Escherichia coli deletes *in vivo* the same domains from a double-mutant leucyl-tRNA synthetase gene that were deleted *in vitro* to make the LeuAC urzyme.

Guo Qing Tang and, Charles W. Carter, Jr[1]

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260

ORCID IDs:

C.W.C: 0000-0002-2653-4452

G.-Q. Tang: 0000-0011-6686-2788

[1]Corresponding Author: Department of Biochemistry and Biophysics University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260

Email: carter@med.unc.ed

SUMMARY

E. coli deletes CP1 and ABD domains from ~40% of plasmids containing leucyl-tRNA synthetase if and only if both active-site signatures are mutated. Shortened ORFs occurred in all reading frames but form three discrete sets in the same frame. One had only the AVGA signature. Two longer ones both retained the same 24-residue segment containing the AMSAS signature. Large pre-steady-state bursts and steady-state acylation assays confirm that they encode active tRNA synthetases. In these respects, the ORFs resemble models for ancestral Class I aminoacyl-tRNA synthetases (AARS). Both signatures thus appear necessary and sufficient for aminoacylation. Growth at 4° C produced most of the middle-sized ORF, which is incompatible with the others and appears to result from a distinct mechanism. Three different linkages between the two parts of the active site acylate tRNA minihelix at similar rates. That result greatly expands the sequence space of active ancestral AARS. Widely spaced active-site mutants thus trigger deletions of modules acquired as full-length AARS evolved from simpler catalysts. We propose that the deletions survive because they limit mischarging due to disrupted coupling of active-site residues to domain motion. Such deletions may thus be a general phenomenon, opening broad access to primordial gene discovery.

Keywords: Origin of genetic coding, induced plasmid instability, single-turnover kinetics, aminoacylation, protein modularity, bidirectional coding

GRAPHICAL ABSTRACT

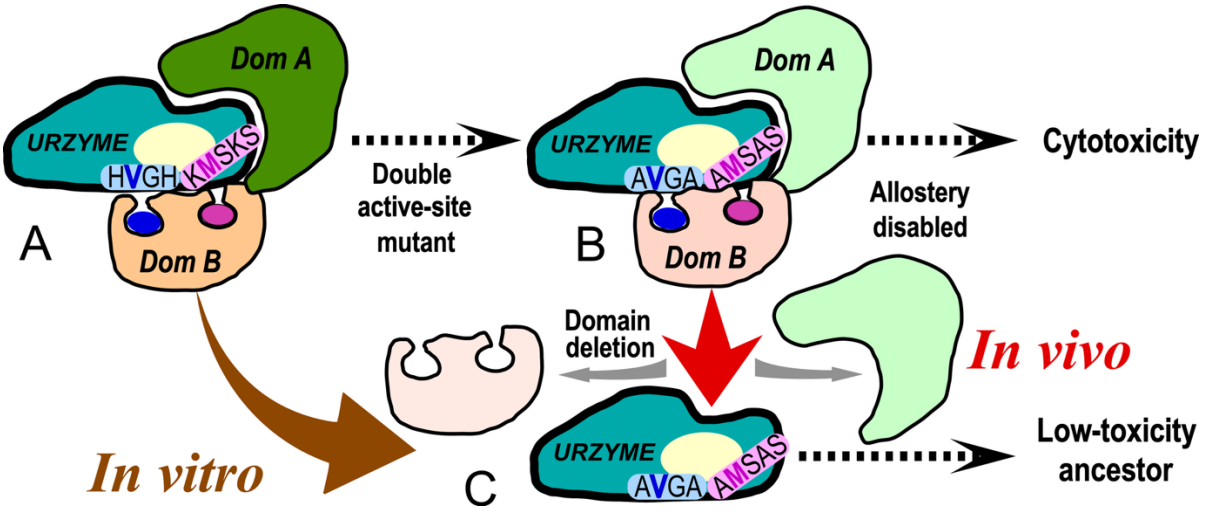


Figure caption. Schematic rationale for formation and selection of LeuRS urzyme-like deletions. **A.** Full length LeuRS uses a complex allosteric network of interactions between Dom A (CP1), Dom B (ABD) and WT active-site catalytic histidine and lysine residues. **B.** Creation of the double mutant corrupts the allosteric effects of the two domains (faded colors). This creates a cytotoxic protein. **C.** Deletion of the domains whose functions have been corrupted by disrupting the allosteric network produces variants lacking the inactivated domains. These variants have considerably less cytotoxicity. They likely also resemble evolutionary precursors of the full-length protein. This may represent a general phenomenon for double mutant multi-domain protein genes.

INTRODUCTION

The Last Universal Common Ancestor, LUCA, lived about 4.2 billion years ago and had a nearly modern metabolism. Moody, et. al. proposed that LUCA had a nearly modern metabolism (Moody, et al. 2024). Similar estimates place the origins of organic chemistry on primitive earth to about the same time. Aminoacyl-tRNA synthetases (AARS) and their cognate tRNAs needed for translation appear to have acquired modern sequences before LUCA appeared (Fournier, et al. 2011; Caetano-Anollés, et al. 2013; Fournier and Alm 2015; Wang, et al. 2025). Thus, both nucleic acids (i.e., genes) and genetically-templated proteins must have achieved modern lengths within the errors of the various dates (Ball 2025). How did that happen?

To address that question we developed models of primordial AARS called "urzymes" (Pham, et al. 2010; Li, et al. 2011; Carter 2014) and "protozymes" (Martinez-Rodriguez, et al. 2015; Onodera, et al. 2021) to help define the transition leading to the code (Carter and Wills 2021; Carter, et al. 2025). Those models for very early stages of AARS evolution (Li and Carter 2013; Weinreb, et al. 2014; Carter, et al. 2022; Tang, et al. 2023; Douglas, Bouckaert, et al. 2024) are based on structural biology. "Urzymes" have 120-130 residues and nearly intact catalytic sites. "Protozymes" are ~50-residue peptides containing only the ATP binding sites. The chief advantage of these experimental models is that their catalytic activities allow us still to study them experimentally. Some modern AARS, however, are nearly an order of magnitude longer than their presumptive primordial ancestors. That gap seems to have been filled by adding new, modular blocks of genetic information to the ancestral forms (Douglas, Bouckaert, et al. 2024).

A novel *in vivo* process in *Escherichia coli* may, at least in part, help to open a new window on how that happened. To better define the functional properties of AARS urzymes (Tang, et al. 2023), we had constructed a plasmid containing a mutant full-length *Pyrococcus horikoshii* leucyl-tRNA synthetase (LeuRS) gene in pET-11a (https://addgene.org/214193). The mutant gene had alanine mutations to both HVGH(AVGA) and KMSKS(AMSAS) signatures *(*Tang, et al. 2023*). E. coli* generated many discrete deletions from that exogenous, nicked plasmid output from high-fidelity PCR amplification. Curation of

66 those deletions revealed sequences similar to those of the protozymes and urzymes we had made on purpose
67 to mimic ancestral AARS.

68 It is routine to isolate and sequence plasmids from multiple transformants to identify one with the desired
69 sequence (Green and Sambrook 2012). The polymerase chain reaction (PCR; Mullis, et al. 1986) can allow
70 single point mutations to accrue. Mutant plasmids can also rearrange in cells grown after transformation.
71 Heretofore, such variation has generally been thought to be a nuisance, as for example were the earliest
72 observations of microRNAs in gels and northern blots (Varani 2015). This is especially true in cases such
73 as ours, that may involve the orthogonality of the respective AARS (Anderson and Schultz 2003). Various
74 precautions can increase the frequency of the desired plasmid (Green and Sambrook 2012) (see also
75 (Tomoiaga, et al. 2021)).

76 To our knowledge, nobody has considered abundance profiling of plasmids produced after transformation
77 as a coherent source of insight about the evolution of protein modularity. In our case, however, the surviving
78 plasmids contained uniquely useful data. These data suggest several novel and functional configurations of
79 the modules we previously showed to be models for ancestral AARS. We outline reasons to think this *in*
80 *vivo* system may be generally useful in the discovery of primordial genes and thus offer an important new
81 window on the modular assembly of protein genes.

82 MATERIALS AND METHODS

83 Materials

84 Genes for Clones 21 and 22 were inserted into pMal-c2x (Plasmid #75286. New England Biolabs, Ipswich,
85 MA.). DNA oligos were ordered from IDT (Integrated DNA Technologies, Coralville, IA, United States).
86 The detailed sequences of oligo DNAs used for mutagenesis (*a1/b1*) and for sequencing (*seqn1/seqn2*)
87 are provided in the Table 2, together with the sequence of full-length LeuRS gene. Phusion™ Plus PCR
88 Master Mix (Cat# F631S) was purchased from Thermo Fisher Scientific (Waltham, MA, United States). *E*
89 *coli* competent cells were partially from Agilent (XL10-Gold ultracompetent cells, Cat# 200315, Santa

90 Clara, CA, United States), and partially using DH5a, home-made, prepared according to the method
91 described by Sharma in 2017. Restrict enzyme DpnI (Cat# 500402, United States) was purchased from
92 Agilent. Purifications and handling of DNA fragment and plasmid used the QIAquick Gel Extraction Kit
93 (Cat# 28704, Qiagen, Hilden, Germany) and QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden,
94 Germany) according to instruction, unless stated otherwise.

95 Table 1. Summary of DNA template sequences of *P. horikoshii* Ph-FL-LeuRS and Oligos for the study.
96 Sequences encoding WT LeuAC urzyme are highlighted in yellow.

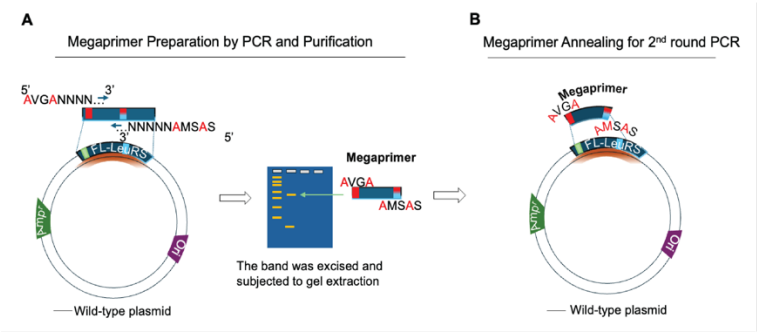
Oligo Identity	**Sequence**
Full-length Leucyl-tRNA synthetase	catatgcatcaccatcaccatcacgcggaactgaactttaaagcgattgaagaaaaatggcagaaacgctggctggaagc gaaaatttttgaaccgaacattcgcgataaaccgaaagaaaaaaaatttatattaccgtggcgtttccgtatctgagcggcca tctgcatgtgggccatgcgcgcacctataccattccggatgtgattgcgcgctttaaacgcatgcagggctataacgtgctgt ttccgatggcgtggcatattaccggcagcccgattgtgggcattgcggaacgcattaaaaaccgcgatccgaaaaccattt ggatttatcgcgatgtgtataaagtgccggaagaaattctgtggacctttgaagatccgattaacattgtgaaatattttatgaa agcggcgaaagaaacctttattcgcgcgggctttagcgtggattggagccgcgaattttataccaccagcctgtttccgccg tttagcaaatttattgaatggcagttttggaaactgaaagaaaaaggctatattgtgaaaggcgcgcatcgcgtgcgctggga tccggtggtgggcaccccgctgggcgatcatgatctgatggaaggcgaagatgtgccgattctggattatattattattaaatt tgaactgcgcgaaaacggcgaagtgatttatctgccggcggcgaccctgcgcccggaaaccgtgtatggcgtgaccaac atgtgggtgaacccgaacgcgacctatgtgaaagcgaaagtgcgccgcaaagataaagaagaaacctggattgtgagca aagaagcggcgtataaactgagctttcaggatcgcgaaattgaagtgattgaagaatttaaaggcgaaaaactgattggca aatatgtgcgcaacccggtgagcggcgatgaagtgattattctgccggcggaatttgtggatccggataacgcgaccggc gtggtgatgagcgtgccggcgcatgcgccgtttgatcatgtggcgctggaagatctgaaacgcgaaaccgaaattctgga aaaatatgatattgatccgcgcattgtggaaaacattacctatattagcctgattaaactggaaggctatggcgattttccggc ggtggaagaagtgaacaaactgggcattaaaagccagaaagataaagaaaaactggaacaggcgaccaaaaccatttat aaagcggaatatcataaaggcatttttaaagtgccgccgtatgaaggcaaaccggtgcaggaagtgaaagaagcgattgc

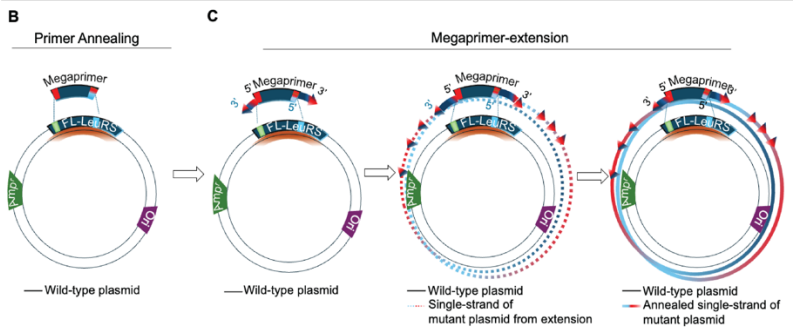
	gaaagaaatgctggaaaaaggcattgcggaaattatgtatgaatttgcggaaaaaaacgtgattagccgctttggcaaccg cgcggtgattaaaattattcatgatcagtggtttattgattatggcaacccggaatggaaagaaaaagcgcgcaaagcgctg gaacgcatgaaaattctgccggaaacccgccgcgcgcagtttgaagcgattattgattggctggataaaaaagcgtgcgc gcgcaaaattggcctgggcaccccgctgccgtgggatccggaatgggtgattgaaagcctgagcgatagcaccatttatat ggcgtattataccattagccgccatattaacaaactgcgccaggaaggcaaactggatccggaaaaactgaccccggaatt ttttgattatatttttctggaagaatttagcgaagataaagaaaaagaactggaaaaaaaaaccggcattccggcggaaattat tcatgaaatgaaagaagaatttgaatattggtatccgctggattggcgctgcagcggcaaagatctgattccgaaccatctga cctttttttatttttaaccatgtggcgatttttcgcgaagaacattggccgaaaggcattgcggtgaacggctttggcaccctgga aggccagaaaatgagcaaaagcaaaggcaacgtgctgaactttattgatgcgattgaagaaaacggcgcggatgtggtg cgcctgtatattatgagcctggcggaacatgatagcgattttgattggcgccgcaaagaagtgggcaaactgcgcaaacag attgaacgcttttatgaactgattagccagtttgcggaatatgaagtgaaaggcaacgtggaactgaaagatattgatcgctg gatgctgcatcgcctgaacaaagcgattaaagaaaccaccaacgcgctggaagaatttcgcacccgcaccgcggtgcag tgggcgttttatagcattatgaacgatctgcgctggtatctgcgccgcaccgaaggccgcgatgatgaagcgaaacgctat gtgctgcgcaccctggcggatgtgtgggtgcgcctgatggcgccgtttaccccgcatatttgcgaagaactgtgggaaaaa ctgggcggcgaaggctttgtgagcctggcgaaatggccggaaccggtggaagaatggtggaacgaaaccattgaagcg gaagaagaatttattcgcagcgtgatggaagatattaaagaaattattgaagtggcgaaaattgaaaacgcgaaacgcgcg tatatttataccgcggaagattggaaatggaaagtggcggaagtggtgagcgaaaaacgcgattttaaaagcagcatggaa gaactgatgaaagatagcgaaattcgcaaacatggcaaagaagtggcgaaaattgtgcagaaactgattaaagaacgcac ctttgatgtgaaacgcattaacgaagaaaaagcgctgcgcgaagcgaaagaatttatggaaaaagaactgggcattgaaat tattattaacccgaccgaagataaaggcggcaaaaaaaaacaggcgatgccgctgaaaccggcgatttttattgaatgataa gctgagc
LeuRS_AVGAa1_F	cgtttccgtatctgagcggccatctgGCGgtgggcGCGgcgcgcacctataccattccgg

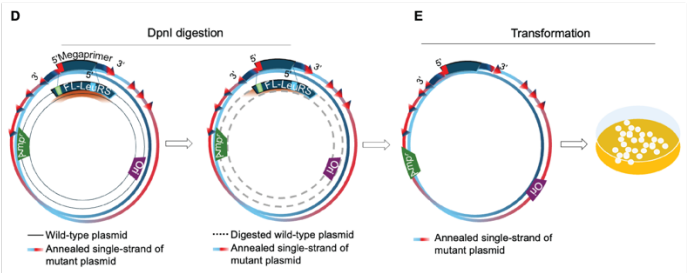
(Forward mutagenesis primer)	
LeuRS_AMSASb1_R (Reverse mutagenesis primer)	caataaagttcagcacgttgcctttgctTGCgctcatCGCctggccttccagggtgccaa
seqn1_AVGA (Forward Sanger sequencing primer)	atgcatcaccatcaccatcacgcggaactg
AMSAS_seqn2rv (Reverse Sanger sequencing primer)	tcatgttccgccaggctcataatatacag

97 **Methods**

98 *Method for preparation of double mutant plasmid.* DNA oligo sequences were designed according to the
99 manual of QuikChange II-E Site-Directed Mutagenesis Kit (Santa Clara, CA, United States) as published
100 online at https://www.agilent.com/cs/library/usermanuals/public/200555.pdf. In order to create the double
101 mutant plasmid DNA, we have implemented a megaprimer PCR method based on that described by Picard
102 (Picard, et al. 1994) (Fig. 1.).







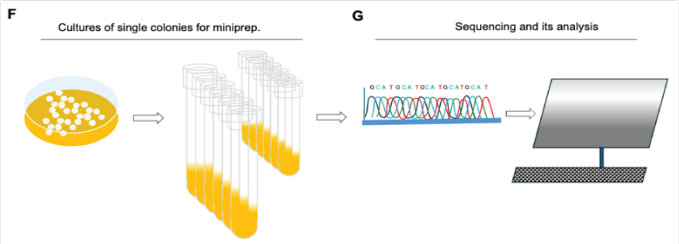


Figure 1. A flowchart illustration of megaprimer-PCR. The method employs a combination of the "Megaprimer-PCR" technique (Picard, et al. 1994) with conventional mutagenesis products (e.g., Agilent, Santa Clara, CA (Technologies)). **A**. Following a first round of PCR (left) with releasing a megaprimer from a target wild-type plasmid DNA, an agarose gel and the PCR product (namely, megaprimer) purification was carried out, in which the propriate mutations AVGA and AMSAS were installed as indicated in bended color squares and amino acid sequences. **B.** For the second round of PCR, annealing the newly purified and resultant megaprimer onto a target plasmid DNA. **C.** PCR extension of the second round of PCR is indicated. The desired PCR extension is initiated and proceeded using the sense-single strand and antisense-single strand of the megaprimer as indicated as colored dot circular lines. **D.** Restriction enzyme DpnI digestion of the second round PCR product, the destroy of original intact wild-type plasmid DNA is indicated as dash lines in grey. **E.** After DpnI digestion of the second round PCR mixture, single colonies of transformants on agar plates can be obtained via conventional transformation with DpnI-digested PCR mixture into competent

114 *E. coli* cells. **F.** Following transformation, a variety of mutant plasmids can be identified through conventional plasmid miniprep.
115 **E.** sequencing, sequence analysis and its identification.

116 First, megaprimer was prepared in a 1st round of PCR. Phusion™ Plus PCR Master Mix (Cat# F631S) was
117 employed according to instruction from manufacture (ThermoFisher Scientific, Waltham, MA, United
118 States). The primer pairs for the 1st round PCR was LeuRS_AVGAa1_F and LeuRS=AMSASb1_R as listed
119 in Table 1 above. In the PCR reaction mix, the plasmid template concentration was 1-4 ng/mL. The primers
120 were used at a concentration of 0.66 mM. The PCR was started at 98 C, 30 sec for initial template
121 denaturation and was following by 44 cycles (98 ºC, 12 sec; 61 ºC, 30 sec; 72 ºC, 3 min 20 sec). After the
122 final extension at 72 ºC, 5 min, the PCR reactions were kept at 4 ºC overnight. After the PCR run, following
123 by horizontal electrophoresis in a 1xTAE, 1.2% Agarose-gel, and visualization by UV transmitter, the
124 megaprimer fragment from the 1st round of PCR was collected from gel cutting out the desired band using
125 a clean razor blade. After DNA fragment purification procedures using QIAquick Gel Extraction Kit (Cat#
126 28704, Hilden, Germany), to yield the megaprimer for the downstream operations.

127 In the second phase, the entire plasmid DNA was amplified using the megaprimer to generate the resultant
128 plasmid harboring the mutant. A 2nd round of PCR was performed using Phusion™ Plus PCR Master Mix
129 (Cat#F631S) and the parameters were setup according to the instructions from the manufacture
130 (ThermoFisher Scientific, Waltham, MA). The PCR procedures were the same as the 1st PCR run. Very
131 importantly, instead of oligo primers, the megaprimer was added at specific desired concentration.
132 Basically, in a 25-40 mL PCR reaction volume, the amount of megaprimer fragment was within a range of
133 0.16mM-1.25 mM, and the template plasmid DNA was 0.5-1.5 ng.

134 After DpnI digestion of the PCR mixture at 37ºC from hours to overnight according to instruction provided
135 by Agilent (Santa Clara, CA, United States), the DpnI-digested PCR mixture is ready for *E. coli.*
136 transformation. The transformation was done using 0.3 mL DpnI-digested PCR mixture per transformation.

137 *Method for E. coli. transformation and the sample submission for sequencing. E coli* transformation was
138 proceeded according to conventional procedure (Fritsch, et al. 1982) and spread onto LB (Lysogeny broth)
139 agar plates with antibiotics (Ampicillin 50 mg/mL, or Carbenicillin 100 mg/mL). After incubation overnight

140 at 37 °C, individual colonies on the LB agar plate were readied for miniprep by following the standard
141 protocol described by QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden, Germany)

142 The resultant plasmid DNA samples were subjected to sequencing by following standard sequencing
143 sample preparation and submission described by Eton Bioscience (San Diego, CA) online
144 https://www.genewiz.com/en/Public/Resources/Sample-Submission-Guidelines/Sanger-Sequencing-
145 Sample-Submission-Guidelines/Sample-Preparation#sanger-sequence with minor modifications.
146 Basically, each sequencing sample was in 15 mL, containing 100-200 ng plasmid DNA with 5-10 picomole
147 sequencing primer, i.e. seqn1_AVGA or AMSAS_seqn2rv as detailed sequences in Table 1 above.

148 *Method for plasmid sequencing and sequencing analysis.* Sequence analysis was performed using
149 conventional methods as follows. First, all the DNA sequences were translated into 6 reading frames in
150 amino acid sequence using Expasy translate online at https://web.expasy.org/translate/. For screening the
151 tentative urzyme candidates, each amino acid sequence from 6 reading frames from a single sequenced
152 plasmid DNA sample were aligned with the sequence of urzyme template using conventional sequencing
153 alignment software such as MAFFT online at https://www.ebi.ac.uk/jdispatcher/msa/mafft?stype=protein
154 (Katoh, et al. 2019) or MUSCLE online at
155 https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein(Edgar 2004). The list of criteria for selection
156 of tentative candidate was according to the sequence length, similarity and identify in addition to the highly
157 conserved catalytic motifs such as motif I HVGH and motif II KMSKS or their-like.

158 The resulting new candidates were designated protozyme, urzyme-like or Δ2ndXvr (for deleted second
159 crossover) and goldilocks, according to their sequence content and were collected and subjected for further
160 downstream identification at the enzymology level after protein biochemistry procedures.

161 *tRNA^{Leu} and TΨC-minihelix^{Leu} preparation.* A plasmid encoding the *P. horikoshii* tRNA^{Leu} (UAG
162 anticodon) was synthesized by Integrated DNA Technologies and used as template for PCR amplification
163 of the tRNA and upstream T7 promoter and downstream Hepatitis Delta Virus (HDV) ribozyme. The PCR
164 product was used directly as template for T7 transcription. Following a 4-hour transcription at 37°C the

165 RNA was cycled five times (90°C for 1 min, 60°C for 2 min, 25°C for 2 min) to increase the cleavage by
166 HDV. The tRNA was purified by urea PAGE and crush and soak extraction. The tRNA 2'-3' cyclic
167 phosphate was removed by treatment with T4 PNK (New England Biolabs) following the manufacturer's
168 protocol. The tRNA was then phenol chloroform isoamyl alcohol extracted, filter concentrated, aliquoted,
169 and stored at -20°C.

170 The original sequence information for composing and designing *Pyrococcus horikoshii* TΨC-minihelix-
171 Leu was based on the mature *P. horikoshii* tRNA-Leu, the tRNA sequence was collected from the
172 complement strand of the genomic sequence of *P. horikoshii* (OT3, GB# BA000001.2, between nucleotides
173 1448081 and 1448168). The designated minihelix sequence was obtained by combining the acceptor stem
174 sequence including DCCA with the TΨC -stem-loop according to Schimmel (Schimmel and Alexander
175 1998). A plasmid harboring the minihelix, an upstream T7 promoter, and downstream Hepatitis Delta Virus
176 (HDV) ribozyme was acquired from previous co-worker Jessica Elder. To reduce challenges posed by the
177 high GC content in the stem portion of minihelix during the preparation, we implemented the Phi29 DNA
178 polymerase-mediated isothermal amplification approach according to the manufacture's protocol (New
179 England Biolabs, Ipswich, MA). After the purification of the product from isothermal amplification, T7
180 transcription was carried out. Following a 4–6-hour transcription at 37 ˚C, the reaction mixture was
181 subjected to urea PAGE fractionation, crush, and soak extraction. After 2'-3' cyclic phosphate removal by
182 T4 PNK, the minihelix RNA was then phenol chloroform isoamyl alcohol extracted, filter-concentrated,
183 quantitated, and aliquoted, and stored at −80 ˚C.

184 *Expression and purification of deletion variants.* We expressed Clones 21 and 22 as MBP fusions from
185 pMAL-c2x in BL21Star (DE3) (Invitrogen) as described in (Tang, et al. 2024). The lysis buffer contained
186 20 mM Tris, pH 7.4, 1 mM EDTA, 5 mM β-mercatoethanol (βME), 17.5% Glycerol, 0.1% NP40, 33 mM
187 $(NH_4)_2SO_4$, 1.25% Glycine, 300 mM Guanidine Hydrochloride) plus cOmplete protease inhibitor (Roche).
188 After elution from Amylose FF resin (Cytiva) with 10 mM maltose in lysis buffer (200 mM HEPES, pH
189 7.4, 450 mM NaCl, 100 mM KCl, 10 mM β-ME. Fractions containing protein were concentrated and mixed

190 to 50% glycerol and stored at -20°C. All protein concentrations were determined using the Pierce™
191 Detergent-Compatible Bradford Assay Kit (Thermo Scientific). We determined purity by running samples
192 on PROTEAN® TGX (Bio-RAD) gels and active fractions for all variants were measured as described in
193 the next section.

194 *Single turnover active-site titration assays*. Active-site titration assays were performed as described
195 (Fersht, et al. 1975; Francklyn, et al. 2008) with the exception that ^{32}P-ATP was labeled in the α position
196 in order to follow the time-dependence of all three adenine nucleotides. of protein was added to 3 μM in 1x
197 reaction mix (50 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 5 μM ATP, 50 mM amino acid, 1 mM DTT, inorganic
198 pyrophosphatase, and α-labeled [^{32}P] ATP) to start the reaction. Timepoints were quenched in 0.4 M sodium
199 acetate 0.1% SDS and kept on ice until all points had been collected. Quenched samples were spotted on
200 TLC plates, developed in 850 mM Tris, pH 8.0, dried and then exposed for varying amounts of time to a
201 phosphor image screen and visualized with a Typhoon Scanner (Cytiva). The time-dependent of loss (ATP)
202 or de novo appearance (ADP, AMP) of the three adenine nucleotide phosphates were estimated using the
203 ImageJ measure function. Time-dependence was fitted using the nonlinear regression module of
204 JMP16PRO™ to equation (1):

205 $$\text{Product(calc)} = A*\exp(-k_{chem}*\text{seconds}) - k_3*\text{seconds} + C \quad (1)$$

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

208 For [ATP] decay curves, the fitted A value estimates the burst size, n, directly as n = A*[ATP]/[Enzyme].
209 C gives n = (1–C) * [ATP]/[Enzyme]. For exponentially increasing concentrations of product, the situation
210 reverses, and n = (1–A) * [ATP]/[Enzyme]and n = C*[ATP]/[Enzyme]. These approximations are justified
211 by the very small variance of multiple estimates.

212 *Aminoacylation assays*. We determined the active fraction of minihelix RNA by following extended
213 acylation assays using the AVGA LeuAC mutant until they reached a plateau. That plateau value was used
214 to compute minihelixconcentrations in assays with all variants.

215 Aminoacylations withs minihelix were performed as described (Tang, et al. 2024) in 50 mM HEPES, pH
216 7.5, 10 mM $MgCl_2$, 20 mM KCl, 5 mM DTT with indicated amounts of ATP and isoleucine. The high
217 affinity of Clones 21 and 22 for minihelix meant that we had to mix increased amounts of [α^{32}P] A76-
218 labeled tRNA for assays. The RNA substrates were heated in 30 mM HEPES, pH 7.5, 30 mM KCl to 90°C
219 for 2 minutes and then cooled linearly (drop 1°C/30 seconds) until it reached 80°C when $MgCl_2$ was added
220 to a final concentration of 10 mM. The tRNA continued to cool linearly until it reached 20°C. Michaelis-
221 Menten experiments were performed by repeating these assays at the indicated RNA concentrations.
222 Concentration-dependence fitted to $K_{sp} = k_{cat}/K_M$ and kcat according to the modified formula (Eqn. 2)
223 introduced by Johnson (Johnson 2019).

224 $$Rate_{calc} = \frac{\frac{k_{cat}}{K_M} \times [RNA,M]}{\left(1 + \frac{k_{cat}}{K_M} * \frac{[RNA,M]}{k_{cat}}\right)} \qquad (2)$$

225 *Data processing and Statistical analysis*. Phosphor imaging screens of TLC plates were densitometered
226 using ImageJ. Data were transferred to JMP16PRO™ Pro 16 via Microsoft Excel (version 16.49), after
227 intermediate calculations. We fitted both active-site titration curves and Michaelis-Menten assays using the
228 JMP16PRO™ nonlinear fitting module.

229 Factorial design matrices (e.g. Table 1) were processed using the Fit Model multiple regression analysis
230 module of JMP16PRO™ Pro, using an appropriate form of equation (3) (Box, et al. 1978)ΔG‡

231 $$Y_{obs} = \beta_0 + \Sigma\beta_i * P_i + \Sigma\beta_{ij} * P_i * P_j + \varepsilon \qquad (3)$$

232 where Y_{obs} is a dependent variable, usually an experimental observation, β_0 is a constant derived from the
233 average value of Y_{obs}, β_i and β_{ij} are coefficients to be fitted, $P_{i,j}$ are independent predictor variables from the
234 design matrix, and ε is a residual to be minimized. All rates and apparent affinities were converted to free

235 energies of activation or binding, $\Delta G = -RT\ln(k)$, before regression analysis. Free energies are additive,
236 whereas rates and apparent affinities are multiplicative. For example, the activation free energy for the first-
237 order decay rate in single-turnover experiments is $\Delta G\ddagger(kc_{hem})$.

238 Multiple regression analyses of factorial designs exploit the replication inherent in the full collection of
239 experiments to estimate experimental variances. In most cases we estimated errors on the basis of t-test P-
240 values. In some cases, we added error bars to histograms to show the variance of individual datapoints.
241 Multiple regression analyses reported here also entail triplet experimental replicates, which enhance the
242 associated analysis of variance.

243 **RESULTS**

244 We describe a novel fragmentation of the leucyl-tRNA synthetase gene from *P. horikoshii* that occurs in
245 ~40% of the plasmids recovered from *E. coli* cells transformed with a double active-site mutant LeuRS
246 plasmid. We described this plasmid as part of an earlier mutagenesis experiment (Tang, et al. 2023). There
247 is nothing subtle about what we did. It is a standard mutagenesis experiment. No selection was applied
248 other than antibiotic resistance. The *E.* coli strain lacked the RecA recombinase. To preclude adventitious
249 loss of sequences during the PCR the double mutant full-length sequence served as a megaprimer (Picard,
250 et al. 1994), as described in Methods. The expression vector, pET-11a, greatly reduces the leaky synthesis
251 of mRNA. Single colonies were sampled without inducing protein expression and grown for sequencing.
252 We had no reason to expect any unusual results.

253 However, the plasmid abundance profile had highly unusual aspects. To understand these, we summarize
254 the LeuRS structural biology and discrete products formed after transformation, in Fig. 2.

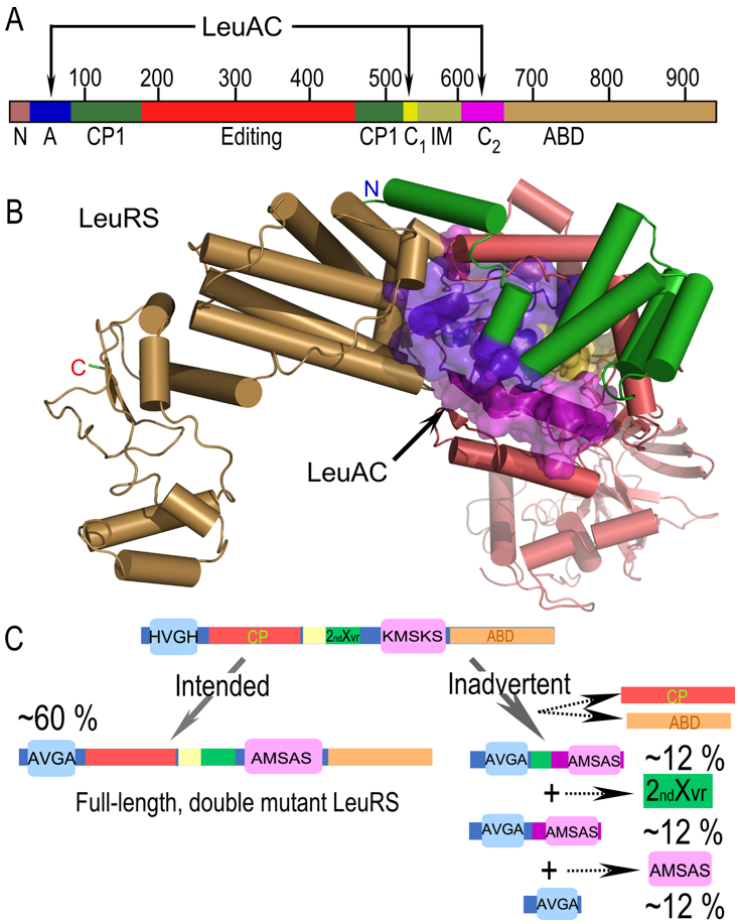


Figure 2. Reference structural schematic of the experiments described herein. **A**. Linear schematic of the 3D structure of the mosaic structure of *P. horikoshii* LeuRS and the LeuAC urzyme. CP1 denotes connecting peptide 1 as described (Burbaum, et al. 1990; Burbaum and Schimmel 1991). Editing denotes the editing domain embedded in CP1. ABD denotes the anticodon-binding domain. A, C1 and C2 are discontinuous segments of LeuRS that form LeuAC. They each are joined by single peptide bonds following excision of the intervening sequences to make a single polypeptide. **B.** 3D cartoon of LeuRS, colored as in A. The LeuAC urzyme we study as a model for the early ancestor of LeuRS is shown by the blue and magenta surface. The components of LeuAC are also indicated above **A. C.** Comparison of the intended double mutant and the three subsets of short, fragmentary ORFs formed *in vivo* from the starting full-length double mutant gene. Note that the 2nd Crossover connection of the Rossmann fold (2ndXvr) is colored green here for clarity. The different N-terminal (blue) segments in truncated fragments are not to scale.

E. coli creates a discrete population of LeuAC urzyme-like coding regions *in vivo*.

Here, we describe the abundance profiling of the *in vivo* deletions (Fig. 2C, Fig. 3) we observed as part of the previous work (Tang, et al. 2023). We sequenced 180 plasmids from transformed single colonies to assess the degree of homogeneity in the ensemble. About 60% (108) of sequenced plasmids isolated had the desired, full-length double mutant. The remaining ~40% had long deletions of coding sequences. The deletion plasmids all contained short stretches of LeuRS coding sequences. Rather than ~2900 nt, they have between 290 – 450nt, far shorter than the full length LeuRS gene. These were re-sorted in idiosyncratic

272 ways. Coding sequences occurred in all six reading frames. Yet, when annotated for continuous open
273 reading frames and aligned to that of the full length LeuRS gene, they form the coherent alignment shown
274 in Fig. 3A. The raw sequence data for the insertions in 34 of these plasmids from Fig. 3A and 3C are in
275 Supplemental Section A. A fasta file with these and other, similar deletion plasmids is in Supplemental
276 Section B. Genomic features, together with the temperature at which they were incubated and the respective
277 reading frames are shown schematically in Supplementary Figs. S1-S7 and collected in Supplementary
278 Table S1.

279 We describe here only the first 34 such sequences analyzed. ORFs from all the additional, subsequently
280 sequenced plasmids fall into the same three categories as those highlighted by black vertical arrows in Fig.
281 3A.

282 Each clade differs in a distinct way from the analytical LeuAC urzyme (Carter, et al. 2014; Hobson, et al.
283 2022) we studied earlier (Tang, et al. 2023; Tang, et al. 2024). Fig. 3A summarizes this unexpectedly
284 coherent mosaic. A neighbor joining tree based on the alignment in Fig. 3A reveals only three clades (Fig.
285 3B). All shortened ORFs deleted major segments from the full-length LeuRS gene. Notably, there are no
286 remnants of the two large domains we deleted to generate LeuAC. The first missing domain is the 444-
287 residue Connecting Peptide 1 (CP1; residues 83-527) that provides the editing function. The second domain
288 missing in all deletions is the 304-residue anticodon-binding domain (ABD; residues 663-967).

289 The coding sequences retained in the shortened ORFs comprise three discrete sets, each with different
290 properties. The shortest subset (Supplemental Figs S2-S4) retains only a 28-residue segment (residues 41-
291 68) containing the AVGA sequence and which forms the ATP binding site. About 11% (8) of the 34
292 sequences encode truncated protozymes, (blue backgrounds) retain only these sequences. We refer to this
293 set as protozyme-like (from "Proto", relating to a precursor). These sequences are highlighted in blue in
294 Fig. 1A. These residues are present in virtually all deletions.

295 A longer subset (Supplemental Figs. S1,4) contains, in addition, to the intact, 50-residue protozyme,
296 residues 650-674 that contain the AMSAS motif. This 25-residue segment is against a sand background in

Fig. 3A. Although it contains the entire length of the protozyme, it is the shorter of the two deletions that contain both mutant catalytic signature motifs. About ~10% (7) of the sequences in Fig. 3A have this intermediate length. It is ~81 residues long. We expressed this ORF as Clone 21 and for reasons outlined in detail below, we refer to it as the Goldilocks urzyme because it is just the right length to align antiparallel to minimal Class II urzymes.

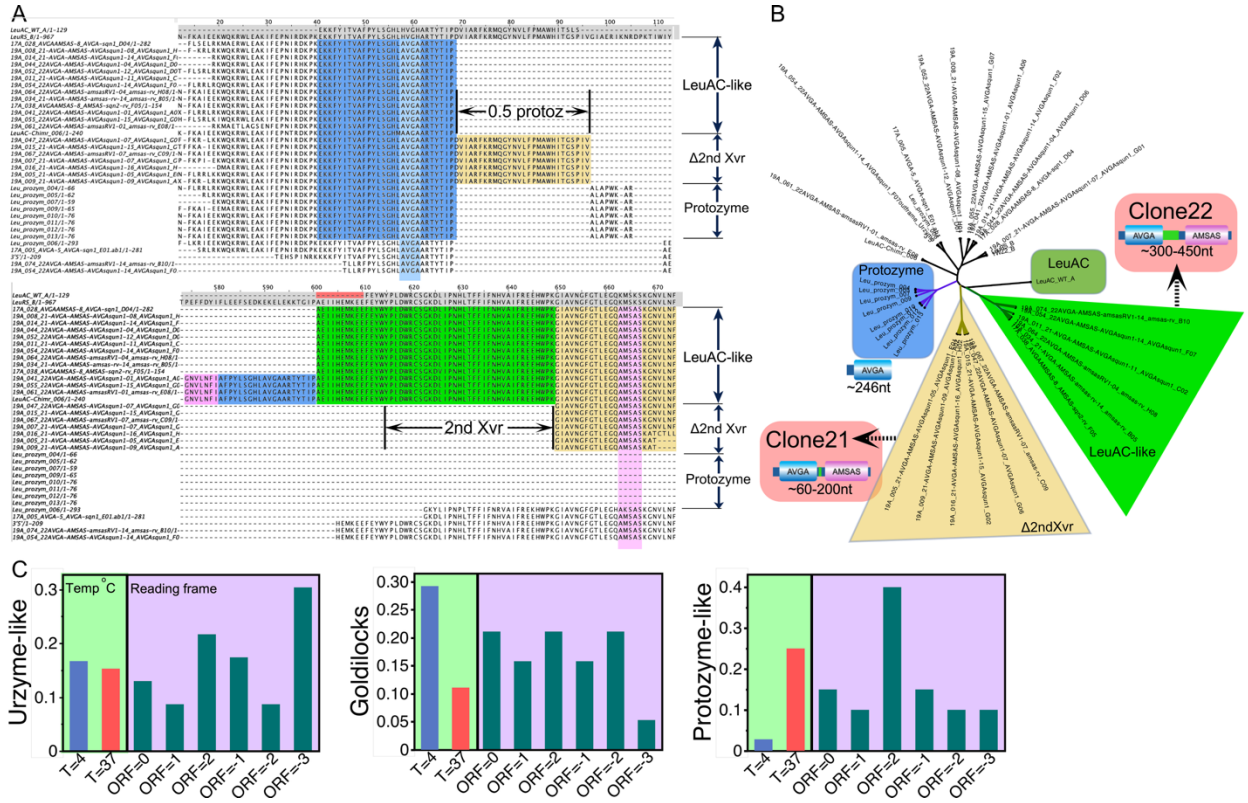


Figure 3. Shortened ORFs generated *in vivo* from full-length LeuRS double active-site mutants. **A.** Modular deletions and shuffling in excerpts from the multiple sequence alignment (Tang and Carter 2025). Both panels together show the LeuAC sequence and corresponding segments from full-length *P. horikoshii* leucyl-tRNA synthetase in rows 1 and 2 (light grey background at the top). Vertical light blue and magenta columns indicate the two mutant signatures. Green horizontal backgrounds highlight uniquely retained sequences of the LeuAC-like urzyme, Clone 22. These lack the second half of the protozyme (blue background). Sand backgrounds highlight urzyme sequences uniquely retained in the Goldilocks variant, Clone 21. It lacks the second crossover connection of the Rossman dinucleotide-binding fold (Buehner, et al. 1973; Burbaum and Schimmel 1991). **B.** Neighbor-joining tree from the alignment in A highlights the three principal clades of related sequences. Coloring is that in A. For subsequent reference Clone21, the Goldilocks deletion variant, was excerpted from one of the second crossover deletions (denoted Δ2nd Xvr). Clone 22 contains the second crossover lacking in Clone21 and has an extended segment preceding it (red band under sequence numbers). Clones 21 cannot be a subset of Clones 22.**C.** Genomic properties show that incubating transformant colonies at lower temperatures leads to the appearance of Goldilocks deletions (green panels). Magenta panels show the distribution of reading frame as follows. ORF 1 is the same reading frame as the intact LeuRS plasmid. ORFs 1,2 are advanced by 1 and 2 frames. ORFs -1, -2, -3 are corresponding frames in the 3'-5' direction. The different temperature-dependence and mutually exclusive sequences in Goldilocks and Urzyme-like coding regions imply that they arise via different intracellular DNA rearrangements. Further details on the incidence of tandem repetition are included in Supplementary Table S1.

The longest ORF in Fig. 3A (300 – 450nt; 19%; 14 sequences; Supplemental Figs. 1, 6,7), includes residues 588-636 which code for the second crossover (Richardson 1973) of the Rossmann fold (Buehner, et al.

321 1973) (green backgrounds in Fig. 3A). That segment contributes a third motif, **GKD**L (**G**x**D**Q in other
322 Class I AARS). It contributes the aspartate side chain to the active site and we presume it coordinates the
323 active-site Mg^{++} ion (Weinreb and Carter 2008; Weinreb, et al. 2009; Williams, et al. 2015). These
324 sequences also have only a half-protozyme and an extra segment that is not part of LeuAC (red band just
325 under the sequence numbers). We expressed this ORF as Clone 22 and refer to it as "urzyme-like".

326 Curiously, the lengths of the truncated protozyme ORFs match exactly that portion of the protozyme in the
327 LeuAC-like ORFs. The protozyme segment is twice as long in the intermediate-sized ORF which lacks the
328 entire second crossover segment of the Rossmann fold (Buehner, et al. 1973). Each clade in the ensemble
329 is remarkably discrete and reproducible. Occasional point mutations outside the LeuRS coding regions that
330 differentiate the different truncated plasmids from one another are not highlighted.

331 We call the 81-residue ORF the "Goldilocks" urzyme because it is "just right" for antiparallel alignment
332 with the shortest Class II urzymes (Patra, et al. 2024). We return to the Goldilocks urzyme in a subsequent
333 section, because of its inherent compatibility with the shortest Class II AARS urzymes.

334 Supplementary Section C contains figures showing the variety of urzyme, Goldilocks, and protozyme
335 sequences that appeared in sequenced plasmid samples as summarized in Fig. 3C. Curious genomic features
336 include single open-reading frame deletions, (Supplemental Figure S1, S5); tandem-configurations,
337 (Supplemental Figures S2, S3, S7); and multi-frame embedding within the same plasmid (Supplemental
338 Figures S3, S4, S6, S7). The bottom four sequences in the second half of the LeuAC-like deletions illustrate
339 such features. They have an inserted segment that derives from two different regions of the full-length
340 sequence (highlighted in the bottom panel in Fig. 3A). Blue background shading represents a duplication
341 of the first half of the protozyme. It does not vary in either protozyme-like or LeuAC-like sequences. A
342 short N-terminal segment (magenta background) derives from the C-terminus of both the LeuAC-like and
343 Goldilocks urzymes, C-terminal to the AMSAS signature. The last 5 sequences in Fig. 3A are aberrant
344 deletions and difficult to classify. They contain additional deletions in both N- and C- terminal segments.
345 Their coding sequences are nevertheless free of point mutations.

346 **The deletions must occur *in vivo*.**

347 Full-length mutants at only one site do not produce any deletions. The process thus requires both sets of
348 mutations. Nor does the reciprocal mutagenesis of double mutant plasmid with WT primers. These control
349 experiments also mean that deletions derive not from the PCR mutagenesis itself. Rather, they must arise
350 *in vivo,* after transformation. The cellular process is dependent on dual mutagenesis to the active site, hence
351 therefore also one-way. We note as well that most, if not all ORFs have open reading frames in frame on
352 the opposite strand. These ORFS cannot be annotated in terms of known contemporary protein sequences.
353 We have not yet investigated whether or not the opposite strand ORFs might encode proteins homologous
354 to Class II AARS as predicted by the Rodin-Ohno hypothesis (Rodin and Ohno 1995).

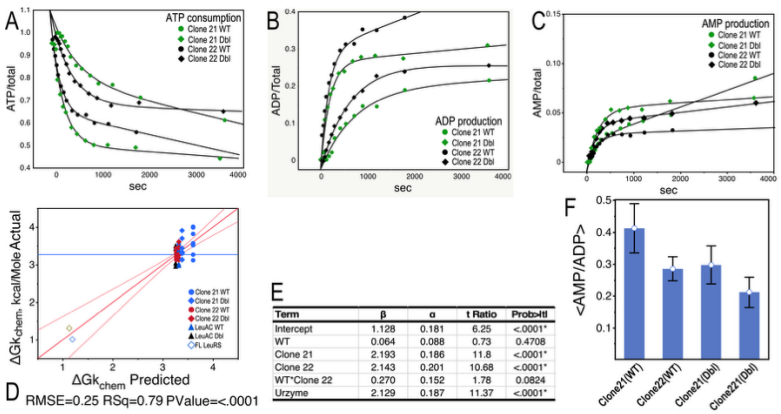
355 We cannot suggest a detailed mechanism, but the deletions must have been created and selected *in vivo*.
356 We consider factors influencing selection of the coding sequences retained in the Discussion. We note here
357 that they are likely coupled to DNA replication because mRNA synthesis from the pET-11a host plasmid
358 is minimal without induction. Notably, those incubated exclusively at 37º C had no Goldilocks deletions
359 and fewer tandemly repeated and multi-frame encoding than did those at 4º C. The enrichment of deletions
360 at low temperature (Fig. 3C) is consistent with the known accessibility of D-loops to recombination proteins
361 and replication restart at low temperature (Michel and Sandler 2017).

362 **The deletions encode active tRNA synthetases.**

363 AARS urzymes retained three catalytic properties seen in full length AARS. (i) They accelerate amino acid
364 activation. (ii) They show significant pre-steady state bursts in single-turnover kinetic assays (Fersht, et al.
365 1975; Francklyn, et al. 2008). (iii) They accelerate aminoacylation of tRNA and minihelix substrates (Tang,
366 et al. 2024). Clone21 and Clone22 (Fig. 3B), retain all three catalytic activities at levels comparable to those
367 observed for the urzymes generated intentionally on the basis of structural biology (Carter 2014).
368 Experimental kinetics data confirm all three criteria (Figs. 4-6).

Fig. 4 summarizes single-turnover experiments on wild type (WT), double (AVGA/AMSAS; Dbl) Clone 21 and 22, as well as LeuAC urzyme (Hobson, et al. 2022; Tang, et al. 2023; Tang, et al. 2024) variants. Single-turnover experiments (Fersht 1975; Francklyn, et al. 2008) use large enzyme concentrations to build the first-round product formation into a measurable signal from the first round of catalysis. They use only a small excess of substrate; Most kinetic events are first order; hence they measure unimolecular reactions. If product release is rate-limiting, the bound product builds up in the enzyme. Such experiments show a pre-steady-state burst. Burst sizes represent the fraction of active enzyme molecules contributing to the experimental signal. These are shown in Supplementary Fig. S8. They range from ~0.3 for the WT Clone 22 to ~0.6 for the double mutant Clone 22 with both WT and double mutant Clone 21 values both close to 0.55. Single turnover data in Fig. 3 reveal differences between Clones 21 and 22.

Time courses of the concentrations of all three adenine nucleotides for all four variants (Fig. 4A-C) are similar to those observed previously for LeuAC (Tang, et al. 2024). The first-order rate constants for amino acid activation (Supplementary Table S1) are very nearly the same as those for LeuAC (Fig. 4D, E) and orders of magnitude slower than those of full-length LeuRS. Moreover, as we found for LeuAC (Tang, et al. 2023) the AVGA and AMSAS mutations actually enhance $k_{chem.}$ for Clone 21. But for Clone 22, which is slightly more active overall, that effect was reversed. P-values for the β-coefficients of the regression model in Figs. 4D and 4E) suggest that although the difference between the two variants is significant, the difference between the effects of mutant and WT active-site residues is smaller.



Term	β	α	t Ratio	Prob>|t|
Intercept	1.128	0.181	6.25	<.0001*
WT	0.064	0.088	0.73	0.4708
Clone 21	2.193	0.186	11.8	<.0001*
Clone 22	2.143	0.201	10.68	<.0001*
WT*Clone 22	0.270	0.152	1.78	0.0824
Urzyme	2.129	0.187	11.37	<.0001*

Figure 4. Single turnover analysis of the recombinant deletion urzymes. **A.** Time course for ATP consumption by all four variants. **B.** Time course for ADP production. **C.** Time course for AMP production. **D.** Regression model for the dependence of the first-order rate constant, k_{chem}, on variant and mutant properties. **E.** Table of regression coefficients for the model in **D**, their estimated errors and P values. **F.** Histogram of the ratio of AMP to ADP produced by the four urzymes. WT Clone 21 has the highest ratio, hence uses ATP most efficiently. This ratio is consistent with the fact that the burst size of this variant in ADP production is lower than the others and the fact that the steady-state rate of AMP production is higher.

Single-turnover experiments yield two rate constants. The first-order rate constant, k_{chem}, refers to the depletion of ATP in the chemical step. The turnover rate is k_3. The mean ratios k_{chem}/k_3 for Clones 21 and 22 (230) are similar to that for LeuAC (223). Thus, the pre-steady state bursts are comparable. All retain the activated Ile-5'AMP with an increased affinity of ~ –3.2 kcal/mole.

The ratio of AMP to ADP produced in such assays measures the efficiency ATP usage to produce activated amino acid (Patra, et al. 2024). Interestingly, the WT Clone 21 uniquely makes less ADP (green dots in Fig 3B) and more AMP. It makes more efficient use of ATP than the others (Fig. 4F).

Steady-state kinetics confirm that Clones 21 and 22 catalyze aminoacylation. The essence of translation in biology is forming a covalent bond between an amino acid and an RNA substrate. Our previous work with the LeuAC urzyme derived from LeuRS showed three crucial details (Hobson, et al. 2022; Tang, et al. 2023; Tang, et al. 2024). (i) The preferred RNA substrate, by a factor of ~10, is the 35-nucleotide minihelix containing the acceptor stem and the TΨC-minihelix. (ii) The preferred amino acid substrate is isoleucine, not leucine. (iii) Catalysis of acyl transfer to RNA is minimally impacted by alanine substitution of the polar side chains in HVGA and KMSKS signatures. For these reasons, we measured rates for isoleucine and minihelix substrates for the new double mutant (Dbl) constructs generated *in vivo* as well as their counterparts (WT) with histidine and lysine in the respective signatures. Triplicate assays of acylation are shown in Fig. 5 and shown in Supplementary Table S2.

The distinction between single-turnover and steady state assays becomes blurred because we study catalysts that represent increasingly ancestral forms. As turnover decreases, steady-state assays require longer times and higher enzyme concentrations to build an adequate signal-to-noise ratio. Unpublished time-course data also suggest that the observed rates are, at least partially, first-order processes. $\Delta G(k_{cat})$ values from aminoacylation correlate with a linear combination of $\Delta G^{\ddagger}(k_{chem})$ and $\Delta G^{\ddagger}$ (k_3) values from single turnover

experiments ($R^2 \sim 0.97$; $P \sim 0.03$). Regression coefficients (see Supplemental Section E and Fig. S8) suggest that $\Delta G(k_{chem})$ contributes about 65% of the correlation and $\Delta G(k_3)$ 35%. We have not found any way around this. It is a cost of working close to the boundary where experiments are no longer possible.

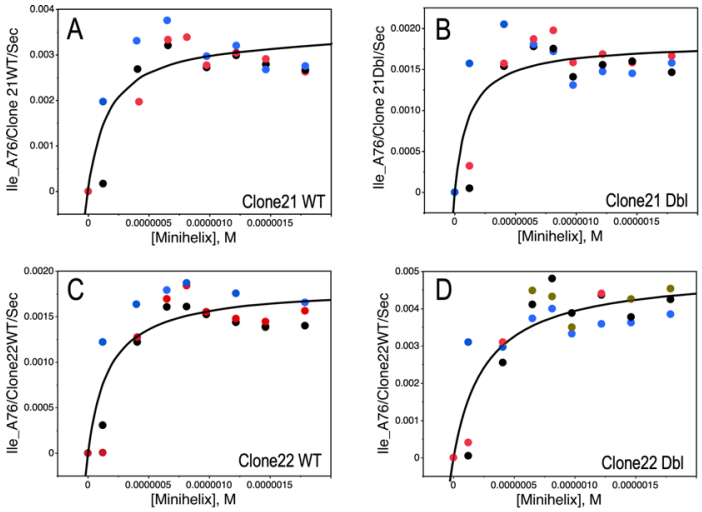


Figure 5. Michaelis-Menten plots of replicated aminoacylation assays. Aminoacylation of the leucine TΨC minihelix by Clone21, the Goldilocks urzyme with WT HVGH and KMSKS active site signatures (**A**) and the double mutant (**B**) sequences. **C, D.** Aminoacylation of the minihelix by Clone22 WT and DBL, respectively. Note that the vertical scales in **A-D** are different. The relative activities are compared in Fig. 5.

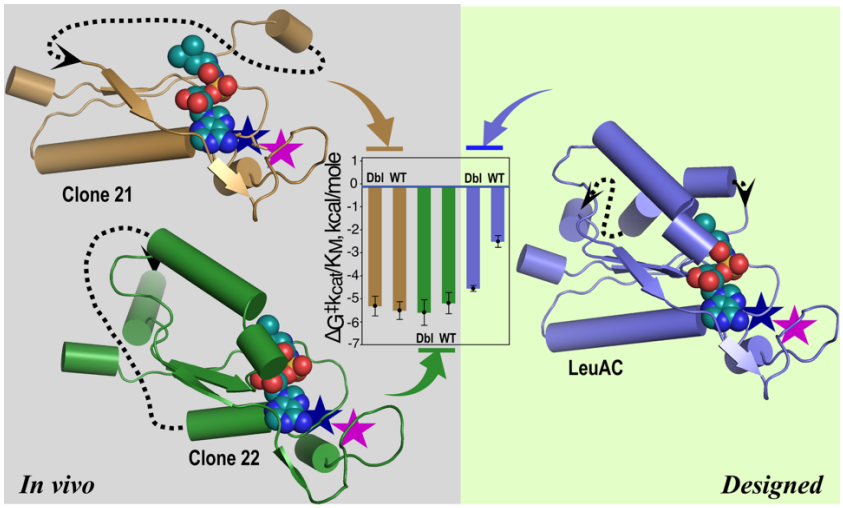
The curves in Fig. 5A, 5B, 5C do indeed suggest a decrease in observed rates at higher [minihelix]. The [substrate] necessary to detect dissociation between enzyme and minihelix RNA, and the apparent dissociation constants for the enzyme•RNA complex, ~0.2 μMare an order of magnitude lower than the enzyme concentration. The RNA substrate is therefore rapidly consumed, and the acylated RNA product may remain bound, inhibiting further reaction by reducing $[E_{free}]$.

Clones 21 and 22 connect the two active-site signatures in two, entirely new, functional ways.

Fig. 3A shows that segments derived from the LeuRS active site are always present with their original sequences. That is what allows the neighbor joining algorithm used in Fig, 3B to identify only three clades. The original LeuAC design strategy preserved as much of the active site as possible. Only two short connections were made in order to delete the CP1 and CP2 peptides. Clones 21 and 22 have configurations that omit longer and different segments from those present in LeuAC. This leaves us puzzled as to how the

435 two active-site signatures can assume similar configurations to those we imagine from the 1WZ2 crystal
436 structure.

437 In turn, the significantly different linkages of the two parts of the active site introduce quite a new element
438 into our thinking about ancestral Class I AARS (Fig. 6). LeuAC, designed *in vitro*, has 129 amino acid
439 residues. The two *in vivo* LeuRS urzymes are both smaller. Clone 22 has residues 41-68 + 601-674 = 102
440 residues. Clone 21 has 41-95 + 650-674 = 81 residues. Yet, the averages of the four bars representing the
441 WT and double mutant *in vivo* constructs are ~ 2 kcal/mole more negative than the corresponding average
442 for LeuAC (Tang, et al. 2024). Thus, they both catalyze minihelix acylation about 30 times better than the
443 original LeuAC.



444
445 Figure 6. The N- and C-terminal parts of the LeuRS active site can be connected three distinct ways (dashed arrows) without
446 changing the catalytic activity in RNA minihelix aminoacylation by either WT or double mutant forms (bar graph). The three active
447 tRNA synthetases have, respectively, 81 (sand), 102 (green), and 129 (blue) residues. All cartoons are based on PDB 1WZ2. The
448 construct on the right was previously designed to mimic the ancestral LeuRS. The two on the left were created *in vivo* by deletions
449 to a full-length double mutant LeuRS plasmid. Wild type histidine (blue stars) and lysine (magenta stars) residues were added back
450 by PCR mutagenesis for purposes of comparison.

451 **Regression modeling identifies and attributes selectable functions.**

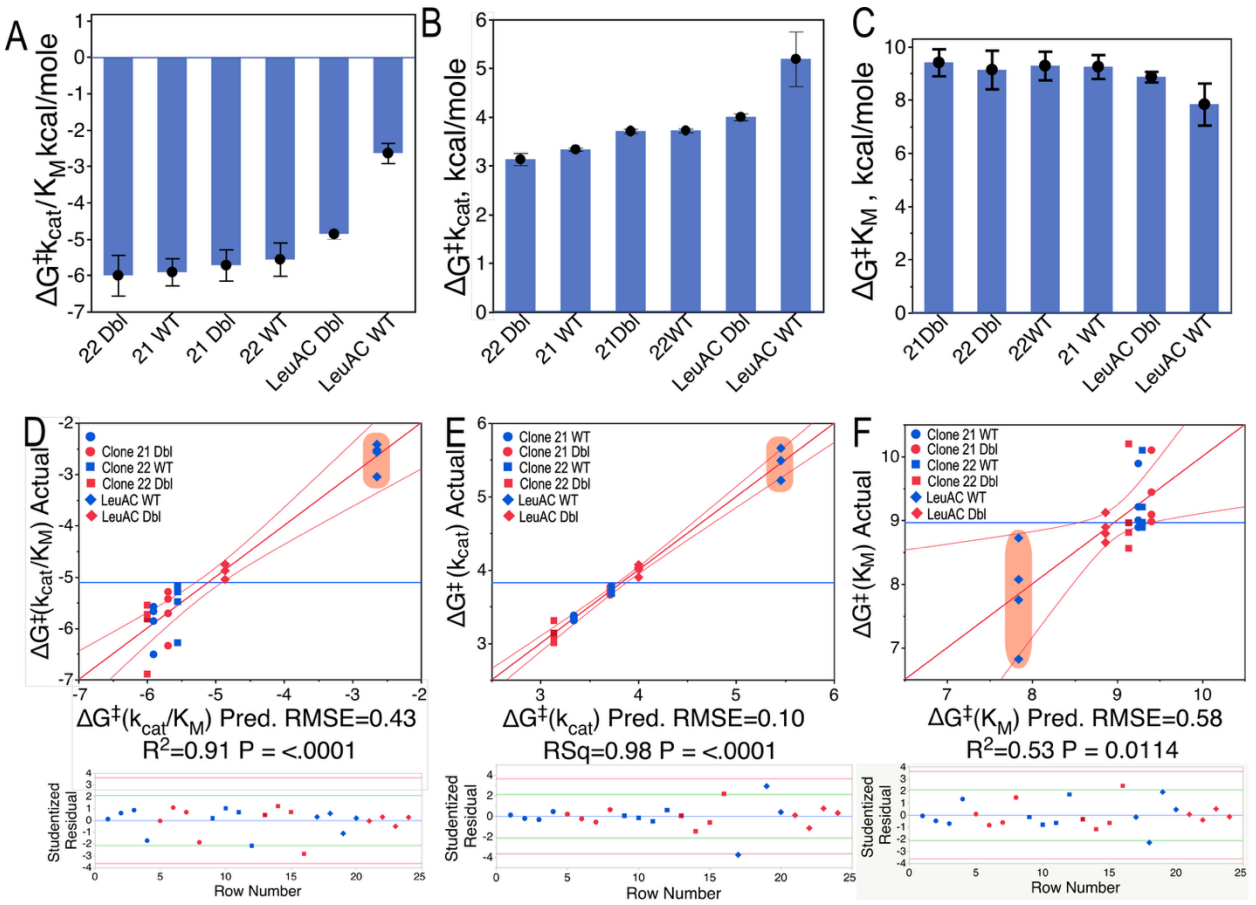
452 The *in vivo* fragmentation allows us to analyze the enzymology of LeuAC at unprecedented resolution. The
453 second crossover connection (missing in Clone 21) and second β-strand of the protozyme (missing in Clone
454 22), together with the WT vs double mutant active site residues form an incomplete factorial design (Carter
455 and Carter 1979). The logarithmic transformation from rates to free energies of activation allow testing of
456 linear models (Weinreb, et al. 2012a; Li and Carter 2013; Tang, et al. 2023; Tang, et al. 2024). Regression

457 models for the Michaelis-Menten constants in Supplementary Table S2 show that the second half of the
458 protozyme and the 2nd crossover of the Rossmann fold have significant effects on the $\Delta G^{\ddagger}k_{cat}$, the *in* situ
459 first-order process that transfers the acyl group to the cognate minihelix. Such effects offer an experimental
460 basis on which to assess why particular modules might be selected.

461 Clones 21 and 22 are slightly better than the AVGA/AMSAS double mutant of LeuAC at aminoacylation.
462 They aminoacylate minihelices significantly better than WT LeuAC does (Fig. 7). The enhanced values of
463 k_{cat}/K_M result both from increased k_{cat} (Fig. 7B) and higher affinity for minihelix RNA (Fig. 7C). Active-
464 site mutations have opposite effects on the two *in vivo* constructs. The double mutant (Dbl) Clone 22 is the
465 best catalyst of aminoacylation. It is nearly matched by the WT Clone 21. The WT Clone 22 is the least
466 active variant. The decreased activity of WT variants that contain the second crossover connection of the
467 Rossmann fold (i.e. LeuAC and Clone 22) has a P value $<10^{-4}$. We noted a similar effect on the burst size
468 (Supplementary Fig. S8).

469 The wild-type active-site variants bind more tightly to substrate minihelix in both Clone 21 and Clone 22,
470 (Fig. 7C), but not in LeuAC. All four *in vivo* variants are better catalysts in all respects than the double
471 mutant LeuAC. All of the effects influencing $\Delta G^{\ddagger}k_{cat}$ have highly significant P-values (Fig. 7B). The WT
472 active side residues change the effects of all three parameters. The second crossover connection reverses
473 the effect of active-site mutations on Clone 22. This is apparent from the placement of the various symbols
474 on the graph in both Fig. 7D and 7E.

475 The rose shading in all three plots shows that the kinetic properties of WT LeuAC are all significantly worse
476 than all of the constructs produced *in vivo*. Estimation of higher-order effects is limited by the poor
477 reproducibility of K_M. Notwithstanding, significant evidence remains for a negative interaction between the
478 second crossover connection and the second crossover connection of the Rossmann fold (Fig. 7D, E). That
479 interaction reduces both $\Delta G^{\ddagger}k_{cat}$ and $\Delta G^{\ddagger}k_{cat}/K_M$ by about 2 kcal/mole. Clones 21 and 22 therefore enhance
480 studies of the evolution of protein modularity at unrivalled resolution.



Term	β	σ	t Ratio	Prob>|t|
Intercept	-6.83	0.37	–18.36	<.0001*
WT	-1.99	0.53	–3.78	0.0014*
2ND XVR	0.83	0.30	2.74	0.0134*
WT*2ND XVR	2.43	0.43	5.66	<.0001*
Protoz	1.14	0.30	3.75	0.0015*
WT*Protoz	1.78	0.43	4.13	0.0006*

Term	β	σ	t Ratio	Prob>|t|
Intercept	2.85	0.09	32	<.0001*
WT	-1.24	0.13	-10	<.0001*
2ND XVR	0.29	0.07	4	0.0009*
WT*2ND XVR	1.83	0.11	17	<.0001*
Protoz	0.87	0.07	12	<.0001*
WT*Protoz	0.86	0.11	8	<.0001*

Term	β	σ	t Ratio	Prob>|t|
Intercept	10.36	0.37	28.05	<.0001*
WT	-0.34	0.23	-1.45	0.1635
2ND XVR	-0.97	0.29	-3.41	0.0032*
WT*2ND XVR	-0.87	0.57	-1.52	0.1466
Protoz	-0.86	0.29	-3.01	0.0075*
WT*Protoz	-1.18	0.57	-2.07	0.0533

Figure 7. Michaelis-Menten parameters of AARS urzymes depend on the structural differences between constructs. Each plot is ordered in descending favorability from left to right. **A**. Bar plot of $\Delta G^{\ddagger}$ (k_{cat}/K_M), the free energy for the second order rate constant for aminoacylation by reaction of free TΨC-minihelix with free substrate at saturating concentration of isoleucine, the amino acid substrate. **B**. $\Delta G^{\ddagger}(k_{cat})$, the free energy for the first-order rate constant. **C.** $\Delta G^{\ddagger}$ (K_M), the free energy for the apparent dissociation constant for the minihelix substrate. **D. E. F.** Regression models that relate steady-state parameters in **A-C** to the structural differences. Independent variables denote the following: WT refers to the presence of WT active-site histidine and lysine residues; 2ndXvr denotes the presence of the second crossover; Protoz denotes the presence of the intact protozyme at the N-terminus.

Class 1 and 2 AARS urzymes use similar secondary structures for aminoacylation.

Clone 21 is the smallest functional Class I AARS gene ever produced. Clones 21 and 22 share the AMSAS signature. They differ in both the protozyme and second crossover. The C-terminal fragment shared by both Clone22 and Clone21 is key to binding RNA substrates. It is strong evidence that acquisition of that signature converted Class I protozymes into aminoacylating enzymes. Curiously, the secondary structure

of the segment is a β-strand followed by a loop. The Motif 2 loop in Class 2 AARS also forms much of the RNA binding site. A striking resemblance of the RNA-binding secondary structures of Class 1 and 2 AARS is shown in Fig. 8. There are thus extraordinary similarities between the stereochemistry in both sets of polypeptide•RNA interactions leading to aminoacylation. We have suggested elsewhere how the contrasting substrate-binding interactions of Class I and II AARS can arise from the inversion symmetry of base-pairing in bidirectional ancestral genes (Carter and Wills 2019a; Carter and Wills 2019b, 2021; Carter 2024; Carter, et al. 2025).

The modest size of the Class I AMSAS-containing module is reminiscent of the N-terminal fragment of the protozyme, which also is ~25 residues long and has a similar β–>turn secondary structure and is key to the binding of ATP. In fact, the corresponding fragment in Class II AARS is actually the same fragment, residues 54-78 in Fig. 8. As we noted previously (Tang, et al. 2023), neither the histidine side chains in the HVGH nor the lysine residues in the KMSKS signatures are necessary for amino acid activation or aminoacylation. The full functionality of the Lys-to-Ala mutations implies that the function of that segment arises solely from polypeptide secondary structure.

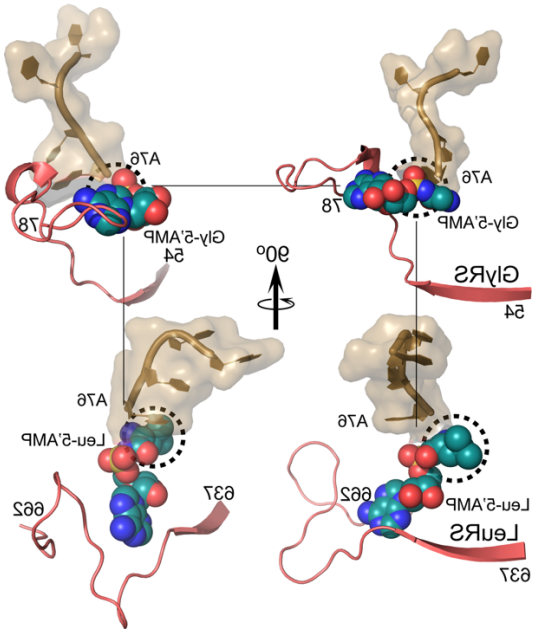


Figure 888. Corresponding segments of the active site that bind RNA substrates in Class 1 LeuRS (PDB ID 1WZ2) and Class II GlyRS (PDB ID 7YSE) have very similar secondary structures. Approximately orthogonal views of the complexes between RNA

binding motifs in the two Classes with their respective activated amino acyl-5'AMP and RNA ligands. Dashed circles highlight the aminoacyl groups. The vertices of the square in the background indicate the 5'-phoshporyl group. They are located remarkably similarly with respect to the β-strand and loop. A chlorine atom that is part of the adenine ring in 7YSE has been omitted.

The Goldilocks urzyme suggests how one gene can encode both Class I, II AARS.

The Rodin-Ohno hypothesis that bidirectional ancestral AARS genes coded for Class I and II urzymes on opposite strands (Rodin and Ohno 1995) led to a surprising number of retrodictions about the Class distinction (Carter, et al. 2025). The bidirectional protozyme gene was readily constructed (Martinez-Rodriguez, et al. 2015) because the corresponding segments from Class I and II AARS have the same number of amino acids. Others have verified its activation activity (Onodera, et al. 2021). We have failed, however, to design a bidirectional gene encoding Class I and II AARS urzymes on opposite strands because there are several loci where the combination of the two signatures has different numbers of amino acids and cannot be fitted into a consistent antiparallel alignment.

Because it lacks the second crossover connection of the Rossmann fold, Clone 21 is the same length (81 residues) as the Class I GlyCA urzyme (Patra, et al. 2024). Further, 81-residue Goldilocks excerpt is a catalytically active AARS. Moreover, the Class-defining signature sequences of Class I and II AARS can both be readily aligned opposite one another in antiparallel alignments by the homology suggested by the GlyCA structure (Carter, et al. 2025). Thus, the Goldilocks urzyme, Clone21, may show how to explore designed bidirectional Class I/II AARS urzyme genes.

We compare in Fig. 9 antiparallel alignments of Goldilocks-length urzymes drawn from 10-12 species of each synthetase from those in aars.online (Douglas, Cui, et al. 2024). The three Class I AARS subclassses match uniquely with Class II subclass urzyme genes. Superior matches entail both the total length and significantly elevated codon middle-base pairing. From the table in Fig. 9B, Class IA IleRS pairs best with IIB LysRS; Class IIA ProRS pairs best with Class IB GlnRS. Class IC TrpRS pairs best with Class IIC HisRS. We previously outlined extensive tests of the pairing frequencies under the null hypothesis of no bidirectional coding ancestry for such alignments (Chandrasekaran, et al. 2013). Standard errors for all mean values in Fig. 9B are derived from ~140 bidirectional alignments. They about 1 percent of the values

themselves. The middle-base pairing frequencies are thus all highly non-random and statistically different one from another. Moreover, the different lengths of the three sets of related pairings make it hard to evaluate cross pairing in the matrix in Fig. 9B.

Pairings in Fig. 9B conflict with the conventional sub-classification by pairing subclass A in Class I with subclass B in Class II and vice versa. Each such pairing brings along two additional AARS from the same respective subclasses. Thus, 16 of the 20 bacterial AARS sequence alignments show analogous pairings between Class I and II AARS genes (Carter, et al. 2025).

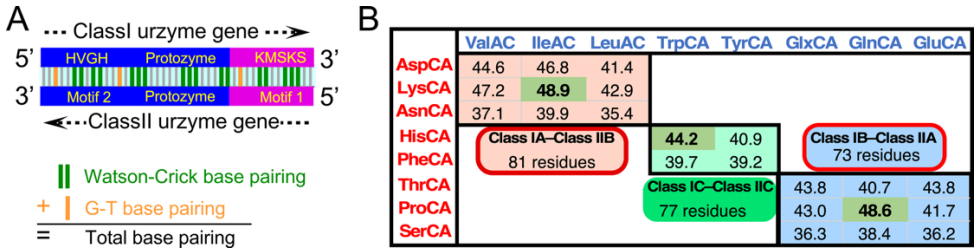


Figure 9. The Goldilocks urzyme, Clone21, suggests a paradigm for designing bidirectional urzyme genes for Class I and II ancestors on opposing DNA strands. **A.** Schematic of sense/antisense encoding of Class I and II AARS urzyme genes, consistent with the architecture of the two AARS Classes. Vertical lines between the two genes suggest bases that would be paired in a bidirectional gene but are only partially paired in sequences derived from contemporary genes. **B.** Percentages of codon middle-base pairing in antiparallel alignments, calculated as indicated in **A** are summarized for three groups derived from the three consensus synthetase subclasses (Carter, et al. 2025). Values greater than ~30% are highly unlikely under the null hypothesis.

DISCUSSION

We describe a novel host-plasmid interaction that occurs when a PCR-generated, nicked plasmid containing a Leucyl-tRNA synthetase double mutant is introduced into *E. coli*. LeuRS is a large, multidomain protein. Transforming *E. coli* with a variant containing mutations in two remote parts of the active site gives rise to a high proportion (~40%) of a discrete set of very long deletions. Most of the short ORFs retain sequences encoding active leucyl-tRNA synthetases that resemble the LeuAC urzyme (Hobson, et al. 2022). The rest encode a shortened variant of the LeuRS protozyme.

The LeuRS ORFs made *in vivo* validate previous studies of AARS ancestors.

This result unexpectedly confirms many aspects of our *in vitro* studies of ancestral AARS (Pham, et al. 2007; Pham, et al. 2010; Li, et al. 2011; Li and Carter 2013; Li, et al. 2013; Carter 2014, 2015; Martinez-

561 Rodriguez, et al. 2015; Carter and Wills 2021). We had thought that the full protozyme and 2nd crossover
562 segments were essential for acylation. Neither is actually required. Catalytic proficiencies of Clones 21, 22
563 are even better assignment catalysts than the LeuAC double mutant (see Fig. 6 and K_{sp} values in
564 SupplementalTable S2). It is hard to overstate the significance of these parallels. They substantiate our
565 assertion that AARS urzymes are valid models for early AARS evolution.

566 Analytical structure-based deconstruction took us ten years. Here, we found that *E. coli* creates ORFs *in*
567 *vivo* and in a week or so, that are uncommonly similar to those we previously described. The Clone 21 and
568 22 deletions represent at least four discoveries we otherwise would not have made: (i) the mutant AVGA
569 and AMSAS active-site signatures can form functional linkages in surprisingly simple ways we had not
570 considered (Fig. 6). (ii) the module that enables the Class I protozyme to recognize and acylate RNA
571 substrates is a short segment including the downstream AMSAS signature. That signature forms a structure
572 quite similar to that of the RNA binding site of Class II AARS. (iii) the Goldilocks urzyme shows that the
573 second crossover connection of the Rossmann fold is not required for catalysis of aminoacylation (Fig. 7).
574 (iv) From that we also discovered how to configure bidirectional urzyme genes (Fig. 9).

575 Each of these new clues would have made it easier for nature to find a reflexive set of protein assignment
576 catalysts to administer the coding table. A coherence of sequence and function unites the data in Figs. 1-5.
577 That coherence (Fig. 6) suggests that although we were certainly on the right track working exclusively
578 from superposing tertiary structures to infer ancestral AARS properties. *E. coli* has created similar
579 catalytically active Class I urzymes that are different from those we had designed computationally and
580 characterized experimentally. In fact, it has shown us how we can do an even better job. The full
581 significance of Fig. 6 is that we were sampling what turns out to be a much broader functional landscape.

582 **New evidence for the modular origin of protein-based urzymes for both AARS Classes**

583 The three new ORFs created *in vivo* complement previously reported hypotheses about the origins of
584 protein-based AARS. We can now infer five new aspects about how amino acid activation and
585 aminoacylation first emerged from the two Class I active-site signatures. (i) They have distinct

586 functionality–ATP utilization (Martinez-Rodriguez, et al. 2015) and substrate RNA recognition (Fig. 8).
587 (ii) They act somewhat independently because they can be linked together differently without disrupting
588 function. (iii) The protozyme and C-terminal fragments (Fig. 8) suffice to catalyze both reactions needed
589 for codon-directed protein synthesis. (iv) They need not even contain the complex amino acid side chains
590 by which they first were identified. Finally, (v) we can now see how genes for Class I and II AARS urzymes
591 can be aligned on opposite strands of a single bidirectional gene (Fig. 9).

592 The longer-term impact for the study of the origin of genetics is hard to underestimate. These results,
593 together with previous work, provide new and broadly-based support for a model for the emergence of the
594 first Class I protein AARS. We envision assembly from three peptide fragments, each about 25 residues
595 long. There is significant evidence for exceptionally high codon middle-base pairing between antiparallel
596 alignments of each fragment. Bidirectional ancestral genes could thus have provided each fragment,
597 establishing a basis for the joint origin of aminoacylating urzymes of both AARS classes (Fig. 10).

598 One fragment (sand background) provides an ATP binding site. The new result that both the half-protozyme
599 and LeuAC-like ORFs retain this fragment unexpectedly supports its key role. To date we have only the
600 structural evidence in the molecular cartoons that this fragment actually binds ATP. It does, however,
601 contain the highly conserved ATP binding surfaces found in crystal structures from both Classes.

602 We proposed the half-protozyme as a progenitor of the Class I AARS protozyme (blue background) via
603 gene duplication and fusion to form an inverted repeat (Chandrasekaran, et al. 2013). Protozymes with both
604 contemporary and bidirectionally coded sequences accelerate amino acid activation by 10^6-fold. The third
605 fragment (magenta background) endows the protozyme with the additional ability to recognize RNA
606 substrates and catalyze acyl-transfer. It is described here for the first time (Fig. 8) showing the similarity
607 between polypeptide:RNA interactions in the two AARS Classes. It appears to be both necessary and
608 sufficient to enable the protozyme to catalyze aminoacylation (Fig. 6).

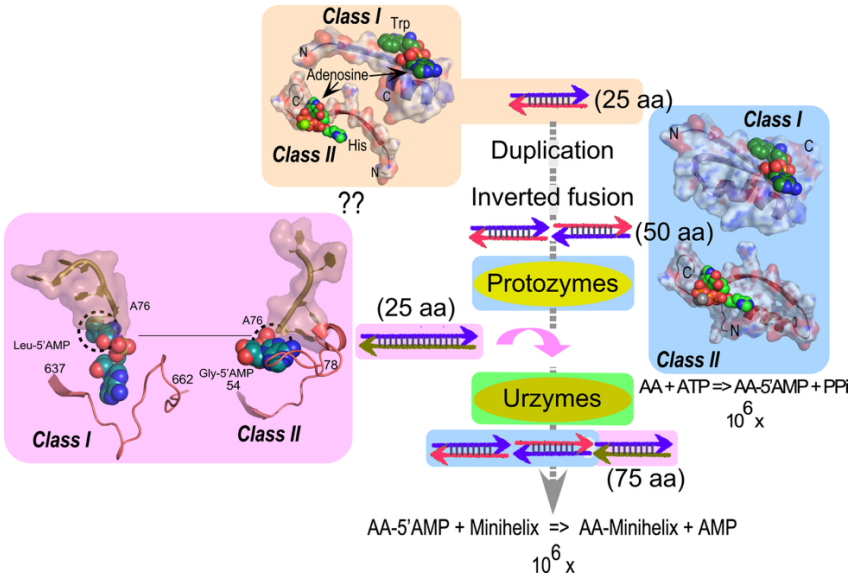


Figure 10. Key steps in the emergence of protein AARS. *In vivo* fragmentation of the LeuRS gene described here provides overlapping reinforcement for the idea that amino acid activating and RNA aminoacylation activities can be attributed to three polypeptide fragments each about 25 amino acids long highlighted by sand, blue, and magenta backgrounds, as described in the text. The two reactions catalyzed by all AARS are indicated with their respective rate accelerations, as shown in Figs. 3,4.

Differences between the three *in vivo* ORFs described here and the corresponding species we created *in vitro* may allow more precise specification of the actual modularity of ancestral AARS. To that end, it will be of interest to deepen understanding of the underlying *in vivo* processes. We plan to follow the distribution of plasmid sizes and the functionality of their apparent coding regions as a function of time following transformation and at different levels of induced expression.

Profiling short plasmids accompanying cloned mutations may help access other primordial genes

A central question about the origin and early evolution of translation is how to connect the much shorter coding sequences of ancestral AARS quasispecies to conventional ancestral sequence reconstructions. The latter are currently limited to essentially full-length species that had already been assembled by the time of the last universal common ancestor, LUCA (Fournier, et al. 2011; Fournier and Alm 2015). It seems likely that some process or processes added new genetic material as modules of significant size. How did AARS genes undergo that modular growth?

We argue that the LeuRS ORFs also signify a new and potentially general approach to the discovery of ancestral genes, both for AARS and perhaps also for other important protein families. Our argument is

628 based on previously published evidence that appropriate double mutants disrupt coupling networks in the
629 intact LeuRS sufficiently to cause metabolic damage. Although pET-11a should minimize expression in
630 the absence of inducer IPTG, it is well-known that variable amounts of leakage invariably occur. For
631 example, few, if any of the wide range of plasmids with mutant AARS we have made in that vector can be
632 maintained as transformed cultures. Expression invariably requires fresh transformation.

633 It is remarkable that two of the three LeuRS ORFs preserve both AARS catalytic activities, albeit at greatly
634 reduced levels. The simple coherence of the LeuRS ORFs (Fig. 9) suggest that some differential fitness
635 may have led to their survival. We consider several questions in turn.

636 *Why do so few variations survive competition with the full-length LeuRS double mutant?* The rate
637 enhancement for tRNA aminoacylation by the double mutant is nearly an order of magnitude less than that
638 for WT LeuRS (Tang, et al. 2024). That loss of function would mean that the WT enzyme present in the
639 transformed cells should produce adequate cellular Leucyl-tRNA^{Leu}. Loss of both domains only when both
640 active site loci are mutated suggests that the double mutation causes a toxic gain of function. Active-site
641 mutations increased acylation with isoleucine by LeuRS (Tang, et al. 2023; Tang, et al. 2024). Similarly,
642 TrpRS mutants lacking either CP1 or the ABD increased activation of tyrosine (Li and Carter 2013;
643 Weinreb, et al. 2014). Mischarging of tRNA is also considered a likely possible mechanism for causing
644 genetic disease by mutant mitochondrial AARS (Meyer-Schuman and Antonellis 2017). Mischarging of
645 tRNA by the double mutant full-length LeuRS is thus a possible culprit.

646 *Why are both CP1 and ABD domains deleted entirely in all shortened plasmids?* It seems significant to us
647 that the connecting peptide (CP1) and anticodon-binding (ABD) domains are both entirely missing in all
648 deletions. Our earlier work on TrpRS (Weinreb, et al. 2012a; Li and Carter 2013; Weinreb, et al. 2014;
649 Chandrasekaran, et al. 2016; Carter, et al. 2017; Chandrasekaran and Carter 2017) and LeuRS (Tang, et al.
650 2023) quantified an essential energetic coupling of the functional contribution of these two domains to
651 catalysis and specificity to the active-site histidine and lysine residues. The four catalytic side chains
652 increase both the rates and specificities of both Class I AARS by five orders of magnitude. However, the

653 functional contributions of the catalytic side chains occurs if and only if both domains are present (Li and
654 Carter 2013).

655 That linkage between the active-site and the two domains appears to be mediated by two different aspects
656 of the tertiary structure in the full-length proteins. The first is a switching interaction within the urzyme.
657 We identified (Kapustina, et al. 2007) and analyzed (Weinreb, et al. 2012a, b) that switching interaction in
658 considerable detail in TrpRS. The second part of the coupling is a packing interaction between non-polar
659 side chains in the ABD that provides "sockets" for the unique non-polar side chains, H**V**GH and K**M**SKS,
660 in each catalytic signature (Tang, et al. 2023). Its role is to assure that the active site is sensitive to the
661 arrangement of the domains. Both aspects appear characteristic of other (Alexander and Schimmel 2001;
662 Budiman, et al. 2007; Li, et al. 2015) and perhaps all Class I AARS.

663 The linkage described in the previous paragraphs provides all-or-none functionality. If either histidine or
664 lysine residues, or if one of the two domain is missing, the substantial rate enhancement and specific
665 substrate recognition are lost. Neither domain can provide its normal function if both active-site loci are
666 mutated. That interdependence suggests why both domains are lost.

667 Mischarging of tRNA is possible toxic gain of function in the present case. It would explain the discrete
668 spectrum of deletions. The ten-fold preference for TΨC minihelix over tRNA (Tang, et al. 2024) means
669 that deletion of both domains to produce urzyme-like species would be expected to reduce mischarging of
670 full-length tRNA 10^7-fold. That also shows that mischarging can be minimized by deleting major parts of
671 the mature protein. Our studies of the Class I AARS mechanistic enzymology thus suggest a rationale for
672 which variants survive.

673 *What mechanisms might have favored the removal of such long segments of genetic information?* The
674 deletions we observe must result from DNA metabolism, because they are inherited in the recovered
675 plasmids. The genomic features summarized earlier (e.g. Fig. 3C and Supplemental Figs. 1-7) suggest
676 multiple exchanges between different segments of coding information. These events likely occur during
677 DNA replication, perhaps by formation of secondary or tertiary structures that promote stem-loop formation

678 (Bikard, et al. 2010). The *E. coli* strain carries a RecA⁻ gene. Any recombination likely entailed other
679 pathways (Ivancic-Bace, et al. 2005).

680 Whatever the mechanism, we note that such processes have been observed in other systems. A notable and
681 functionally significant example is the proposed evolutionary path of the *Drosphophila melanogaster*
682 SLIMP•SerRS heterodimer observed in mitochondria from many metazoans. The SLIMP (SerRS-Like
683 Insect Mitochondrial Protetin) evolved by domain collapse and active-site ablation (de Potter, et al. 2023).

684 The ORFs in Clones 21, on the one hand, and the protozyme-like and Clones 22 on the other hand, are not
685 compatible with successive recombinational events. The former contains the intact protozyme whereas the
686 latter contains only the N-terminal half of the protozyme (Figs. 3A and 6). They must therefore arise
687 independently. These distinctly different configurations of the two active site signatures and the fact that
688 the Goldilocks urzyme (Clone21) appears mostly after prolonged incubation of colony plates at 4° C (Fig.
689 3C) suggest that multiple mechanisms may be required to explain the variety of deletions.

690 Another possible mechanism involves the inadvertent induction of the CRISPR/Cas defense machinery. As
691 noted by Markulin and colleagues (Markulin, et al. 2020) CRISPR systems are activated only below 30° C.
692 None of these arguments provide full explanations of the observations we report here. For example, we
693 cannot explain why we do not see mutations that create pseudogenes that are no longer expressed.

694 *Why might other complex proteins behave similarly?* What we observed may occur when similar
695 experiments are done with a wider range of AARS and other complex multi-domain proteins. Many
696 important protein families—enzymes, motor proteins, regulatory proteins—share a complex multidomain
697 configuration with LeuRS. It has long been apparent that using intramolecular communication to induce
698 functional cooperativity is an emergent property of such proteins (Monod, et al. 1963; Monod, et al. 1965;
699 Changeux and Edelstein 2005; Sadovsky and Yifrach 2007; Zandany, et al. 2008 ; Ben-Abu, et al. 2009 ;
700 Carter 2017). It appears to be an important source of the fitness that supported the (rapid) evolutionary gain
701 of new genetic sequences.

Although the details of coupling differ in many ways from protein family to protein family, it is irrelevant precisely how they function. The specific fragmentation we observed suggests that disrupting internal coupling networks can perturb the fitness of cloned proteins. Other protein families may use coupling mechanisms that can be made cytotoxic by widely spaced active-site mutations. Our experiments show that one way to eliminate cytotoxic effects of cloned variant full-length proteins is to reduce those variants to their simplest functional forms.

Our models for ancestral AARS show that the most highly conserved modules retain substantial catalytic activity. They also reduce putative cytotoxicity. Experiments similar to those we describe here may produce similar sets of much shorter, but still functional variants of many other proteins. The possibility of a wider generality of our protocol could provide a rapid access to urzyme-like variants. The eventual utility of such variants would have evolutionary significance. Such variants could also provide a means to reduce the large size of genes for delivery by viral capsid systems (Hirsch, et al. 2010; Lewis 2014). Protein engineering (Kuhlman and Baker 2004; Romeroa, et al. 2018; Dauparas, et al. 2022; Dauparas, et al. 2023) can further enhance function (Patra, et al. 2025).

DATA AVAILABILITY

Data that are not provided in the supplement will be made available upon reasonable request.

ACKNOWLEDGMENTS

This work was supported by the Matter-to-Life program of the Alfred P Sloan Foundation, Grant G-2021-16944. We are grateful for generous assistance from P. R. Wills and J. Douglas on the alignments in Fig. 8. We also thank H. Fried (Cursor Scientific Editing and Writing) for stylistic advice on the manuscript.

SUPPLEMENTARY INFORMATION

Complete sequence alignments and additional information are provided in the file:
Recombinant_deletions_supplement.pdf

REFERENCESReferences

Alexander R, Schimmel P. 2001. Domain-domain communication in aminoacyl-tRNA synthetases. Prog Nucleic Acid Res Mol Biol. 69:317-349.

Anderson JC, Schultz PG. 2003. Adaptation of an Orthogonal Archaeal Leucyl-tRNA and Synthetase Pair for Four-base, Amber, and Opal Suppression. Biochemistry 42:9598-9608.

Ball P. 2025. Luca is the progenitor of all life on Earth. But its genesis has implications far beyond our planet. In. The Guardian. London, UK: Guardian News & Media Limited.

Ben-Abu Y, Zhou Y, Zilberberg N, Yifrach O. 2009 Inverse coupling in leak and voltage-activated K+ channel gates underlies distinct roles in electrical signaling. Nature Structural & Molecular Biology 16 71-79.

Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, 74:570–588.

Box GEP, Hunter WG, Hunter JS. 1978. Statistics for Experimenters. New York: Wiley Interscience.

Budiman M, Knaggs MH, Fetrow JS, Alexander RW. 2007. Using molecular dynamics to map interaction networks in an aminoacyl-tRNA synthetase. PROTEINS, Structure, Function and Bioinformatics 68:670-689.

Buehner M, Ford GC, Moras D, Olsen KW, Rossmann MG. 1973. D-Glyceraldehyde 3-Phosphate Dehydrogenase: Three Dimensional Structure and Evolutionary Significance. Proc. Nat. Acad. Sci. USA 70:3052-3054.

Burbaum J, Schimmel P. (Minireview co-authors). 1991. Structural Relationships and the Classification of Aminoacyl-tRNA Synthetases. The Journal of Biological Chemistry 266:16965-16968.

Burbaum JJ, Starzyk RM, Schimmel P. (Review Article co-authors). 1990. Understanding Structural Relationships in Proteins of Unsolved Three-Dimensional Structure. PROTEINS: Structure, Function, and Genetics 7:99-111.

Caetano-Anollés G, Wang M, Caetano-Anollés D. 2013. Structural Phylogenomics Retrodicts the Origin of the Genetic Code and Uncovers the Evolutionary Impact of Protein Flexibility. PLoS ONE 8:e72225.

Carter CW, Jr. 2015. What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. Life 5:294-320.

Carter CW, Jr, Tang GQ, Patra SK, Betts L, Dieckhaus H, Kuhlman B, Douglas J, Wills PR, Bouckaert R, Popovic M, Ditzler M. 2025. Structural Enzymology, Phylogenetics, Differentiation, and Symbolic Reflexivity at the Dawn of Biology. Genome Biology and Evolution 17.

Carter CW, Jr, Wills PR. 2019a. Class I and II aminoacyl-tRNA synthetase tRNA groove discrimination created the first synthetase•tRNA cognate pairs and was therefore essential to the origin of genetic coding. IUBMB Life 71:1088–1098.

Carter CW, Jr. 2024. Base Pairing Promoted the Self-Organization of Genetic Coding, Catalysis, and Free-Energy Transduction. MDPI Life 14:199.

Carter CW, Jr. 2017. High-Dimensional Mutant and Modular Thermodynamic Cycles, Molecular Switching, and Free Energy Transduction. Annual Review of Biophysics 46:433-453.

Carter CW, Jr. 2014. Urzymology: Experimental Access to a Key Transition in the Appearance of Enzymes. J. Biol. Chem. 289:30213–30220.

Carter CW, Jr., Carter CW. 1979. Protein Crystallization Using Incomplete Factorial Experiments. Journal of Biological Chemistry 254:12219-12223.

Carter CW, Jr., Chandrasekaran SN, Weinreb V, Li L, Williams T. 2017. Combining multi-mutant and modular thermodynamic cycles to measure energetic coupling networks in enzyme catalysis Structural Dynamics 4:032101.

Carter CW, Jr., Li L, Weinreb V, Collier M, Gonzales-Rivera K, Jimenez-Rodriguez M, Erdogan O, Chandrasekharan SN. 2014. The Rodin-Ohno Hypothesis That Two Enzyme Superfamilies Descended from One Ancestral Gene: An Unlikely Scenario for the Origins of Translation That Will Not Be Dismissed. Biology Direct 9:11.

Carter CW, Jr., Popinga A, Bouckaert R, Wills PR. 2022. Multidimensional Phylogenetic Metrics Identify Class I Aminoacyl-tRNA Synthetase Evolutionary Mosaicity and Inter-modular Coupling. International Journal of Molecular Sciences 23: 1520.

Carter CW, Jr., Wills PR. 2019b. Experimental Solutions to Problems Defining the Origin of Codon-Directed Protein Synthesis. Biosystems 183:103979.

Carter CW, Jr., Wills PR. 2021. The Roots of Genetic Coding in Aminoacyl-tRNA Synthetase Duality Annual Review of Biochemistry 90:349-373.

Chandrasekaran SN, Carter CW, Jr. 2017. Adding torsional interaction terms to the Anisotropic Network Model improves the PATH performance, enabling detailed comparison with experimental rate data Structural Dynamics 4:032103.

Chandrasekaran SN, Das J, Dokholyan NV, Carter CW, Jr. 2016. A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms. Structural Dynamics 3:012101.

Chandrasekaran SN, Yardimci G, Erdogan O, Roach JM, Carter CW, Jr. 2013. Statistical Evaluation of the Rodin-Ohno Hypothesis: Sense/Antisense Coding of Ancestral Class I and II Aminoacyl-tRNA Synthetases. Molecular Biology and Evolution 30:1588-1604.

Changeux J-P, Edelstein SJ. 2005. Allosteric Mechanisms of Signal Transduction. Science 308:1424-1428.

Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, et al. 2022. Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, :49–56 (2022)

Dauparas J, Lee GR, Pecoraro R, An L, Anishchenko I, Glasscock C, Baker D. 2023. Atomic context-conditioned protein sequence design using LigandMPNN. BioRxiv.

de Potter B, Vallee I, Camacho N, P ´ovoas LıFC, Bonsembiante A, Pons APi, Ec khar d U, Gomis-R ¨uth F-X, Yang X-L, Schimmel P, et al. 2023. Domain collapse and active site ablation generate a widespread animal mitochondrial seryl-tRNA synthetase. Nucleic Acids Research 51:10001–10010.

Douglas J, Bouckaert R, Carter CWJ, Wills PR. 2024. Enzymic recognition of amino acids drove the evolution of primordial genetic codes. Nucleic Acids Research 52:558–571.

Douglas J, Cui H, Perona JJ, Vargas-Rodriguez O, Tyynismaa H, Carreño CA, Ling J, Ribas-de-Pouplana L, Yang X-L, Ibba M, et al. 2024. AARS Online: a collaborative database on the structure, function, and evolution of the aminoacyl-tRNA synthetases. Life 4:1091–1105.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Fersht AR, Ashford JS, Bruton CJ, Jakes R, Koch GLE, Hartley BS. 1975. Active Site Titration and Aminoacyl Adenylate Binding Stoichiometry of Amionacyl-tRNA Synthetases. Biochemistry 14:1-4.

Fersht AR, Ashford, J.S., Bruton, C.J., Jakes, R., Koch, G.L.E. and Hartley, B.S. 1975. Active Site Titration and Amino Acyladenylate Binding Stoichiometry of Amino Acyl Synthetases. Biochemistry 14:1-4.

Fournier GP, Alm EJ. 2015. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. J. Mol. Evol. 80:171-185.

Fournier GP, Andam CP, Alm EJ, Gogarten JP. 2011. Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life. Orig Life Evol Biosph 41 621–632

822 Francklyn CS, First EA, Perona JJ, Hou Y-M. 2008. Methods for kinetic and thermodynamic
823 analysis of aminoacyl-tRNA synthetases. Methods 44:100–118.
824 Fritsch EF, Sambrook J, Maniatis T. 1982. Molecular Cloning: A Laboratory Manual. New York.:
825 Cold Spring Harbor Laboratory.
826 Green M, Sambrook J. 2012. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, NY:
827 Cold Spring Harbor Laboratory Press.
828 Hirsch ML, Agbandje-McKenna M, Samulski RJ. 2010. Little Vector, Big Gene Transduction:
829 Fragmented Genome Reassembly of Adeno-associated Virus. Molecular Therapy 18 6-8.
830 Hobson JJ, Li Z, Carter CW, Jr. 2022. A leucyl-tRNA synthetase urzyme: authenticity of tRNA
831 Synthetase urzyme catalytic activities and production of a non-canonical product. International
832 Journal of Molecular Sciences 23:4229.
833 Ivancic-Bace I, Salaj-Smic E, Brcic-Kostic K. 2005. Effects of recJ, recQ, and recFOR Mutations on
834 Recombination in Nuclease-Deficient recB recD Double Mutants of Escherichia coli. Journal of
835 Bacteriology 187:1350–1356.
836 Johnson KA. 2019. New standards for collecting and fitting steady state kinetic data. Beilstein J. Org.
837 Chem. 15:16-29.
838 Kapustina M, Weinreb V, Li L, Kuhlman B, Carter CW, Jr. 2007. A Conformational Transition
839 State Accompanies Tryptophan Activation by *B. stearothermphilus* Tryptophanyl-tRNA Synthetase.
840 Structure 15:1272-1284.
841 Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment,
842 interactive sequence choice and visualization. Brief. Bioinform. 20:1160-1166.
843 Kuhlman B, Baker D. 2004. Exploring folding free energy landscapes using computational protein
844 design. Current Opinion in Structural Biology 14:89-95.
845 Lewis R. 2014. Gene Therapy's Second Act. Scientific American 310:52-57.
846 Li L, Carter CW, Jr. 2013. Full Implementation of the Genetic Code by Tryptophanyl-tRNA
847 Synthetase Requires Intermodular Coupling. J. Biol. Chem. 288:34736–34745.
848 Li L, Francklyn C, Carter CW, Jr. 2013. Aminoacylating Urzymes Challenge the RNA World
849 Hypothesis. J. Biol. Chem. 288:26856-26863.
850 Li L, Weinreb V, Francklyn C, Carter CW, Jr. 2011. Histidyl-tRNA Synthetase Urzymes: Class I
851 and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for Cognate
852 Amino Acid Activation. J. Biol. Chem. 286:10387-10395.
853 Li R, Macnamara LM, Leuchter JD, Alexander RW, Cho SS. 2015. MD Simulations of tRNA and
854 Aminoacyl-tRNA Synthetases: Dynamics, Folding, Binding, and Allostery. Int. J. Mol. Sci. 16:15872-
855 15902.
856 Markulin D, Miti´c D, Radov´ci´c M, Ivan´co´c-Bace I. 2020. CRISPR-Cas in Escherichia coli:
857 regulation by H-NS, LeuO and temperature. PERIODICUM BIOLOGORUM UDC 57:61 121-
858 122:155–160.
859 Martinez-Rodriguez L, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, Weinreb V,
860 Chandrasekaran SN, Collier M, Ambroggio X, Kuhlman B, et al. 2015. Functional Class I and II
861 Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. J. Biol.
862 Chem. 290:19710–19725.
863 Meyer-Schuman R, Antonellis A. 2017. **Emerging mechanisms of aminoacyl-tRNA synthetase**
864 **mutations in recessive and dominant human disease**. Human Molecular Genetics 26, :R114–R127.
865 Michel B, Sandler SJ. 2017. Replication Restart in Bacteria. Journal of Bacteriology 199:e00102-
866 00117
867 Monod J, Changeux J-P, Jacob F. 1963. Allosteric Proteins and Cellular Control Systems. Journal of
868 Molecular Biology 6: 306-329.
869 Monod J, Wyman J, Changeux J-P. 1965. On the nature of allosteric transitions: A plausible model.
870 Journal of Molecular Biology 12:88-118.

871 Moody ER, Álvarez-Carretero S, Mahendrarajah TA, Clark JW, Betts HC, Dombrowski N, Szánthó
872 LL, Boyle RA, Daines S, Chen X, et al. 2024. The nature of the last universal common ancestor and
873 its impact on the early Earth system. Nature Evolution and Ecology 8:1654-1666.
874 Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of
875 DNA in vitro: the polymerase chain reaction. Cold Spring Harbor Symposia On Quantitative Biology
876 51:263-273.
877 Onodera K, Suganuma N, Takano H, Sugita Y, Shoji T, Minobe A, Yamaki N, Otsuka R, Mutsuro-
878 Aoki H, Umehara T, Tamura K. 2021. Amino acid activation analysis of primitive aminoacyl-tRNA
879 synthetases encoded by both strands of a single gene using the malachite green assay. Biosystems
880 208:104481.
881 Patra SK, Betts L, Tang GQ, Douglas J, Wills PR, Bouckeart R, Carter CW, Jr. . 2024. A genomic
882 database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class
883 II urzymes. Nucleic Acids Research 52:13305-13324.
884 Patra SK, Randolph N, Dieckhaus H, Kuhlman B, Douglas J, Carter CW, Jr. 2025. Aminoacyl-
885 tRNA synthetase urzymes optimized by deep
886 learning behave as a quasispecies. Structural Dynamics 12:024701.
887 Pham Y, Kuhlman B, Butterfoss GL, Hu H, Weinreb V, Carter CW, Jr. 2010. Tryptophanyl-tRNA
888 synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular
889 complementation. J. Biol. Chem. 285:38590-38601.
890 Pham Y, Li L, Kim A, Weinreb V, Butterfoss G, Kuhlman B, Carter CW, Jr. 2007. A Minimal
891 TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA
892 Synthetases. Molecular Cell 25:851-862.
893 Picard V, Ersdal-Badju E, Lu A, Bock SC. 1994. A rapid and efficient one-tube PCR-based
894 mutagenesis technique using Pfu DNA polymerase. Nucleic Acids Research 22:2587-2591.
895 Richardson JS. 1973. Handedness of Crossover Connections in B Sheets. Proc. Nat. Acad. Sci. USA
896 73:2619-1623.
897 Rodin SN, Ohno S. 1995. Two Types of Aminoacyl-tRNA Synthetases Could be Originally Encoded
898 by Complementary Strands of the Same Nucleic Acid. Origins of Life and Evolution of the Biosphere
899 25:565-589.
900 Romeroa MLR, Yang F, Lind Y-R, Toth-Petroczy A, Berezovsky IN, Goncearenco A, Yang W,
901 Wellner A, Kumar-Deshmukh F, Sharon M, et al. 2018. Simple yet functional phosphate-loop
902 proteins. PNAS 115:E11943-E11950.
903 Sadovsky E, Yifrach O. 2007. Principles underlying energetic coupling along an allosteric
904 communication trajectory of a voltage-activated K channel. Proceedings of the National Academy of
905 Sciences, USA 104:19813-19818.
906 Schimmel P, Alexander R. 1998. Diverse RNA substrates for aminoacylation: Clues to origins? Proc.
907 Nat. Acad. Sci. USA 95:10351-10353.
908 Tang GQ, Carter CW, Jr. 2025. *Escherichia coli deletes in vivo* the same domains from a double-
909 mutant leucyl-tRNA synthetase gene that were deleted *in vitro* to make the LeuAC urzyme. BioRxiv
910 doi: https://doi.org/10.1101/2025.04.05.647346.
911 Tang GQ, Elder JJH, Douglas J, Carter CW, Jr, . 2023. Domain Acquisition by Class I Aminoacyl-
912 tRNA Synthetase Urzymes Coordinated the Catalytic Functions of HVGH and KMSKS Motifs.
913 Nucleic Acids Research 51:8070-8084.
914 Tang GQ, Hu H, Douglas J, Carter CW, Jr. 2024. Primordial aminoacyl-tRNA synthetases preferred
915 tRNA minihelix substrates over full-length tRNA. Nucleic Acids Research 52:7096-7111.
916 QuikChange II site-directed mutagenesis kit. [Internet]. Available from https://www.agilent.com.
917 Tomoiaga D, Bubnell J, Herndon L, Feinstein P. 2021. High rates of plasmid cotransformation in E.
918 coli overturn the clonality myth and reveal colony development. doi: 10.1101/2021.03.19.434223
919 Varani G. 2015. Twenty years of RNA: the discovery of microRNAs. RNA

920 21:751–752.
921 Wang M, Aziz MF, Caetano-Anollés G. 2025. Tracing the origin of the genetic code and
922 thermostability to dipeptide sequences in proteomes. Journal of Molecular Biology In Press.
923 Weinreb V, Carter CW, Jr. 2008. Mg^{2+}-free *B. stearothermophilus* Tryptophanyl-tRNA Synthetase
924 Activates Tryptophan With a Major Fraction of the Overall Rate Enhancement. Journal of the
925 American Chemical Society 130:1488-1494.
926 Weinreb V, Li L, Carter CW, Jr. 2012a. A Master Switch Couples Mg^{2+}-Assisted Catalysis to
927 Domain Motion in B. stearothermophilus Tryptophanyl-tRNA Synthetase. Structure 20:128-138.
928 Weinreb V, Li L, Carter CW, Jr. 2012b. A Master Switch Couples Mg^{2+}-Assisted Catalysis to
929 Domain Motion in B. stearothermophilus Tryptophanyl-tRNA Synthetase. In. Structure. p. 128-138SI.
930 Weinreb V, Li L, Chandrasekaran SN, Koehl P, Delarue M, Carter CW, Jr 2014. Enhanced Amino
931 Acid Selection in Fully-Evolved Tryptophanyl-tRNA Synthetase, Relative to its Urzyme, Requires
932 Domain Movement Sensed by the D1 Switch, a Remote, Dynamic Packing Motif J. Biol. Chem.
933 289:4367-4376.
934 Weinreb V, Li L, Kaguni LS, Campbell CL, Carter CW, Jr. 2009. Mg^{2+}-Assisted Catalysis by *B.*
935 *stearothermophilus* TrpRS is Promoted by Allosteric Effects. Structure 17:952-964.
936 Williams T, Yin WY, Carter Jr. CW. 2015. Selective Inhibition of Bacterial Tryptophanyl-tRNA
937 Synthetases by Indolmycin is Mechanism Based. Journal Biologial Chemistry 291:255–265.
938 Zandany N, Ovadia M, Orr I, Yifrach O. 2008 Direct analysis of cooperativity in multisubunit
939 allosteric proteins. Proc Nat. Acad. Sci, USA 105 11697–11702.

940

941