything (and nothing)," Acta Crystallogr., Sect. A **79**, C1 (2023).
[56]R. Andino and E. Domingo, "Viral quasispecies," Virology **479–480**, 46–51 (2015).
[57]M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," Nat. Biotechnol. **35**, 1026–1028 (2017).
[58]J. Velecky, M. Hamsikova, J. Stourac, M. Musil, J. Damborsky, D. A. Bednar, A. Evans, and M. Stanislav, "SoluProtMutDB: A manually curated database of protein solubility changes upon mutations," Comput. Struct. Biotechnol. J. **20**, 6339–6347 (2022).
[59]A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," PLoS Pathog. **6**(7), e1001005 (2010).
[60]E. Domingo and S. Wain-Hobson, "The 30th anniversary of quasispecies meeting on 'Quasispecies: Past, present and future'," EMBO Rep. **10**(5), 444–448 (2009).
[61]L. E. Orgel, "Evolution of the genetic apparatus," J. Mol. Biol. **38**, 381–393 (1968).
[62]L. E. Orgel, "The maintenance of the accuracy of protein synthesis and its relevance to ageing," Proc. Natl. Acad. Sci. U. S. A. **49**, 517–521 (1963).
[63]P. Wills, "Scrapie, ribosomal proteins and biological information," J. Theor. Biol. **122**, 157–178 (1986).
[64]P. R. Wills, "Does information acquire meaning naturally?," Ber. Bunsengesellschaft Phys. Chem. **98**, 1129–1134 (1994).
[65]G. W. Hoffmann, "On the origin of the genetic code and the stability of the translation apparatus," J. Mol. Biol. **86**, 349–362 (1974).
[66]P. R. Wills and C. W. Carter, Jr., "Impedance matching and the choice between alternative pathways for the origin of genetic coding," Int. J. Mol. Sci. **21**, 7392 (2020).
[67]J. Douglas, H. Cui, J. J. Perona, O. Vargas-Rodriguez, H. Tyynismaa, C. A. Carreño, J. Ling, L. Ribas-de-Pouplana, X.-L. Yang, M. Ibba, H. Becker, F. Fischer, M. Sissler, C. W. Carter, Jr., and P. R. Wills, "AARS online: A collaborative database on the structure, function, and evolution of the aminoacyl-tRNA synthetases," Iubmb Life **76**(12), 1091–1105 (2024).
[68]L.-T. Nguyen, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," Mol. Biol. Evol. **32**(1), 268–274 (2015).
[69]N. V. Dokholyan and E. I. Shakhnovich, "Understanding hierarchical protein evolution from first principles," J. Mol. Biol. **312**, 289–307 (2001).