Structural Enzymology, Phylogenetics, Differentiation, and Symbolic Reflexivity at the Dawn of Biology

Charles W. Carter, Jr[1]*, Guo Qing Tang[1], Sourav Kumar Patra[1], Laurie Betts[1], Henry Dieckhaus[1,2], Brian Kuhlman[1,3], Jordan Douglas[4,5], Peter R. Wills[4], Remco Bouckaert[5], Milena Popovic[6], and Mark A. Ditzler[6]

[1]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260; [2]Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC [3]Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC [4]Department of Physics, Auckland University, Auckland, NZ; [5]Department of Computer Science, Auckland University, Auckland, NZ; [6]Blue Marble Space Institute of Science, Ames, CA; [7]Center for the Emergence of Life, NASA Ames Research Center, Moffett Field, CA

*Author for Correspondence> Charles W. Carter, Jr. Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill. Telephone: 919 259 2558. FAX: (919) 843-3328**.** Email: carter@med.unc.edu

ORCID IDs:

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

S.K.P: 0000-0002-9314-5641

L.B: 0000-0001-7594-3028

H.D: 0000-0003-1390-2444

B.K: 0000-0003-4907-9699

J.D: 0000-0003-0371-9961

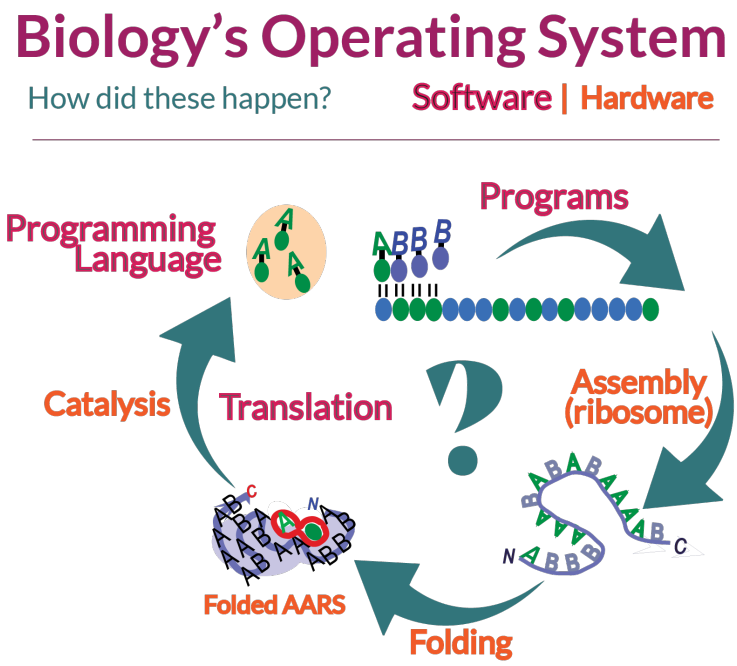
P.R.W: 0000-0002-2670-7624

M.P: 0009-0007-9914-5558

M:D: 0000-0003-3108-4596

Keywords: Origin of genetic coding; Aminoacyl-tRNA synthetase•RNA cognate pairs; Bidirectional genetic coding; ancestral genes; substrate specificity; phylogenetics

31 **Graphical Abstract**
32



33
34
35

Abstract

The reflexive translation of symbols in one chemical language to another defined genetics. Yet, the co-linearity of codons and amino acids is so commonplace an idea that few even ask how it arose. Readout is done by two distinct sets of proteins, called aminoacyl-tRNA synthetases (AARS). AARS must enforce the rules first used to assemble themselves. The roots of translation lie in experimentally testing the structural codes that the earliest AARS•tRNA cognate pairs used to recognize both amino acid and RNA substrates. We review here new results on five different facets of that problem. (i) The surfaces of structures coded by opposite strands of the same gene have opposite polarities. The corresponding proteins then fold up "inside out" relative to one another. The inversion symmetry of base pairing thus projects into the proteome. That leads in turn to contrasting amino acid and RNA substrate binding modes. (ii) *E. coli* reproduces *in vivo* the nested hierarchy of active excerpts we had designed as models—protozymes and urzymes—for ancestral AARS. (iii) A third novel deletion produced *in vivo* and a new Class II urzyme suggest how to design bidirectional urzyme genes. (iv) Codon middle-base pairing provides a basis to constrain Class I and II AARS family trees. (v) AARS urzymes acylate Class-specific subsets of an RNA library, showing RNA substrate specificity for the first time. Four new phylogenetic routines augment these results to compose a viable platform for experimental study of the origins of genetic coding.

Significance Statement

The origin of genetic coding poses questions distinct from those faced studying the evolution of enzymes since the first cells. Modern enzymes that translate the code range in size from ~330 to ~970 amino acids. Ancestral forms cannot have been nearly as complex. Moreover, such primitive enzymes likely could enforce only a much-reduced coding alphabet. Structural and molecular biology data point to a broad sketch of events leading to the code. That research platform will enable us to see how Nature came to store information about the physical chemistry of amino acids in the coding table. That, in turn, allowed searching of a very broad amino acid sequence space. Selection could then learn how to assemble amino acids into functional, reflexive catalysts. Those catalysts had rates and fidelities consistent with bootstrapping the modern coding alphabet. New phylogenetic algorithms need to be developed to fully test that putative sketch experimentally.

Introduction

Genetic coding requires a programming language (the codon table) and a set of programs written in that language to enable it. How did physical chemistry enable discovery about 4Ga ago of the first genetic coding rules from among so many others? Those rules were needed to write blueprints into the sequences of the first genes whose translated produced could then impose the same rules. The double helix (Watson and Crick 1953) inspired a series of brilliant studies. These quickly revealed details of the genetic coding table and the key roles of aminoacyl-tRNA synthetase (AARS)•tRNA cognate pairs (Hoagland, et al. 1956; Berg and Ofengand 1958; Nirenberg and Mattaei 1961; Trupin, et al. 1965; Jones and Nirenberg 1966a; Jones and Nirenberg 1966b; Giegé 1972; Ostrem and Berg 1974).

Such rapid successes obscured questions about how the translation machinery might have arisen from the prebiotic world. For decades, answers to such work remained as disconnected theories (Crick 1968; Yarus, et al. 2009). Most (Koonin and Novozhilov 2017), ignored the crucial role played by aminoacyl-tRNA synthetases. As recently as 2000, and faced with very complex indications from the most recent phylogenetic analyses, an authoritative review by experts including specialists in the area of AARS began their conclusion as follows: "*It is unlikely that the aminoacyl-tRNA synthetases played any specific role in the evolution of the genetic code; their evolutions did not shape the codon assignments.*" (Woese, et al. 2000). Our view is just the opposite. Their statement is logically the same as if we were to proclaim that tRNA played no specific role in the evolution of the genetic code. The present-day code relies solely on the specific fidelity of amino acid binding by AARS•tRNA cognate pairs. Neither statement accounts for the way the present-day code relies solely on the specific fidelity of amino acid binding by AARS•tRNA cognate pairs. Both imply that it emerged from a precursory system that vanished without leaving a trace. That, in turn, would distinguish coding from virtually all other aspects of biology.

We chose to address the problem by excerpting model systems from full-length AARS. Those excerpts include urzymes (~120-130 residues) and protozymes (~50 residues). They retain sufficient catalytic proficiencies to enable us still to use enzymology at the limits of what we could retrieve about the past from AARS structural biology (Pham, et al. 2007; Pham, et al. 2010; Li, et al. 2011; Carter 2014; Martinez-Rodriguez, et al. 2015; Carter 2016; Carter and Wills 2019b; Carter 2022). Only the most surprising aspect of that work has been replicated (Onodera, et al. 2021). Nevertheless, we argue that such models for ancestral AARS allow us to venture further into the past, beyond barriers that seemed evident to earlier authors.

Translation requires a unique sequence of events. The first activates amino acids with ATP so that they can join together spontaneously. The second links the amino acid covalently to tRNA. Doing so assigns a set of chemical symbols—the anticodons—to represent each side chain. How did both activities arise at the same time? That question is closely tied to a related one. The readout from genes to catalytic function is done by proteins called aminoacyl-tRNA synthetases (AARS). What taught the ancestral AARS genes to enforce those assignment rules? Indeed, the AARS gene products must read blueprints templated in the nucleotide sequences of their own genes. That AARS can implement the rules first used to assemble themselves is a property we call *reflexivity* (Carter and Wills 2018b). We will understand the origins of reflexivity only when we can describe the earliest AARS•tRNA cognate pairs and the structural codes they used to recognize both amino acid and RNA substrates (Carter and Wills 2021b). Their rates and specificities must also support a cogent narrative that accounts for selection.

Recently, we have been able to form collaborative teams to improve the early model systems and leverage them into a story that matches structural changes to increasing function (Carter and Wills 2018b; Carter and Wills 2018a; Wills and Carter 2018; Carter and Wills 2019a; Carter and Wills 2019b; Wills and Carter 2020; Carter and Wills 2021a, b; Tang, et al. 2023; Carter 2024; Patra, et al. 2024; Tang, et al. 2024). At the same time, we are enhancing those studies with new phylogenetics algorithms that sharpen the synthetase and tRNA family trees (Douglas, Bouckaert, Carter and Wills 2024; Douglas, Bouckaert, Harris, et al. 2024; Douglas, Cui, et al. 2024). From these and other studies we sketch an answer to the rhetorical questions

posed above. Nature had to discover the coding rules by teaching a set of genes how to read their own blueprints. We summarize here a new paradigm that suggests we can now begin to understand how that happened.

That paradigm includes the following elements.

(i) The AARS are the central assignment catalysts necessary, along with their cognate tRNA molecules, to link amino acids specifically to their symbolic representations in all protein-coding genes. The earliest synthetase genes likely encoded Class I and II synthetases on opposite strands (Rodin and Ohno 1995).

(ii) The genetic code assigns nonpolar and polar amino acids to complementary codons. That gives protein sequences from bidirectional AARS genes inverted side-chain polarity (Zull and Smith 1990; Carter 2024). Thus, the encoded Class I and II AARS pairs folded inside-out relative to one another into different 3D structures. That, in turn, leads to different amino acid and RNA substrate binding modes (Carter and Wills 2019a; Carter and Wills 2019b). These data thus argue that initial substrate recognition was rooted in the base pairing between their coding sequences. We reinforce here the explanatory power of this observation (Carter and Wills 2021b).

(iii) AARS urzymes actually prefer to acylate TΨC-minihelices, instead of full-length tRNAs (Tang, et al. 2024). Both protein and RNA components therefore appear to have functioned first as simpler, single domains. Class I and II urzymes retain amino acid specificities consistent with enforcing a code of ~4 letters. Studies of RNA recognition confirm details of the code used by urzymes to recognize cognate minihelices (Carter and Wills 2018a, 2019a).

(iv) Catalysis by urzymes does not need active-site amino acids that were not present at early times (Tang, et al. 2023). Phylogenetics and AI helped identify new AARS "urzymes" from novel sources (Patra, et al. 2024). *E. coli* makes the same nested hierarchy of synthetase protozymes and urzymes *in vivo* that it took two decades of analytical work to deconstruct.

(v) To be successful, further work requires new phylogenetic algorithms to solve unique problems posed by the AARS family trees (Carter, et al. 2022; Douglas, Bouckaert, Carter and Wills 2024; Douglas, Cui, et al. 2024).

(vi) Logic gating shaped the energy landscape on which coding emerged (Carter and Wills 2021a).

These aspects of our ongoing work now make a full platform for exploring the origins of genetics.

Results

Synthetase genes must interpret their own blueprints.

AARS acquired an enchanting natural property that was novel and apparently unique in the universe so far known to us. In a manner akin to the training of artificial neural networks, they learned to read the instructions for their own assembly. That property is called reflexivity (Fig. 1) because it entails self-reference (Hofstadter 1979, 2007). Genetic coding stores the properties of the amino acid side chains—their size and polarity—by matching them to a set of symbols (codons) (Fig. 1A). To implement coding, Nature then put that physical chemistry to use. It discovered a set of gene sequences that initiates the reaction cycle shown by the broad red arrows in Fig. 1B. Peptides with those sequences could fold in water into 3D structures. Cavities in those 3D structures could recognize specific sets of amino acids and their corresponding symbols. They also could speed up forming a chemical bond between them. Remarkably, at the same time Nature also found how to immortalize those sequences as the ancestral AARS genes.

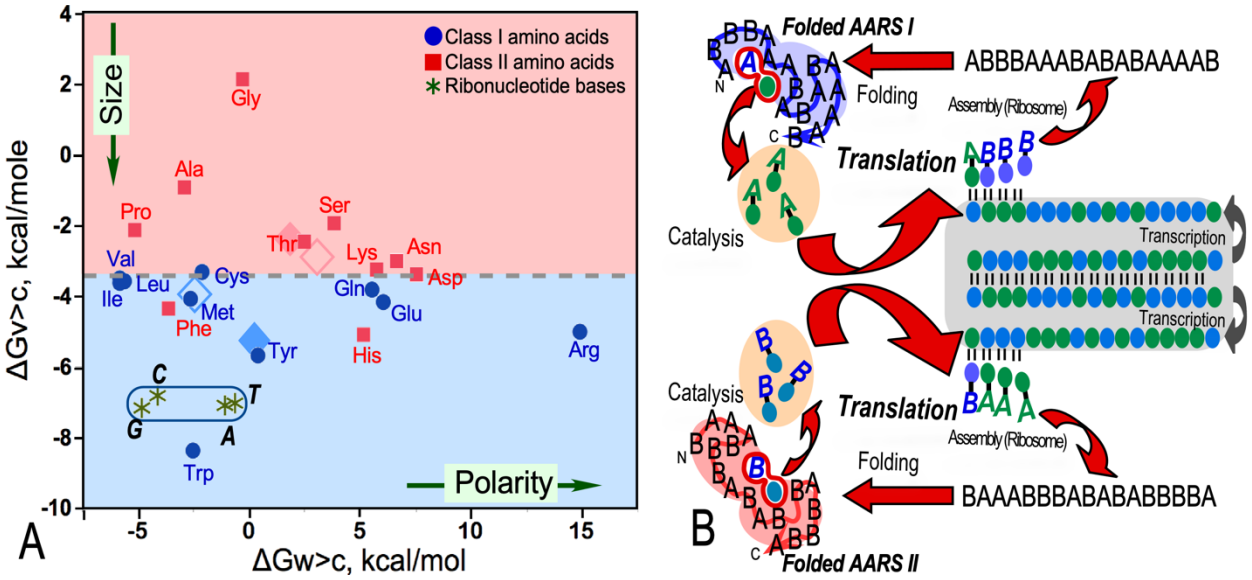


Figure 1. The origins of AARS reflexivity. **A.** Free energies of transfer for amino acids and ribonucleotide bases. These are the building blocks for proteins and nucleic acids. The Y axis is the free energy for transferring the side chain from vapor to cyclohexane. It is thus a surrogate for size. The X axis is the corresponding free energy for transferring the side chain from water to cyclohexane. It is a surrogate for polarity. The plot thus compares the physical chemistry of the nucleic acid and protein alphabets. Class I amino acids are blue dots; Class II amino acids are red squares. The colored background shows that Class I amino acids are predominantly bigger. They also span a larger range of polarity, although most are nonpolar Mean (solid) and median (outline) values for each Class are shown as diamonds of the same color. **B.** Schematic of the role of AARS in the information flow in genetics. A bidirectional ancestral gene and its mRNA transcripts are inside the gray panel. The gene is a bidirectional gene encoding Class I and II AARS on opposite strands. The respective translated peptides are written with a binary alphabet, with amino acids A, B activated respectively by Class I and Class II synthetases. Acylated RNAs are shown as capital letters linked to a green or blue ellipse, representing the symbolic codon representation. A folded conformation is essential for recognition of both amino acid and RNA substrates, and for stabilizing the two transition states for carboxyl group activation and acyl-transfer. Paired cycles of large red arrows define the reflexivity of AARS within each Class. Supplies of building blocks (acylated RNAs within amber ellipses) must be created by the two proteins, which must fold to catalyze the crucial reactions. Selection and gene replication are implicit in the cycle labeled "transcription".

Converting the information in genes into proteins vastly expands it. The standard nucleic acid alphabet has only four letters, and these four letters have almost the same physical properties (Fig. 1A). Their sidechain volumes are both similar and larger than those of all amino acids except tryptophan. Their size greatly reduces the number of ways they can pack into tertiary structures. They also have almost the same polarity. Because both scales in Fig 1A are both logarithmic, the 20 (smaller) amino acids span a vastly greater range of both size and polarity. Thus, the structural and chemical roles of amino acids are far more diverse than those of the four bases. We have estimated on that basis that chemical engineering by proteins is roughly a billion times more diverse and proficient than that by ribozymes (Carter and Wills 2018b; see Supplementary material). For that reason, the invention of coding signals a major advance in life's agency.

AARS come in two different evolutionary Classes (Eriani, et al. 1990). Their respective specificities are shown in Fig. 1A as blue (Class I) and red (Class II). That partition has long puzzled the field. Many have tried to rationalize or explain the partition (Delarue and Moras 1992; Cusack 1994; Rodin and Ohno 1995; Ribas de Pouplana and Schimmel 2001a; Ribas and Schimmel 2001; Klipcan and Safro 2004; Rodin and Rodin 2006; Delarue 2007; Safro and Klipcan 2013; Takénaka and Moras 2020). We consider the class partition to be an essential and penetrating clue to how genetic coding emerged, as outlined in the following section.

The earliest synthetase genes *encoded Class I and II synthetases on opposite strands*.

Rodin and Ohno proposed that the first Class I and II genes arose on opposite strands of the same nucleic acid (Rodin and Ohno 1995). They saw that antiparallel coding sequences for conserved "signature

sequences" defining the two Classes have far more base pairing than expected for random alignments. Three orthogonal kinds of evidence validate the hypothesis: (i) The complementary regions do indeed encode catalytically active synthetases AARS (Carter, et al. 2014; Carter 2015). (ii) Independently reconstructed ancestral Class I and II sequences show increasing base-pairing frequencies as they approach the root of the tree of life (Chandrasekaran, et al. 2013). (iii) Bidirectional genes have been designed and expressed. Translated products from both strands have the same catalytic proficiency, within experimental error, as the native sequences (Martinez-Rodriguez, et al. 2015; Onodera, et al. 2021). Finally, the Rodin-Ohno hypothesis affords a potentially essential metric for constraining the Class I and II evolutionary trees. We show below that the Rodin-Ohno hypothesis has extraordinary retrodictive power.

Two facets of bidirectional coding merit more comment. First, it suggests a contradiction: only one strand of a gene carries unique sequence information. The inversion symmetry of base-pairing means that either strand can be reconstructed if you know the sequence of the other. The prevailing view (Crick 1970) holds for that reason that only one strand does the coding; the other is simply a template for reconstructing the coding strand. How, then, can a bidirectional gene encode two radically different 3D protein structures? We explore that question in the next section. Second, unlike all but one (Wong, et al. 2016; Takénaka and Moras 2020) of the prevailing (Crick 1976; Carter 2008; Yarus, et al. 2009; Koonin and Novozhilov 2017) ideas about origins of coding, the Rodin-Ohno hypothesis serves us metaphorically, as an attractor (Newman 1996). It keeps spawning new, testable predictions. Each corroboration enriches the narrative.

A curious inversion symmetry in the genetic coding table projects base-pairing into the proteome.

Zull and Smith (Zull and Smith 1990) observed that codon assignments are consistent with projection of the nucleic acid inversion symmetry into the proteome (Fig. 2). Genetic codons are assigned to amino acids such that all 14 codons for large, nonpolar residues are anticodons for highly polar residues. Dipeptides that define turns (i.e. Pro-Gly; Gly-Pro) work in both directions on opposite strands. Sequence logos for a set of bidirectional urzyme genes (Fig. 2A) illustrate the variety of sequences enabled by that detail of the coding table. The mean residue by residue transfer free energy for water to cyclohexane partition of amino acid side chains for that alignment, $<\Delta G_{w>chx}>$, reveals a stunning reflection symmetry in antiparallel plots for the 81 amino acids in the designed bidirectional urzyme gene alignments (Fig. 2B). One consequence of that reflection symmetry is that the two signature sequences in each Class (Fig. 2C) behave differently in the folded conformations of Class I and II AARS. The Class I signatures, here HV(I)GH and KMSKS, face into the active-site pocket where they accelerate activation and acylation by factors of 10^5 and 10^3, respectively for the full-length enzymes. The two signatures are linked together only weakly by a nonpolar interaction between the one hydrophobic side chain in each signature. In the absence of the two missing domains, the catalytic contribution of the active site histidine and lysine residues is nearly absent. Their catalytic roles require coupling imposed by domain motions (Tang, et al. 2023). By contrast, the Class II signatures Motifs 1 and 2 form a compact nonpolar packing cluster that provides substantial structural stabilization. In this sense, the ancestral Class I and II synthetase genes folded up inside out.

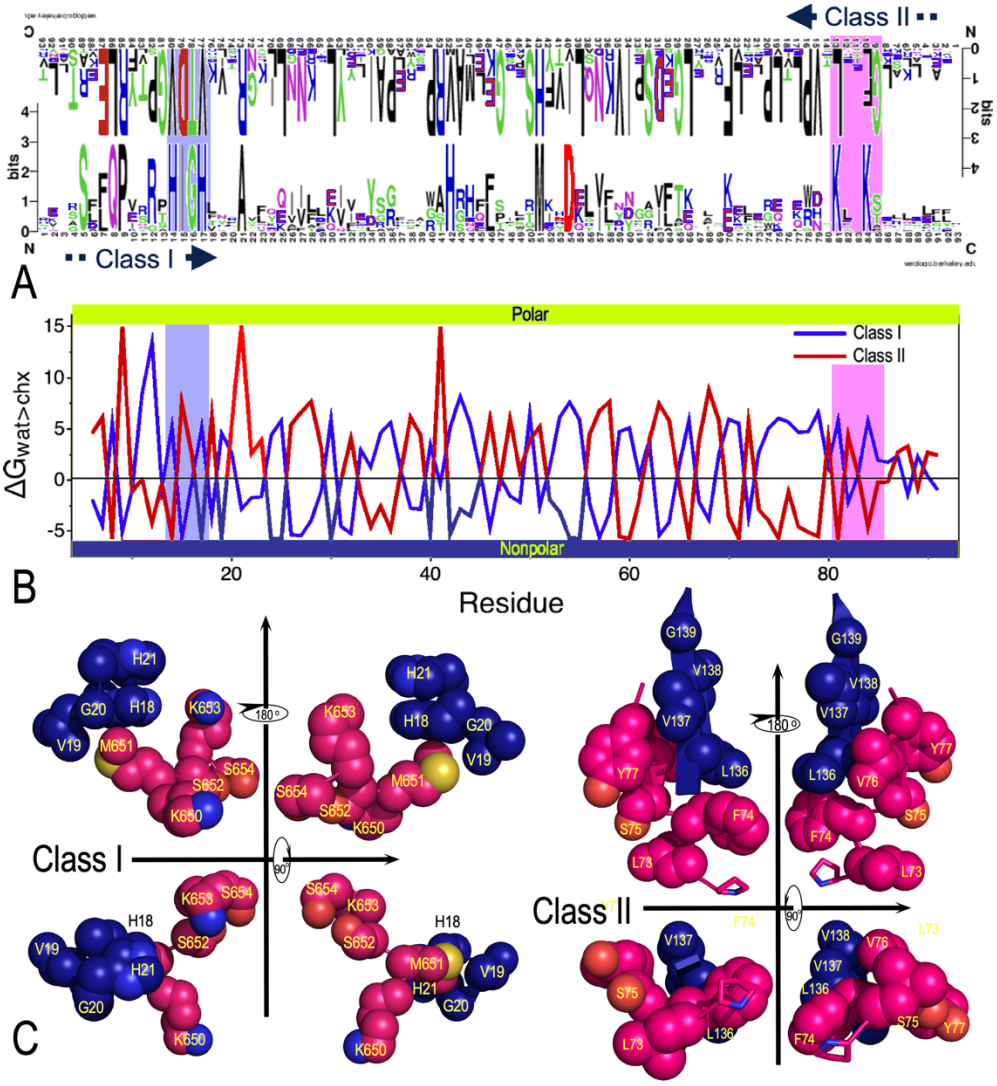


Figure 2. Inversion symmetry of base-pairing projects into the proteome. **A.** Sequence logos (Crooks, et al. 2004) from a set of designed bidirectional genes constrained by the active LeuAC "Goldilocks" and HisCA urzymes both of which have 93 residues. **B.** Plots of $\langle \Delta G_{w>c} \rangle$, the free energy of transfer from water to cyclohexane for each sidechain in the sequence, an experimental metric for the polarity of the sidechain. Sequences of Class I AARS (blue) are plotted; those for Class II (red) are plotted right-to-left. **C.** Consensus Class-defining signature sequences from LeuAC (1WZ2) and HisCA show the impact of the inversion symmetry in the translated proteins. Class I signatures, H(I)GH and K(M)SKS, include mostly polar residues with key catalytic functions. Class II signatures, Motifs 1 and 2, form extensive nonpolar interactions.

*Inside-out folding retrodicts Class I, II amino acid and RNA substrate differentiation***.** Class I, II protozymes provide the AARS ATP binding determinants. When superimposed on their adenosine moieties (Fig. 3A, B), we see that their α-phosphates are prochiral, hence offer non-equivalent binding environments. The amino acids react from opposite sides. Class I side chains, which are always bigger, have more room to grow because they face away from the protein (Fig. 3A). Class II side chains are smaller; they have less room because they would grow into the protein (Fig. 3B).

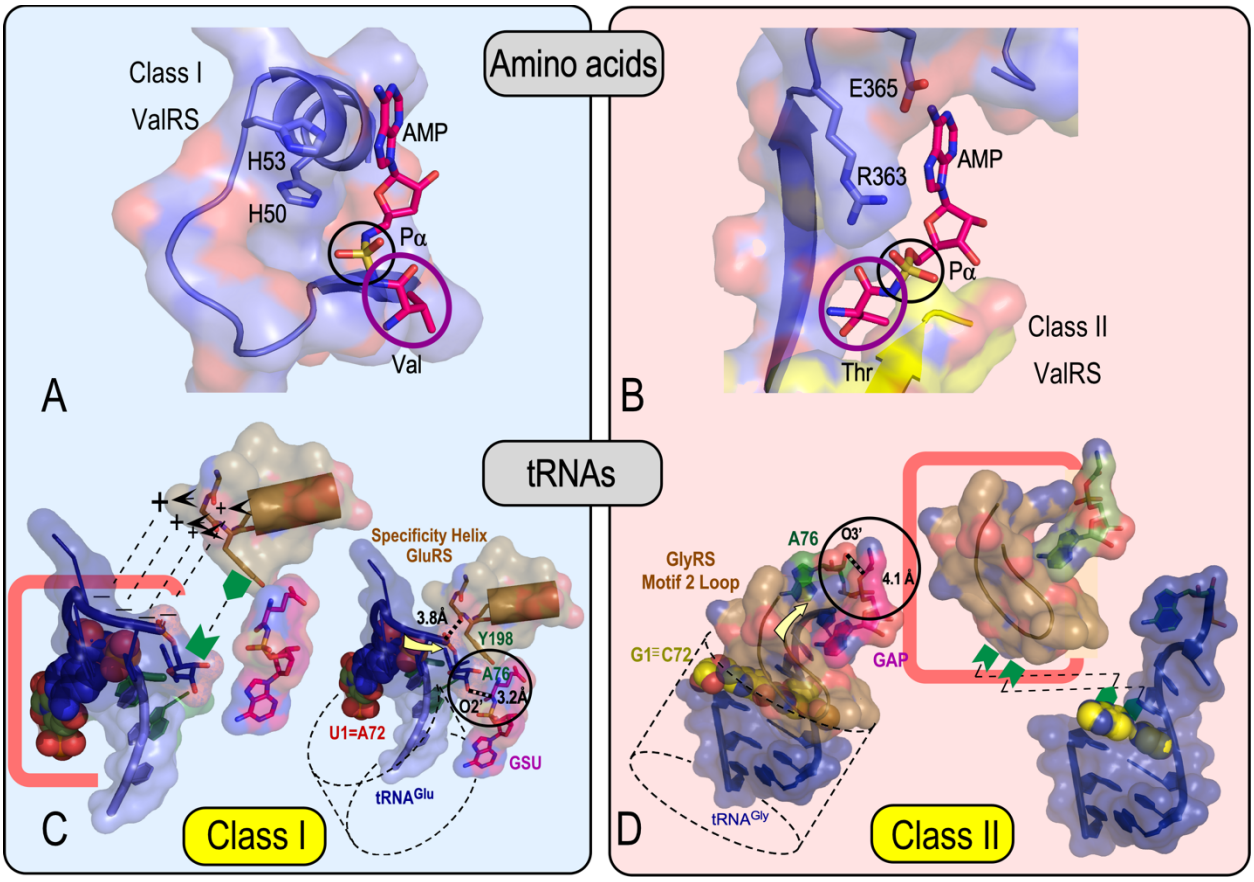


Figure 3. Inside-out folding accounts for both amino acid specificity and tRNA groove recognition. **A, B.** Superposition of the adenosine moiety of Class I ValRS and Class II ThrRS protozymes reveals that amino acid activation sites are diastereroisomeric. The loci of activated amino acids of Valine (A) and Threonine (B) face in opposite directions, accounting for the observation that Class I side chains are uniformly larger than homologous Class II side chains. **C, D.** Class I and II AARS recognize RNA substrates from opposite grooves. The 3'-acceptor stem forms a tight RNA hairpin (red rectangle) in Class I RNA substrates. The N-terminus of the Class I specificity-determining helix forms extensive interactions with that hairpin as indicated by dashed lines between protein positively charged groups and negatively charged phosphate groups in the RNA and nonpolar interactions between a protein aromatic residue and the nonpolar face of the A76 ribose (matching green blocks). Class II RNA 3'-acceptor stems are extended, leaving the initial base pair of the stem to form a platform for recognition by the motif 2 protein hairpin (red rectangle).

Inside-out folding assures that Class I, II synthetases bind to opposite tRNA grooves. A similar asymmetry differentiates tRNA binding by Class I and II AARS. The 3' DCCA terminus of Class I tRNA acceptor stems forms a sharp hairpin. That RNA hairpin is recognized by the N-terminus of the specificity determining α-helix in Class I urzymes. Class II acceptor stems are extended, leaving the initial base-pair as a platform for the Motif 2 protein hairpin. The synthetases for aromatic subclasses Ic (Trp, Tyr) and IIc (Phe) are exceptions (Ribas de Pouplana and Schimmel 2001b).

The duality of amino acid recognition combines with that of the RNA recognition mechanism. The 2D coordinate system in Fig. 3 (A,C vs B,D) underlies the initial discriminations allowing ancestral AARS and their RNA substrates to enforce a rudimentary coding alphabet. The distinctions between the signature sequences in Fig. 2C and the specificities defined in Fig. 3 reinforce the idea that the base-pairing inversion that is fundamental to heredity also projects into the proteome to define substrate recognition by ancestral AARS. We discuss the alphabets that might have grown from that projection in more quantitative detail in a subsequent section.

AARS urzymes and protozymes are more than simple analytical constructs.

We coined the terms "urzyme" and "protozyme" to suggest that the constructs are models for ancestral AARS forms. That notion implies that they may resemble evolutionary intermediates. We have argued that their catalysis and substrate specificities are appropriate to have helped launch genetics. Despite an ample literature, skepticism remains about their authenticity. Only one other group has tested any of our biochemical work (Onodera, et al. 2021). In light of such lingering doubts, this section describes two recent studies suggesting to us that the nested hierarchies may represent more than meets the eye. One entails an urzyme for a new Class II AARS from an annotated eukaryotic genomic database (Patra, et al. 2024). The other describes how *E. coli* generates a similar nested hierarchy of active segments.

AARS urzymes occur in annotated genomic databases. While assembling the aars.online database (Douglas, Cui, et al. 2024), one of us (JD) noted that the Arctic Fox genome has a gene for the α-subunit from the bacterium *Streptococcus alactolyticus GlyRS-B.* That curious orphan gene is a quirky result either of horizontal gene transfer or contamination in the eukaryotic *V. lagopus* genomic data entry. The GlyRS-B clade is found only in bacteria and chloroplasts. It is an unusual heterotetramer in which $tRNA^{Gly}$ binding is relegated to an idiosyncratic β-subunit with homology not to other AARS, but rather to four other cellular proteins with RNA binding functions (Han, et al. 2023). Only the α-subunit is present in the genome. It is not present in genomes from related foxes.

Two purine base insertions create a potentially functional intron in the annotated gene, separating it into two ORFs (Fig. 4A). AlphaFold2 predicts that ORF1 has the same structure as the smallest Class II synthetase urzyme derived from HisRS. A crystal structure for the *E. coli* GlyRS-B confirms the Alphafold prediction (Han, et al. 2023). We expressed constructs with a shortened ORF1 and combining ORF1 with ORF2 (Fig. 4B.) Both these excerpts accelerate aminoacylation of $tRNA^{Gly}$ (albeit with high K_M values). The GlyCA glycine activation rate acceleration is intermediate between those of Class II HisCA and Class I TrpAC and LeuAC urzymes (Patra, et al. 2024).

Fig. 4C shows zymograms for both constructs with accompanying Coomassie-stained native PAGE gels. These zymograms afford new, convincing visual evidence that the catalytic activity migrates with the major band.

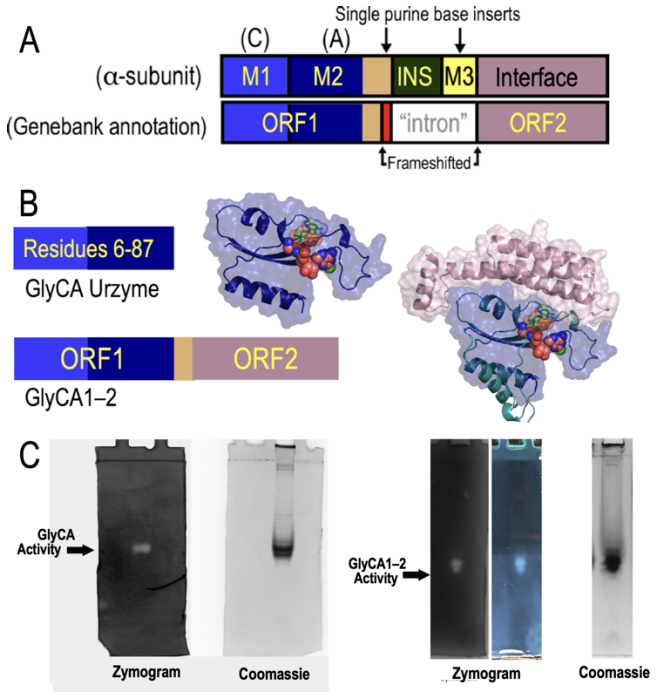
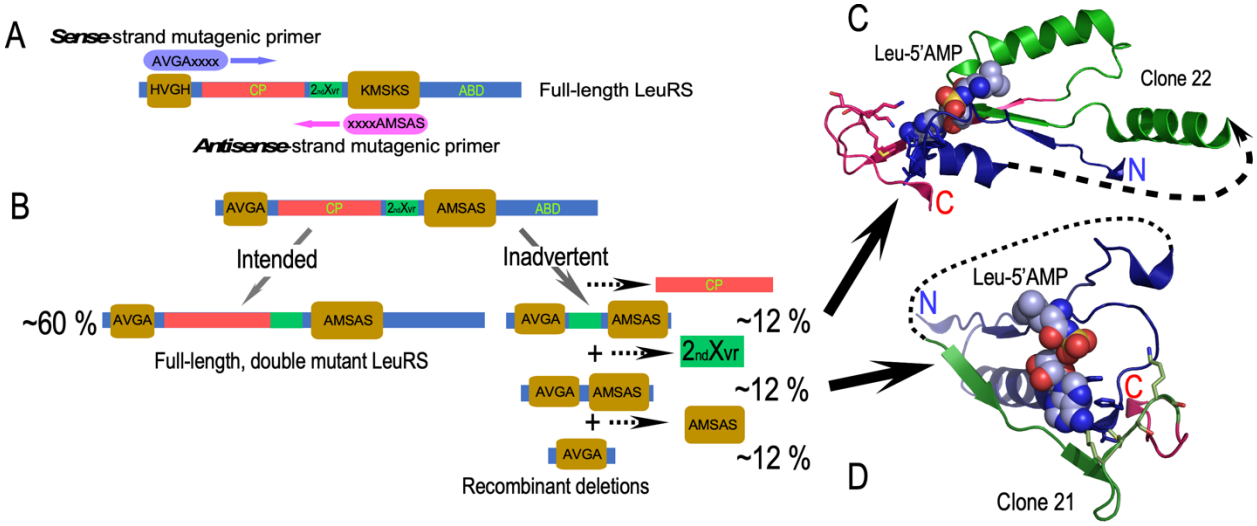


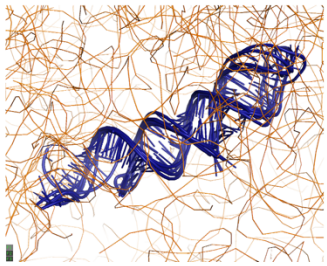
Figure 4. GlyRS urzymes derived from the *V. lagopus* genomic database (Adapted from (Patra, et al. 2024). **A.** Schematic of the annotated genomic entry for the Arctic Fox GlyRS-B α-subunit. Two inserted purine bases (top vertical arrows) create frameshifting (bottom vertical arrows) and an internal stop codon that produce two ORFs. ORF1 ends with a stop codon C-terminal to the red frameshifted sequence. **B.** These differences change the modularity of the *V. lagopus* GlyRS α-chain. 3D structures to the right are AlphaFold2 predictions for both ORFs. Colors are those used in the schematic. AlphaFold2 prediction matches closely that observed in PDB ID 7YSE (Han, et al. 2023). That allows visualization of likely binding geometry for glycine-5 sulfoamyl adenylate (spheres). The annotated intron coding sequence is frame-shifted and corresponds to what is called the insertion domain C-terminal to the active site. Motif 3 and an internal helix-turn-helix motif that covers the terminal adenosine in HisCA2 (Li, et al. 2011) are thus missing in both GlyCA1 and Gly CA1-2. (Created using PYMOL (Pymol 2010)). **C.** Comparisons of a Coomassie stained native gel (right) to a zymogram visualized *în situ* using absorption of the MG orthophosphate complex by GlyCA and GlyCA1–2 (Cestari and Stuart 2013).

E. coli makes in vivo *the same nested hierarchy of synthetase protozymes and urzymes the construction of which required two decades of analytical work*. We designed the LeuAC urzyme by stitching together the conserved cores containing the HIGH and KMSKS signatures and deleting the connecting peptide and anticodon-binding domains. One of us (GQT) mutated the HVGH and KMSKS signatures to AVGA and AMSAS, thereby making a double mutant of full length native LeuRS for use in an earlier publication (Tang, et al. 2023). Only ~60% of the transformed *E. coli* colonies had the expected full-length double mutant plasmid (Fig. 5A). The remaining plasmids all had deletions of almost exactly those variable regions we had deleted *in vitro* (Fig. 5B). Urzymes bearing the two longer deletions catalyze both amino acid activation and aminoacylation. The rest of the recombinant deletions were protozymes. Both recombinants Clone 22 (Fig. 5C) and Clone 21 (Fig. 5D) retain the AMSAS sequence are active leucyl-tRNA synthetase urzymes (Fig. 5E).



E

Variant	ΔG(kcat/KM), kcalmole	Stdev	ΔG(kcat), kcal/Mole	Stdev	ΔG(/KM), kcal/mole	Stdev
Clone21_WT	-5.91	0.42	3.34	0.03	9.25	0.45
Clone21_DBL	-5.69	0.46	3.71	0.05	9.40	0.51
Clone22_WT	-5.56	0.50	3.73	0.04	9.28	0.54
Clone22_DBL	-6.00	0.60	3.14	0.13	9.13	0.73
LeuAC_WT	-2.64	0.28	5.19	0.56	7.84	0.79
LeuAC_DBL	-4.86	0.14	4.00	0.07	8.86	0.20



F

Figure 5. E. coli provides an *in vivo* system to track modular evolution of AARS. **A.** Simplified scheme for mutating both HVGH and KMSKS signatures using bidirectional mutagenic primers. **B**. Schematic representation of plasmids sequenced from single colonies isolated after transforming *E coli* with the full-length double mutant (AVGA/AMSAS) LeuRS plasmid. All recombinant deletions preserve the exact sequences from the input full length double mutant plasmid but are missing extended fragments from the LeuAC urzyme. **C.** Segments of Clone22, the "urzyme-like" deletion. It retains the second crossover connection (2^{nd} Xvr) but

is missing the C-terminal β-strand of the protozyme. It must fold differently from the urzyme because of the deletions, as indicated by the dashed arrow. **D.** Segments of Clone21, the "Goldilocks" urzyme. Deletion of most of the specificity helix and second crossover connection from the Rossmann dinucleotide-binding fold means that this construct folds differently because the green segment containing the KMSKS loop is joined to the protozyme (dashed line). The 3D structures for **C** and **D** are taken from the crystal structure 1WZ2 of the *Pyrococcus horikoshii* LeuRS. **E.** Steady state kinetic parameters of wild-type and double mutant versions of Clone 21 and Clone 22 compared with those previously published for the corresponding LeuAC variants. **F.** Five 3D structures predicted by Alfafold3 (Abramson, et al. 2024) for the mRNA of full-length *P. horikoshii* LeuRS have been aligned by superposing the coding region (dark blue) for the KMSKS-bearing loop (green segment in C). That region is the only motif shared by all five predictions.

The Goldilocks deletion maps to 3D structure in its mRNA. The intermediate deletion is especially interesting because although it retains the AMSAS sequence, it is missing most of the second crossover connection of the Rossmann dinucleotide fold (dashed line in Fig. 5F). We already introduced the descriptor "Goldilocks" for urzymes bearing this deletion because of its intermediate size. Its structure is unknown but must differ substantially from that of the corresponding sequences in the full-length LeuRS (Fig 5D). The Goldilocks deletion is an active, 81 residue synthetase. It represents the addition to the protozyme of the mutant AMSAS signature.

Notably, that short fragment seems to be both necessary and sufficient to enable catalysis of acyl transfer from the activated aminoacyl-5'AMP to a substrate RNA. Further, as noted in Figs. 2A and 7, it is the same length as the Class II HisCA urzyme. That parity appears also to resolve the challenge of designing Rodin-Ohno bidirectional urzyme genes. For these reasons, the Goldilocks urzyme may also represent the minimal evolutionary path from protozyme to urzyme.

The discrete locations and high reproducibility of these deletions strongly suggests that they result from 3D structural properties of the gene itself. We had suggested as much in our initial paper describing the TrpRS urzyme (Pham, et al. 2007). To investigate that possibility, we predicted mRNA structures using AlphaFold3 (Abramson, et al. 2024). AlphaFold3 has only rudimentary capability to predict RNA structures and the absence of many training datasets meant that predicted structures are highly variable. Nonetheless, one long stem-loop was present in each prediction (Fig. 5F). That loop contains the exact codons for the module that converts the LeuRS protozyme into the Goldilocks urzyme. That coincidence is consistent with the hypothesis that the nested recombinant deletions in Fig. 5B, and by implication the synthetase modularity arise from structures within the mRNA that promote recombination. *In vivo* models of this sort could be an invaluable tool to study the enzymology of modular protein evolution (Fig. 5E).

We need to adapt phylogenetics to address the unique problems posed by synthetases.

Synthetases pose four problems for conventional phylogenetics algorithms: Each requires a qualitatively different innovation in phylogenetic software. Our pursuit of more reliable ancestral reconstructions made it necessary to implement solutions to each of these problems. We discuss these briefly in this section.

AARS exhibit numerous, diverse insertions and deletions. Indeed, both superfamilies likely grew to their contemporary form from quite small protozymic ancestors largely by assimilating modular bits of genetic information (Fig. 6). The gradual accumulation of insertion modules on the enzyme surface cannot be treated adequately using amino acid substitution models alone (Whelan and Goldman 2001b; Le and Gascuel 2008). An additional Bayesian probability model is required to accommodate the birth and death of insertion modules over long evolutionary timescales in both AARS superfamilies.

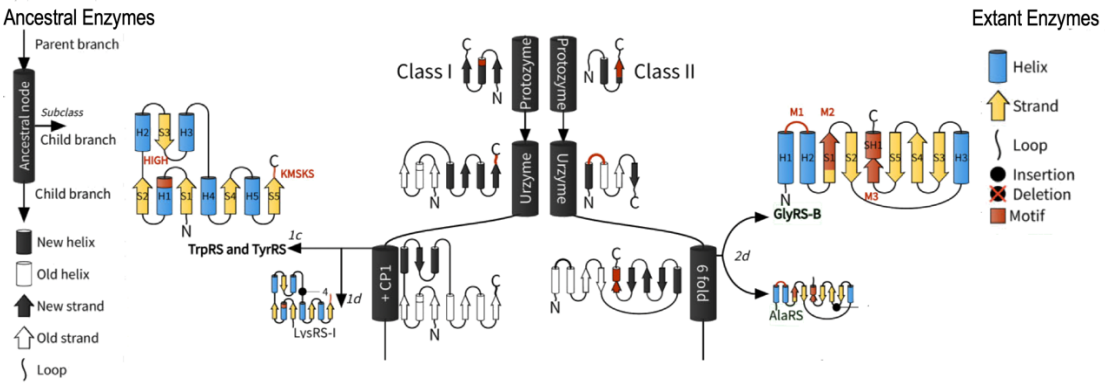


Figure 6. Insertion elements at the root of the AARS family trees [adapted from aars.online (Douglas, Cui, et al. 2024)]. Schematic representation of two extant AARS (color) in the context of putative ancestors (black and white). The succession shown here is truncated to include only the transition from protozyme to urzyme to catalytic domain. The purpose is to illustrate how important modular accretion was to the evolution of both AARS superfamilies. Note the succession converting ancestral (black) to more modern (white) secondary structures. Modular accretion continued to occur throughout the evolution of both superfamilies (see Figure 4 of (Douglas, Bouckaert, Carter and Wills 2024).

We addressed this problem in two stages. First, we created an online database with access to data on secondary and tertiary structure alignments for both superfamilies (Douglas, Cui, et al. 2024). Many authors in the field contributed to this database, which now provides more rigorously curated alignments. That database helped us to identify how indels differ systematically from one synthetase to another. The second stage was to build a Bayesian probability and a Dollo birth/death model for BEAST2 (Douglas, Bouckaert, Carter and Wills 2024).

AARS for the most complex amino acids, tryptophan and tyrosine, resemble the oldest synthetases (Fig. 6). TrpRS and TyrRS appear to be the most recent additions to the Class I superfamily (Ribas de Pouplana, et al. 1996). TrpRS and TyrRS also are the smallest of all AARS, and also have the simplest modularity (Doublié, et al. 1995). Thus, they appear to be most ancient. Occam's razor suggests it is unlikely that earlier versions had additional modules that all were lost in order to enhance their specificity. More likely, a simpler dormant ancestral gene proved a better choice when amounts of the two amino acids increased to the point where it became practical to create a new coding letter, as suggested in Fig. 6. We called such a case "retrofunctionalization" to be compared with "neofunctionalization" and "parafunctionalization" (Lynch and Force 2000; S. Rastogi and D. A. Liberles 2005). Other examples likely occur in both superfamilies.

AARS sequence evolution likely entailed substantial saltation. One of the most problematic aspects of building reliable trees is having to assume a gradualistic clock-like mutation process. Yet, branching often appears to entail multiple correlated changes (Katsnelson, et al. 2019). Localized spikes of evolutionary change have been well-documented experimentally for some more recent protein superfamilies (Bridgham, et al. 2009; Manceau, et al. 2020). Such cases were summarized earlier at the macromolecular level (Eldredge N 1972). Various attempts to solve this problem have had gradually improving success (Bokma 2002; Pagel, et al. 2006; Manceau, et al. 2020).

Consistent probabilistic treatment of rapid changes with birth/death model (Douglas, Bouckaert, Harris, et al. 2024) led to several substantial improvements. We assumed that each lineage experiences a rapid evolutionary spike, whose size is informed by the number of unobserved bifurcations along that lineage (i.e., the number of stubs). Branching and evolution do indeed appear to be tightly coupled. Despite the model's high dimensionality, several results convince us that it has many attractive properties. Most important is enhanced statistical consistency, and a low covariation between rates and spikes. The model consistently recovers correct known models from simulated data.

Unobserved speciation events do seem to have left their footprints on the lineages that have been observed. We hypothesized that abrupt evolution could occur at any level in biology, demonstrating the process in

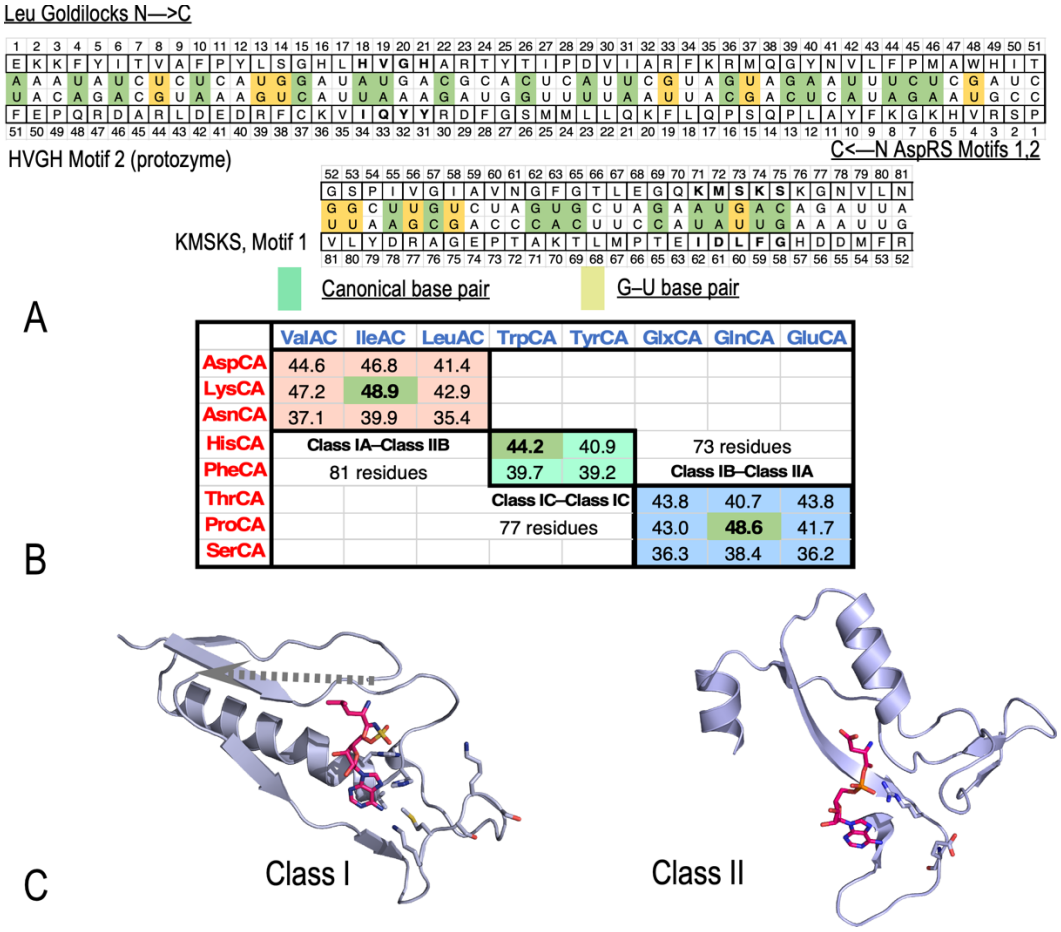
genes, morphologies, and languages (Douglas, Bouckaert, Harris, et al. 2024). Identifying them produced more accurate divergence times. They also resulted in more reliable branchpoints. We demonstrated these enhancements in three detailed case studies spanning the entire spectrum of phylogenies. Inferred AARS molecular, morphological changes in cephalopods, and social evolution of Indo-European cognate word phylogenies all changed differently. In particular, AARS branchpoints are indeed associated with highly elevated fixation of point mutants.

The amino acid substitution matrix must be updated dynamically. Standard amino acid substitution models rely on the critical assumption that the alphabet of amino acids remained constant through time (Whelan and Goldman 2001a; Le and Gascuel 2008). While this assumption might hold for younger proteins (e.g., among plants and animals), the assumption becomes increasingly unjustifiable as the age of the phylogeny approaches the last universal common ancestor, and even more so dealing with the reflexive nature of the AARS (Carter and Wills 2018b; Shore, et al. 2019). This phylogenetic model, under active development, involves the multiplication of a smaller, reduced $(n–1) \times (n–1)$ amino acid substitution matrix near the root of the tree with the full $n \times n$ matrix in the present day, where n is the number of amino acids. Simulation studies can often distinguish subfunctionalization (e.g., IleValRS => IleRS + ValRS) from neofunctionalization (IleRS => IleRS + ValRS).

*Class I, II AARS trees must be coupled***.**

Another basic issue for us is how to link the Class I and II AARS trees. We can make an overwhelming case that even the earliest proteins needed both Class I and II AARS. The *sine qua non* of protein secondary structure is binary patterns of either polarity and/or size (Serrano, et al. 1992; Muñoz and Serrano 1994). Polarity is even more essential for forming 3D structures (Dill and MacCallum 2012; Guseva, et al. 2017). The Rodin-Ohno hypothesis would, if verified, strongly constrain trees for the two superfamilies. We hoped to infer such metric from the base-pairing frequencies of codon middle bases in the ~100 antiparallel alignments of Class I vs Class II coding sequences (Chandrasekaran, et al. 2013). Codon middle-base pairing between important segments of Class I and II urzyme genes could thus measure the strength of that constraint. Assembling the data needed to compile those statistics, however, is hard owing to the numerous subtle indels within and between different AARS families in both Classes.

After puzzling for many years, the Goldilocks LeuAC urzyme (Fig. 5) has told us that the 2nd crossover of the Rossmann fold is not needed for aminoacylation. It was likely a later addition and may not have existed during the brief stage of bidirectional coding. The Goldilocks recombinant provided a crucial datum by revealing that only the short β-strand and KMSKS loop are required for aminoacylation. It has exactly the same length as the recently characterized GlyCA Class II urzyme. That realization reduces our estimate of how long the earliest ancestral AARS were to ~ 80 amino acids. Examples from work in progress are shown in Fig. 7.



	ValAC	IleAC	LeuAC	TrpCA	TyrCA	GlxCA	GlnCA	GluCA
AspCA	44.6	46.8	41.4					
LysCA	47.2	**48.9**	42.9					
AsnCA	37.1	39.9	35.4					
HisCA	Class IA–Class IIB			**44.2**	40.9	73 residues		
PheCA	81 residues			39.7	39.2	Class IB–Class IIA		
ThrCA				Class IC–Class IC		43.8	40.7	43.8
ProCA				77 residues		43.0	**48.6**	41.7
SerCA						36.3	38.4	36.2

Figure 7. Toward a Rodin-Ohno urzyme gene and a metric to link the early evolution of Class I and II AARS evolution. **A.** Antiparallel sequence alignments of designed 81-residue Class I and II Rodon-Ohno urzymes based on Class I *P. horikoshii* LeuRS (1WZ2) and Class II *E. coli* AspRS (1C0A), with codon middle bases interleaved and colored according to base pairing, as indicated. **B.** Partial table of codon middle-base pairing frequencies assembled from diagrams similar to that in **A.** Note in particular the differences in length between the AspRS, GlyRS, and LeuRS urzymes and those for the other four. Also note that the Expected pairing frequency between subclass IA LeuRS and subclass IIA GlyRS is substantially smaller than that with subclass II AspRS. Thus, the subclass pairings suggested by middle base pairing frequencies are not as expected (see central part of the table). **C.** Graphic images of the molecular structures implied by the alignments in **A.** Catalytic residues and both amnoacyladenylate ligands are shown as sticks. The long, dashed arrow in Class I indicates a counterintuitive covalent bond connecting the C-terminal segment to the protozyme.

This preliminary evidence replicates evidence from similar unpublished earlier work done before we found the LeuAC Goldilocks urzyme. It suggests that three ancestral bidirectional genes originally paired Class IA IleRS opposite Class IIB LysRS, Class II ProRS opposite Class I GlnRS, and HisRS opposite TrpRS. \ Completing the partial table in Fig. 7B may constrain the joint evolution of Class I and II AARS enough to allow us to reconstruct ancestral sequences back to the time when urzymes enforced an earlier code of far fewer than 20 amino acids.

These pairings are surprising because they suggest that wholesale sequence changes accompanied the speciation as the earliest genes became increasingly specific AARS. Four of the putative ancestral sequences shown in Fig. 7B resemble most closely those for “phase 2” amino acids—lysine, glutamine, histidine, and tryptophan—that were not used in the earliest proteins {Wong, 2016 #833}. The corresponding GlnAC LysAC, HisAC, and TrpAC urzyme sequences thus all seem to be examples of retrofunctionalization (Douglas, Bouckaert, Carter and Wills 2024).

AARS urzyme•minihelix cognate pairs are sufficiently functional to launch genetic coding.

The phylogenetic studies outlined in the last section can open the door to more robust sequence distributions for AARS families. That is only part of the solution to our problem. Any account of survival will rely on relative fitness. The narrative we seek must also account for the survival of the succession of constructs. Proving fitness is a matter of enzymology. Rates and substrate recognition of each successive stage must be consistent with that narrative. This section recaps experiments that test how well AARS urzyme:minihelix cognate pairs might have functioned.

Enzymological studies confirm that urzyme•minihelix pairs work better than either hybrid. For decades, the tRNA anticodon was assumed to play the pivotal role in tRNA identity. It does, in ribosomal mapping to mRNA. However, a seminal proposal appeared in 1993 that today's dual-domain synthetase•tRNA cognate pairs grew from a much simpler set of single-domain•minihelix pairs (Schimmel, et al. 1993). Urzymes would not be discovered for another decade and a half (Pham, et al. 2007; Li, et al. 2011). It took two and a half decades to work out details of the operational code. It took more than three decades to try to answer the question: can urzymes acylate not only tRNA but also minihelix substrates?

Schimmel's was a prescient idea, fueled by the fact that several groups had shown that full-length AARS could acylate minihelices that lacked the anticodon (Francklyn and Schimmel 1989; Schimmel 1991; Frugier, et al. 1992). The idea that the anticodon and corresponding anticodon-binding domains could have come later also implied that the original basis for tRNA recognition by ancestral synthetases must have depended on a different code embedded in the acceptor stem. The authors proposed no details of how such an operational RNA code might work. Yet, it was the skeleton of the crucial notion that the ancestral coding system might originally have worked with much simpler machines.

For any system to evolve, it has to function. Any ancestral system must have been able to perform two tasks with some degree of substrate specificity: activate amino acids and acylate RNA substrates. It is crucial to know whether or not urzymes can recognize amino acids and acylate tRNA minihelices. We recently answered this question dramatically. Urzymes not only have the capability to acylate cognate minihelices; they actually prefer the minihelix substrate by about tenfold (Fig. 8A).

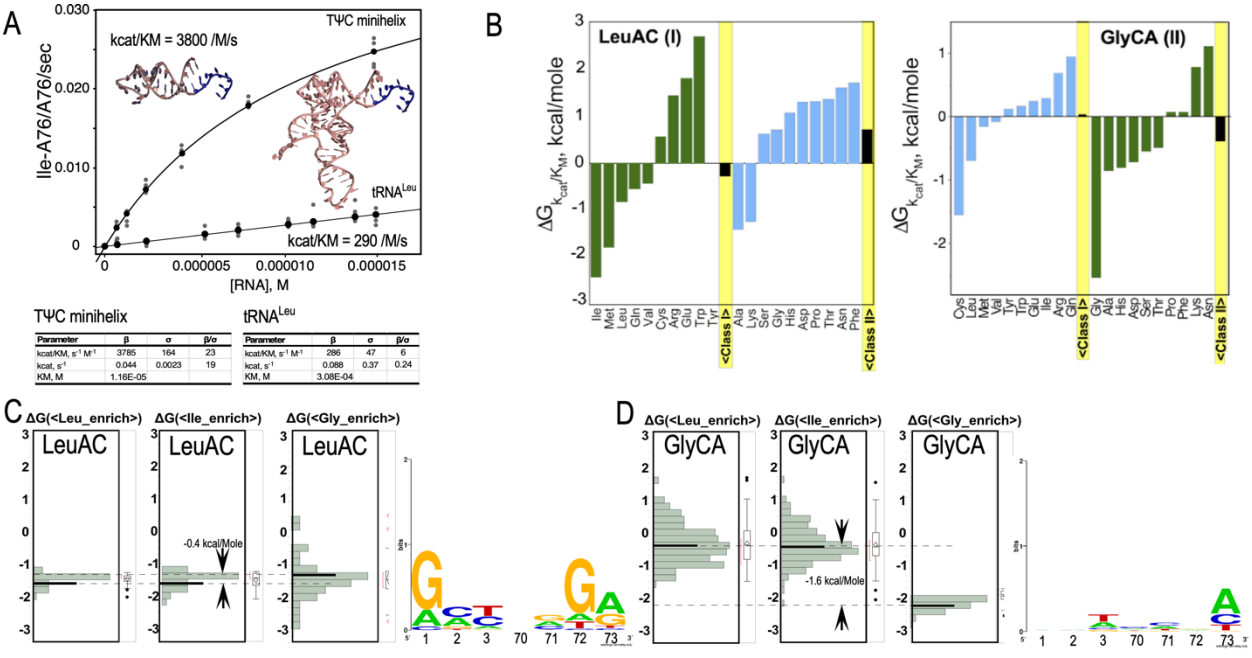


Figure 8. Evidence for the evolutionary fitness of AARS urzyme•minihelix cognate pairs. **A.** Comparison of Michaelis-Menten kinetic data for the LeuAC urzyme acylation of TΨC-minihelix^{Leu} and tRNA^{Leu} [adapted from (Tang, et al. 2024)]. **B.** Amino acid specificity spectra for the LeuAC and GlyCA urzymes [adapted from (Patra, et al. 2024)] **C**. Note in particular that although the

Class I and II Urzymes cannot distinguish amino acids as well as full-length AARS, they tend to favor amino acids from their own Class. **C, D.** RNA specificities of LeuAC and GlyCA urzymes. Histograms for the enrichments of aminoacylated sequences over those inferred from the total library for minihelices acylated with Ile and Leu (**C**; 39 hyperacylated sequences) by AVGA LeuAC mutant and Gly (**D**; 204 hyperacylated sequences) by GlyCA. The vertical axes are expressed as free energy values in kcal/mole, computed on the basis of the enrichment. Black horizontal lines denote the average values. Web Logos compare the RNA specificity of the AVGA LeuAC mutant and GlyCA enriched fractions. The histograms and web logos represent 10-fold (**C**) and 22-fold hyperacylated (**D**) fractions, respectively.

The Michaelis-Menten plot for LeuAC urzyme (Fig. 8A) shows that the Leucine minihelix is an order of magnitude better substrate than $tRNA^{Leu}$. Neither hybrid protein•RNA system works as well. Thus, if the RNA substrate is a minihelix, the best enzyme to acylate it is the cognate urzyme, and conversely. LeuAC has a marked preference the minihelix and LeuRS much prefers $tRNA^{Leu}$. That antisymmetry is unexpected evidence that the two modules co-evolved from the start of genetic coding.

Experimentally, the LeuAC and GlyCA urzymes also differentiate complementary subsets of ~5 similar amino acids (Fig. 8B). The mean free energy difference between the correct and incorrect amino acid Class is about 1 kcal/mole. That means the urzymes mistakenly activate an amino acid from the wrong class about 1 in 5 times. That specificity is comparable to what we observed previously (Carter, et al. 2014).

The same two urzymes also acylate different subsets from a minihelix library (Fig. 8C, D). That library contained all combinations of the seven bases at the top of the acceptor stem. Subsets acylated by LeuAC and GyCA were biotinylated and collected on streptavidin beads. The selected minihelices were then sequenced, as was the un-acylated library itself. As noted in **Methods and Materials**, we used the same library for each urzyme. That library had acceptor stem sequences derived from $tRNA^{Leu}$. For that reason, the two urzymes behave differently. We show in Fig. 8C, D the distributions of the most hyperacylated minihelices from the library, together with web logos for those distributions. The GlyCA urzyme acylated fewer minihelices with greater enrichments. It appears somewhat more selective for minihelices from the library, likely because the constant region of the acceptor stem has some residual specificity.

As with the amino acids, cross-specificity is evident. The mean differences in free energies are ~ –1 kcal/mole for activation by both urzymes for class-specific acylation and –0.4 kcal/mole and –1.6 kcal/mole for minihelix specific aminoacylation by LeuAC and GlyCA.

This experimental platform enables us now to trace the growth of the coding table. As far as we know these are the only data available so far on the likely fidelity of very early translation systems. They raise this new question: Are these rates and fidelities sufficient to compose a self-consistent reflexive system? Answering that question appears now to be within reach. The phylogenetic software set out in the previous section should enable us to estimate which amino acids were in likely ancestral coding tables. They also should allow us to construct reliable ancestral distributions of sequences for the relevant AARS that might have enforced such ancestral codes. We can then construct combinatorial libraries for those very AARS. We have begun to express and characterize such libraries (Patra, S.K. *et. al.* in preparation). We also have written software for treating the kinetics of catalytic mixtures (Douglas, Carter and Wills 2024).

Reciprocally coupled gating suggests the existence of two, new, biological forces.

The reason there are so many different attempts to define the origin of life is that life itself seems so highly improbable. For life to emerge on early post-Hadean earth many, quite different things all had to happen at almost the same time. Translation was only one of the most important and challenging of those things. It could have emerged only under quite special circumstances. Those include a steady input of free energy (Liu, et al. 2020), rudimentary metabolism (Sobotta, et al. 2020; Stubbs, et al. 2020), mutualism relating polypeptides and nucleic acids (Lanier, et al. 2017), and possibly some sort of compartmentation (Zhu, et al. 2013). These areas are outside the scope of this work, but they all must have been provided in some form.

Genetic coding alone seems, at first glance, to be highly improbable. Afterall, within the vast combinatorial space of amino acid-to-codon assignments and gene sequences, only a tiny fraction can enforce reflexive

genetic coding. With more than 10^{84} ways to assign 64 codons to one or more of the 20 amino acids and $>10^{105}$ possible amino acid sequences that are the length of an urzyme, an exhaustive search for suitable subsets of assignments and genes would exceed the resources of the known universe. How then did a suitable combination emerge here on earth?

Such combinatorial problems abound almost anywhere one looks; they are especially common to biology. They can be described as "Levinthal-like" paradoxes because, they share the formulation articulated by Levinthal for protein folding (Levinthal 1968; Dill and Chan 1997). Solutions to the Levinthal paradox take the form of a free energy landscape that adopt the form of a "funnel" (Socci, et al. 1998). The idea of a funnel is that the gradient of the free energy surface guides the search through combinatorial space by making changes favorable if they move toward a free energy minimum. Analogously, the search for suitable combinations of amino acid-to-codon assignments and genes may be resolved by a fitness landscape that takes the form of a funnel that guides the search through combinatorial space towards combinations that support a robust translation system.

Such combinatorial problems abound almost anywhere one looks; they are especially common to biology. They can be described as "Levinthal-like" paradoxes because, they share the formulation articulated by Levinthal for protein folding (Levinthal 1968; Dill and Chan 1997). Solutions to that paradox take the general form of a free energy landscape or "funnel" (Socci, et al. 1998). The idea of a funnel is that the gradient of the free energy surface guides the search through combinatorial space by making changes favorable if they move toward a free energy minimum. An analogous fitness landscape likely guided the search for suitable combinations of amino acid-to-codon matching and genes.

What might have shaped such a funnel in the fitness landscape on which genetic coding emerged? A force is a change in an energy field that changes the direction or velocity of motion. If we view evolution as motion on a landscape then selective constraints imposed on that motion can be seen as biological "forces". One selective constraint is that a translated gene sequence can affect chemistry best if and only if it always folds into about the same 3D structure. Another is that a sequence will fold if and only if its amino acids occur in an appropriate order.

These two constraints are shown by the pairs of horizontal red arrows in Fig. 1B. Each acts as a logical gate, reducing the passage either of peptides that do not fold, or of sequences that do not obey the code. The two logical gates can exert an extraordinary, iterative force if the consequent of the first serves as the antecedent of the second and *vice versa*. We argued earlier (Carter and Wills 2018b; Wills and Carter 2018) that such coupling makes the launch of coding from protein much more probable than from ribozymal synthetases.

Searching at the same time for the coding table and the gene sequences required to implement it also helps solve the Levinthal paradox by shaping the landscape of evolution. Computer modeling of autocatalytic sets shows that side reactions known as "parasites" are the chief threat to survival (Takeuchi, et al. 2017). Reciprocally coupled gating eliminates parasites in both directions.

When two logic gates are coupled tip-to-tail with the consequent of one serving as the antecedent for the other, they make a "strange loop" (Hofstadter 1979, 2007). Tip-to-tail coupling creates a self-referential cycle. Self-reference establishes a threshold beyond which consistent systems become incomplete. Incompleteness, in turn, implies a reservoir of possibilities. The strange loop thus creates a separate force. In chemistry the spatial gradient of a species' chemical potential induces a change in its equilibrium distribution. That gradient of novelty drives discovery in ways analogous to the role of chemical potential.

Searching at the same time for code words (the coding table) and code keys (the AARS gene sequences) adds a cooperative element to the search that enhances its efficiency (Carter and Wills 2018b; Wills and Carter 2018). At the same time, the implicit incompleteness also ensures the emergence of novelty. Novelty, in turn, is the grist on which selection and evolution work.

Conclusions

How do the new results in this paper help answer the questions posed in our opening paragraph? We crafted them with some care, because the authors differ on how to view the question of Nature's agency. When it comes to evolution, agency offers a helpful shorthand to keep sentences from being too verbose. Too much agency can be conflated with creationism. In the central chapter of his book *How Life Works: A User's Guide to the new Biology,* Philip Ball outlines why this question is so vexing (Ball 2023). Here, in a similar vein, we try to emphasize that the question is also intrinsic to our subject. The birth of genetic coding was indeed a remarkable part of biology's transcendence of chemistry. As such, we must recognize how close to optimal the genetic system really is.

Nor can we ignore what appears to be a sense of purpose. Lineages of ever better enzymes built, expanded and refined a table of symbols to capture the physical chemistry of amino acids. As the table grew, so too did the capacity to enhance the catalytic activity and specificity of the genes whose translated products enforced that growing table. Some property of the first genes allowed that to happen.

A growing set of new pilot studies help tie the early evolution of AARS to the emergence of genetic coding. *E. coli* can generate *in vivo* a nested set of active excerpts similar to the protozymes and urzymes we had previously designed as models for ancestral AARS. A mid-sized "Goldilocks" variant made *in vivo* and a short new Class II urzyme suggest how to design bidirectional urzyme genes. The inversion symmetry of base pairing in sense/antisense ancestral AARS genes projects into the proteome. That leads in turn to contrasting amino acid and RNA substrate binding modes. Codon middle-base pairing can help coordinate the building of Class I and II AARS family trees. Acylation of specific subsets of a minihelix library by Class I and II AARS urzymes shows RNA substrate specificity for the first time. Finally, new phylogenetics routines solve four problems that had blocked rooting AARS trees in reduced coding alphabets. Those algorithms will enhance our experimental work. Together, they form a viable platform to study how Nature likely built the earliest genetic coding tables and began to enforce them with quite simple AARS urzyme•minihelix cognate pairs.

Methods and Materials

Zymography

To visualize amino acid activation in a native polyacrylamide gel chromogenically, Zymography was done using a 1.5 mm thick native gel of 8% resolving and 5% stacking which devoid of SDS. Protein samples were prepared without adding SDS and β-ME to maintain the native conformation of protein. Prepared protein (50 µg) was then loaded into two separate wells of native gel. Electrophoresis was done with 40 mA steady current at 4°C. After electrophoresis for 30 minutes the gel was transferred from the glass plates for staining. Then the gel was shaken twice with double distilled water for 5 min. After washing, reaction buffer—50 mM HEPES pH 7.5, 100 mM amino acid (here glycine), 20 mM $MgCl_2$, 50 mM KCl, pyrophosphatase solution [NEB] (0.1 Unit/ml) and polyethylene glycol (PEG-8000; Sigma-Aldrich, Cat. No. 25322-68-3) to a final concentration of 5% (w/v)— was poured over the gel. The gel setup was shaken for 45 minutes at 4°C. This step ensures complete soaking of substrate mixture into the gel; the low temperature reduces inactivation of the enzyme during perfusion.

After the perfusion most of the solution was decanted, leaving minimal solution in a static condition at 37°C. Synthetase activity was activated by adding 5mM ATP solution dropwise onto the gel surface to cover the whole gel. The gel was incubated for 30 minutes. After decanting the reaction mix from the gel box, the staining solution (0.05% Malachite green in 0.1 N HCl and 5 % hexa-ammonium heptamolybdate tetrahydrate solution in 4 N HCl) was added directly onto the gel box. Staining was promoted by shaking for 2 minutes on a gyro-shaker. The staining solution was made as described in (Onodera, et al. 2021). The gradual development of green bands (620 nm) around GlyCA protein present in the gel signified the phosphomolybdate-malachite green complex formation and thus the *in-situ* activity of amino acid activation by GlyCA. The gel was photographed using Gel Doc™ XR+ from BIO RAD imaging machine.

Isolation of recombinant deletions from a plasmid containing a LeuRS double mutant.

We observed recombinant deletions by sequencing plasmids from ~60 colonies of *E. coli* transformed with a double mutant of WT *P. horikoshii* LeuRS. Histidine and lysine residues of both HVGH and KMSKS catalytic signatures were mutated to alanine for an earlier project (Tang, et al. 2023).

Identification of in vivo recombinant deletions. The plasmid vector was pET11a (Novagen, Sacramento, CA, United States). DNA oligos were ordered from IDT (Integrated DNA Technologies, Coralville, IA, United States). Phusion™ Plus PCR Master Mix (Cat# F631S) was purchased from Thermo Fisher Scientific (Waltham, MA, United States). *E coli* competent cells were partially from Agilent (XL10-Gold ultracompetent cells, Cat# 200315, Santa Clara, CA, United States), and partially using DH5a, home-made, prepared according to the method described by Sharma in 2017. Restriction enzyme DpnI (Cat# 500402, United States) was purchased from Agilent. Purifications and handling of DNA fragment and plasmid used the QIAquick Gel Extraction Kit (Cat# 28704, Qiagen, Hilden, Germany) and QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden, Germany) according to instruction, unless stated otherwise.

Method for preparation of double mutant plasmid. DNA oligo sequences were designed according to the manual of QuikChange II-E Site-Directed Mutagenesis Kit (Santa Clara, CA, United States) as published online at https://www.agilent.com/cs/library/usermanuals/public/200555.pdf. In line with the oligo DNA design, and for creating the double mutant plasmid DNA, we have implemented a megaprimer PCR method described by Picard (Picard, et al. 1994).

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

E. coli. transformation and plasmid sequencing. E coli transformation was proceeded according to conventional procedure (Fritsch, et al. 1982) and spread onto LB (Lysogeny broth) agar plates with antibiotics (Ampicillin 50 mg/mL, or Carbenicillin 100 mg/mL). After incubation overnight at 37 ºC, individual colonies on the LB agar plate were readied for miniprep by following the standard protocol described by QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden, Germany)

Plasmid DNA samples were sequenced following standard sequencing sample preparation and submission described by Eton Bioscience (San Diego, CA) online:

https://www.genewiz.com/en/Public/Resources/Sample-Submission-Guidelines/Sanger-Sequencing-Sample-Submission-Guidelines/Sample-Preparation#sanger-sequence with minor modifications. Basically, each sequencing sample was in 15 mL, containing 100-200 ng plasmid DNA with 5-10 picomole sequencing primer, i.e. seqn1_AVGA or AMSAS_seqn2rv as detailed sequences in Table 1 above.

Method for plasmid sequencing and sequencing analysis. Sequence analysis was performed using conventional methods as follows. First, all the DNA sequences were translated into 6 reading frames in amino acid sequence using Expasy translate online at https://web.expasy.org/translate/. For screening the tentative urzyme candidates, each amino acid sequence from 6 reading frames from a single sequenced plasmid DNA sample were aligned with the sequence of urzyme template using conventional sequencing alignment software such as MAFFT online at https://www.ebi.ac.uk/jdispatcher/msa/mafft?stype=protein (Katoh, et al. 2019) or MUSCLE online at https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein (Edgar 2004). Novel plasmids were designated urzyme, urzyme-like and goldilocks, according to their sequence content and were collected and subjected for further downstream identification at the enzymology level after protein biochemistry procedures.

Aminoacylation of minihelix combinatorial libraries

We constructed a minihelix library with sequences 35 nucleotides long derived from the acceptor-stem and TΨC stem loop of $tRNA^{Leu}$. We varied only bases 1-3, 70-73 and these were given all ~16.4 K possible

674 combinations. The acceptor stem of tRNALeu is very GC rich and so will bias recognition by different
675 synthetases.

676 Minihelix library constructs were generated (Integrated DNA Technologies) with seven fully randomized
677 positions in the acceptor stem including the first three base pairs and the Discriminator base (Schimmel, et
678 al. 1993). Minihelix RNA was heated to 90°C for 2 minutes and cooled to 22°C for 10 minutes. Refolded
679 minihelix RNA was resuspended in acylation buffer: 50 mM HEPES, pH 7.5, 20 mM MgCl2, 20 mM KCl,
680 5 mM DTT, 20 mM ATP and 20mM amino acids. Aminoacylations were performed by mixing the minihelix
681 RNA with a specified Urzyme (LeuAC (30uM) and GlyCA (15uM), for 30 and 60 min, respectively).
682 Unreacted amino acids were removed by running each sample through two size-exclusion columns (Zeba,
683 7K, ThermoFisher). NHS-biotin (ThermoFisher) was added in excess on ice to attach biotin to any RNA
684 with a primary amine as previously described (Pütz, et al. 1997; Chumachenko, et al. 2009). Unreacted
685 NHS-biotin was removed by running each sample through two size-exclusion columns. Biotinylated RNAs
686 were captured using streptavidin magnetic beads in 0.1M sodium acetate (pH 5.2) to protect the substrate
687 from degradation. The inactive RNAs were removed by three washes of the beads with sodium acetate.
688 Biotinylated RNAs were released from the beads by breaking the RNA-amino acid bond through incubation
689 in 50mM HEPES pH 7.5 for 30 min at 37°C.

690 After the partitioning steps, short RNA oligonucleotides were ligated to the ends of the active RNAs to
691 facilitate reverse transcription and PCR needed for sequencing using the TruSeq small RNA library prep
692 kit (Illumina). We adapted the ligation steps to 24h, 15°C in the presence of 27% PEG 8000 (Promega) to
693 minimize ligation sequence bias, as previously described (Song, et al. 2014). Duplicate selections and the
694 starting library were multiplexed and sequenced on iSEQ (Illumina), yielding about 1 million sequences
695 per population. The constant regions of the sequences were trimmed off prior to analysis using our python
696 script (Popović, et al. 2015). For each population the filtered reads were counted and compared between
697 populations using FASTAptamer toolkit (Alam, et al. 2015).

698 **Identification of Class I, II bidirectional urzyme genes.**

699 The architecture of the Goldilocks LeuAC urzyme (Fig. 5D) provided the template for constructing and
700 testing possible antiparallel alignments. Pymol (Pymol 2010) representations of candidate urzymes
701 provided the platform for comparing alternative alignments between Class I and II urzymes in which we
702 deleted the second crossover connection. After transferring the sequences saved from pymol as .fasta files
703 to Excel and accessing the codon middle bases, we manually counted the number of middle base pairs in a
704 small number of relative antisense alignments. Having established a secure hypothesis from this preliminary
705 study, we assembled three alignments from aaRS.online (Douglas, Cui, et al. 2024). The amino acid
706 sequences were aligned for both candidate Classes to the secondary structures as well as the codon middle
707 bases, both of which can be represented by a single letter. We used that platform to assemble multiple
708 sequence alignments, which then were analyzed by an automated procedure to yield a matrix of middle-
709 base pairing frequencies.

710 **Data availability**

711 All data underlying this article will be shared on reasonable request to the corresponding author.

712 **Funding**

713 This work was supported by the Alfred P. Sloan Foundation Matter-to-Life Program (Grant G-2021-
714 16944); the National Institutes of Health (Grant R35GM131923 (B.K.) and NSF fellowship DGE-2040435
715 (H.D.). H.D. acknowledges support by a Pre-doctoral Fellowship from the American Foundation for
716 Pharmaceutical Education. This work utilized the resources of the UNC Longleaf high-performance
717 computing cluster.

718 **Design of bidirectional genes with Class I and II AARS urzymes on opposite strands.**

719 Initial working models for bidirectional gene construction was performed with Pymol on grounds suggested
720 by the Leucyl Goldilocks urzyme. The corresponding Class II AARS urzyme was previously thought to be

continuous between Motif 1 and Motif 2. The discovery that the second crossover connection of the Rossmann fold is not needed for catalytic activity required some accommodations in the Class II complement because of the overlap between it and the continuous sequence linking Motifs 1 and 2. These accommodations were identified by anchoring the two Class I and II signatures. The Class II N-terminus was then determined by the length of sequence included following the KMSKS sequence in the Class I partner.

Structures of the TrpRS (Class I) and HisRS (Class II) urzymes were obtained, and any extant strand breaks were repaired using Rosetta loop modeling (Mandell et al., 2009, https://doi.org/10.1038/nmeth0809-551). To design bidirectional genes, we modified sequence design model ProteinMPNN (Dauparas et al., 2022, https://doi.org/10.1126/science.add2187) to enforce bidirectional genetic coding constraints. To preserve the consensus Class I and II sequence signatures, we fixed various active site residues (Class I: HIGH motif, S5, Q8, P9, A21, H42, M51, D54, K81, K84; Class II: VTDV motif, R86, E87) and designed all remaining positions. We generated 10,000 sequences using the default sampling temperature of 0.1. Code for our modified version of ProteinMPNN capable of designing bidirectional genes can be found at https://github.com/Kuhlman-Lab/proteinmpnn.

References

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature https://doi.org/10.1038/s41586-024-07487-w

Alam KK, Chang JL, Burke DH. 2015. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. Molecular Therapy—Nucleic Acids 4:e230.

Ball P. 2023. How Life Works: A User's Guide to the New Biology: University of Chicago.

Berg P, Ofengand EJ. 1958. An Enzymatic Mechanism for Linking Amino Acids to RNA. Proc Nat Acad Sci USA 44:78-85.

Bokma F. 2002. Detection of punctuated equilibrium from molecular phylogenies. .Journal of Evolutionary Biology 15:1048–1056.

Bridgham JT, Ortlund EA, Thornton JW. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461:515-519.

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

Carter CW, Jr , Wills PR. 2018a. Hierarchical groove discrimination by Class I and II aminoacyl-tRNA synthetases reveals a palimpsest of the operational RNA code in the tRNA acceptor-stem bases. Nucleic Acids Research 46:9667–9683.

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, Wills PR. 2018b. Interdependence, Reflexivity, Fidelity, and Impedance Matching, and the Evolution of Genetic Coding. Molecular Biology and Evolution 35:269-286.

Carter CW, Jr. 2016. An Alternative to the RNA World. Natural History 125:28-33.

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

763 Carter CW, Jr. 2022. How did the proteome emerge from pre-biotic chemistry? In: Fiore M, editor. Pre-
764 Biotic Chemistry and Life's Origin. London, UK: The Royal Society of Chemistry. p. 317-346.

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

767 Carter CW, Jr. 2008. Whence the Genetic Code?: Thawing the 'Frozen Accident'. Heredity 100:339-340.

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

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

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

777 Carter CW, Jr., Wills PR. 2021a. Reciprocally-coupled Gating: Strange Loops in Bioenergetics, Genetics,
778 and Catalysis. Biomolecules 11:265.

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

781 Cestari I, Stuart K. 2013. A spectrophotometric assay for quantitative measurement of aminoacyl-tRNA
782 synthetase activity. J Biomol Screen. 18:490–497.

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

786 Chumachenko NV, Novikov Y, Yarus M. 2009. Rapid and simple ribozymic aminoacylation using three
787 conserved nucleotides. . Journal of the American Chemical Society 131:5257-5263.

788 Crick FHC. 1970. Central Dogma of Molecular Biology. Nature 227:561-563.

789 Crick FHC. 1968. The Origin of the Genetic Code. Journal of Molecular Biology 38:367-379.

790 Crick FHC, Brenner, S., Klug, A., and Pieczenik, G. 1976. A Speculation on the Origin of Protein Synthesis.
791 Origins of Life 7:389-397.

792 Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A Sequence Logo Generator. Genome
793 Research 14:1188-1190.

794 Cusack S. 1994. Evolutionary Implications. Nature Structural and Molecular Biology 1:760.

795 Delarue M. 2007. An asymmetric underlying rule in the assignment of codons: Possible clue to a quick
796 early evolution of the genetic code via successive binary choices. RNA 13:1-9.

797 Delarue M, Moras D. 1992. Aminoacyl-tRNA synthetases: Partition into two classes. In: Eckstein F, Lilley
798 DMJ, editors. Nucleic Acids and Molecular Biology. Berlin, Heidelberg: Springer-Verlag. p. 203-224.

799 Dill K, Chan HS. 1997. From Levinthal to pathways to funnels. Nature Structural Biology 4:10-19.

800 Dill KA, MacCallum JL. 2012. The Protein-Folding Problem, 50 Years On. Science 338:1042-1046.

801 Doublié S, Gilmore CJ, Bricogne G, Carter CW, Jr. 1995. Tryptophanyl-tRNA synthetase crystal structure
802 reveals an unexpected homology to tyrosyl-tRNA synthetase. STRUCTURE 3:17-31.

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

Douglas J, Bouckaert R, Harris S, Carter CW, Jr, Wills PR. 2024. Evolution is coupled with branching across many granularities of life. Proceedings of the Royal Society B In Review.

Douglas J, Carter CW, Jr, Wills PR. 2024. HetMM: A Michaelis-Menten model for non-homogeneous enzyme mixtures. iScience 27:108977.

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 In Press.

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

Eldredge N GSJ. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf TJM, editor. Models in paleobiology. San Francisco: Freeman. p. 82-115.

Eriani G, Delarue M, Poch O, Gangloff J, Moras D. 1990. Partition of tRNA synthetases into two classed based on mutually exclusive sets of sequence motifs. Nature 347:203-206.

Francklyn C, Schimmel P. 1989. Aminoacylation of RNA Minihelices with Alanine. Nature 337:478-481.

Fritsch EF, Sambrook J, Maniatis T. 1982. Molecular Cloning: A Laboratory Manual. New York.: Cold Spring Harbor Laboratory.

Frugier M, Florentz C, Giegé R. 1992. Anticodon-Independent Valylation of an RNA Minihelix. Proceedings of the National Academy of Sciences USA 89:3900-3904.

Giegé R. 1972. Recherches sur la spécificité de reconnaissance des acides ribonucléiques de transfert par les aminoacyl-tRNA synthétases [Study on the specificity of recognition of transfer ribonucleic acids by aminoacyl-tRNA synthetases]. [Thèse de Doctorat d'Etat]. [Strasbourg, France]: Université Louis Pasteur.

Guseva E, Zuckermann RN, Dill KA. 2017. Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. Proc. Nat. Acad. Sci. USA 114: E7460–E7468.

Han L, Luo Z, Ju Y, Chen B, Taotao Z, Wang J, Xu J, Gu Q, Yang X-L, Schimmel P, Zhou H. 2023. The binding mode of orphan glycyl-tRNA synthetase with tRNA supports the synthetase classification and reveals large domain movements. Science Advances 9:eadf1027.

Hoagland MB, E. B. Keller, Zamecnik. PC. 1956. Enzymatic Carboxyl Activation of Amino Acids. . J. Biol. Chem. 21:345-358.

Hofstadter DR. 1979. Gödel, Escher, Bach: an eternal golden braid. New York: Basic Books, Inc.

Hofstadter DR. 2007. I Am A Strange Loop. Philadelphia, PA: Basic Books.

Hordijk W. 2019. A History of Autocatalytic Sets: A Tribute to Stuart Kauffman. Biological Theory 14:224–246.

Jones OW, Jr. , Nirenberg MW. 1966a. Degeneracy in the amino acid code. Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis 119:400-406.

Jones OW, Nirenberg MW. 1966b. Degeneracy in the amino acid code", . Biochim. Biophys. Acta 2:400–406.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20:1160-1166.

Katsnelson MI, Wolf YI, Koonin EV. 2019. On the feasibility of saltational evolution. Proc. Nat. Acad. Sci. USA 116 21068–21075.

Kauffman SA. 1986. Autocatalytic Sets of Proteins. J. Theor. Bio., 119:1–24.

Klipcan L, Safro M. 2004. Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. Journal of Theoretical Biology 228:389–396.

Koonin EV, Novozhilov AS. 2017. Origin and Evolution of the Universal Genetic Code. Annu. Rev. Genet. 51:45–62.

Lanier KA, Petrov AS, Williams LD. 2017. The Central Symbiosis of Molecular Biology: Molecules in Mutualism. J. Mol. Evol. 85.

Le SQ, Gascuel O. 2008. An Improved, General Amino-Acid Replacement Matrix. Molecular Biology and Evolution. 2008 25:1307-1320.

Levinthal C. 1968. ARE THERE PATHWAYS FOR PROTEIN FOLDING ? Journal de Chimie Physique 65:44.

Li L, Weinreb V, Francklyn C, Carter CW, Jr. 2011. Histidyl-tRNA Synthetase Urzymes: Class I and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for Cognate Amino Acid Activation. J. Biol. Chem. 286:10387-10395.

Liu Z, Wu L-F, Xu J, Bonfio C, Russell DA, Sutherland JD. 2020. Harnessing chemical energy for the activation and joining of prebiotic building blocks. Nature Chemistry 12:1023–1028.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Manceau M, Marin J, Morlon H, Lambert A. 2020. Model-Based Inference of Punctuated Molecular Evolution. Molecular Biology and Evolution 37:3308–3323.

Martinez-Rodriguez L, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, Weinreb V, Chandrasekaran SN, Collier M, Ambroggio X, Kuhlman B, et al. 2015. Functional Class I and II Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. J. Biol. Chem. 290:19710–19725.

Muñoz V, Serrano L. 1994. Intrinsic Secondary Structure Propensities of the Amino Acids, Using Statistical f-y matrices: Comparison with Experimental Scales. PROTEINS: Structure, Function, and Genetics 20:301-311.

Newman DV. 1996. EMERGENCE AND STRANGE ATTRACTORS. Philosophy of Science 63:245-261.

Nirenberg MW, Mattaei JH. 1961. The Dependence of Cell-Free Protein Synthesis in E. coli Upon Naturally Occurring or Synthetic Polyribonucleotides. Proceedings of the National Academy of Sciences USA 47:1588-1602.

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

Ostrem DL, Berg P. 1974. Glycyl Transfer Ribonucleic Acid Synthetase from Escherichia coli; Purification, Properties, and Substrate Binding! Biochemistry 13:1338

Pagel M, Venditti C, Meade A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 5796:119–121.

Patra SK, Betts L, Tang GQ, Douglas J, Wills PR, Bouckeart R, Carter CW, Jr. . 2024. A genomic database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class II urzymes. Nucleic Acids Research In Press.

Pham Y, Kuhlman B, Butterfoss GL, Hu H, Weinreb V, Carter CW, Jr. 2010. Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular complementation. J. Biol. Chem. 285:38590-38601.

Pham Y, Li L, Kim A, Erdogan O, Weinreb V, Butterfoss G, Kuhlman B, Carter CW, Jr. 2007. A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA Synthetases. Mol. Cell 25:851-862.

Picard V, Ersdal-Badju E, Lu A, Bock SC. 1994. A rapid and efficient one-tube PCR-based mutagenesis technique using Pfu DNA polymerase. Nucleic Acids Research 22:2587-2591.

Popović M, Fliss PS, Ditzler MA. 2015. In vitro evolution of distinct self-cleaving ribozymes in diverse environments. . Nucleic Acids Res. 43:7070-7082.

Pütz J, Wientges J, Schwienhorst A, Sissler M, Giegé R, Florentz C. 1997. Rapid selection of aminoacyl-tRNAs based on biotinylation of α-NH2 group of charged amino acids. Nucleic Acids Research 25:1862-1863.

Pymol. 2010. The PyMOL Molecular Graphics System. New York, NY: Schrödinger, LLC.

Ribas de Pouplana L, Frugier M, Quinn C, Schimmel P. 1996. Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. Proceedings of the National Academy of Sciences, USA 93:166-170.

Ribas de Pouplana L, Schimmel P. 2001a. Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends in Biochemical Sciences 26:591-596.

Ribas de Pouplana L, Schimmel P. 2001b. Two Classes of tRNA Synthetases Suggested by Sterically Compatible Dockings on tRNA Acceptor Stem. Cell 104:191-193.

Ribas L, Schimmel P. 2001. Two Classes of tRNA Synthetases Suggested by Sterically Compatible Dockings on tRNA Acceptor Stem. Cell 104:191-193.

Rodin SN, Ohno S. 1995. Two Types of Aminoacyl-tRNA Synthetases Could be Originally Encoded by Complementary Strands of the Same Nucleic Acid. Origins of Life and Evolution of the Biosphere 25:565-589.

Rodin SN, Rodin A. 2006. Partitioning of Aminoacyl-tRNA Synthetases in Two Classes Could Have Been Encoded in a Strand-Symmetric RNA World. DNA and Cell Biology 25:617-626.

S. Rastogi, D. A. Liberles. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology 5: 28.

Safro M, Klipcan L. 2013. The mechanistic and evolutionary aspects of the 2′- and 3′-OH paradigm in biosynthetic machinery. Biology Direct 8:17.

Schimmel P. 1991. RNA minihelices and the decoding of genetic information. The FASEB Journal 5:2180-2187.

Schimmel P, Giegé R, Moras D, Yokoyama S. 1993. An operational RNA code for amino acids and possible relationship to genetic code. Proceedings of the National Academy of Sciences, USA 90:8763-8768.

Serrano L, Sancho J, Hirshberg M, Fersht AR. 1992. Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol 227:544-559.

Shore J, Holland BR, Sumner JG, Nieselt K, Wills PR. 2019. The Ancient Operational Code is Embedded in the Amino Acid Substitution Matrix and aaRS Phylogenies. Journal of Molecular Evolution 88:136–150.

Sobotta J, Geisberger T, Moosmann C, Scheidler CM, Eisenreich W, Wächtershäuser Gn, Huber C. 2020. A Possible Primordial Acetyleno/Carboxydotrophic Core Metabolism. Life 10:35.

Socci ND, Onuchic JN, Wolynes PG. 1998. Protein Folding Mechanisms and the Multidimensional Folding Funnel. PROTEINS: Structure, Function, and Genetics 32:136–158.

Song Y, Liu KJ, Wang T-H. 2014. Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. PLoS ONE 9:e94619.

Stubbs RT, Yadav M, Krishnamurthy R, Springsteen GR. 2020. A plausible metal-free ancestral analogue of the Krebs cycle composed entirely of α-ketoacids. Nature Chemistry 12:1016–1022.

Takénaka A, Moras D. 2020. Correlation between equi-partition of aminoacyl-tRNA synthetases and amino-acid biosynthesis pathways. Nucleic Acids Research, 48:3277–3285.

Takeuchi N, Hogeweg P, Kaneko K. 2017. Conceptualizing the origin of life in terms of evolution. Phil. Trans. R. Soc. A 375:20160346.

Tang GQ, Elder JJH, Douglas J, Carter CW, Jr, . 2023. Domain Acquisition by Class I Aminoacyl-tRNA Synthetase Urzymes Coordinated the Catalytic Functions of HVGH and KMSKS Motifs. Nucleic Acids Research 51:8070–8084.

Tang GQ, Hu H, Douglas J, Carter CW, Jr. 2024. Primordial aminoacyl-tRNA synthetases preferred tRNA minihelix substrates over full-length tRNA. Nucleic Acids Research 52:1-24.

Trupin JS, Rottman FM, Brimacome R, Leder P, Bernfield MR, Nirenberg M. 1965. RNA Codewords and Protein Synthesis, VI. On the Nucleotide Sequences of Degenerate Codeword Sets for Isoleucine, Tyrosine, Asparagine, and Lysine. Proc. Natl. Acad. Sci. U.S.A. 53 807–811.

Watson JD, Crick FHC. 1953. A Structure for Deoxyribose Nucleic Acid. Nature 171:737-738.

Whelan S, Goldman N. 2001a. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18:691-699.

Whelan S, Goldman N. 2001b. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18:691–699.

Wills PR. 2001. Autocatalysis, information, and coding. Biosystems 50:49-57.

Wills PR, Carter CW, Jr. 2018. Insuperable problems of an initial genetic code emerging from an RNA World. Biosystems 164:155-166.

Wills PR, Carter CW, Jr. 2020. Impedance matching and the choice between alternative pathways for the origin of genetic coding. International Journal of Molecular Sciences 21:7392.

Woese CR, Olsen GJ, Ibba M, Soll D. 2000. Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process. Microbiology and Molecular Biology Reviews 64:202–236.

Wong JT-F, Ng S-K, Mat W-K, Hu T, Xue H. 2016. Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life. Life 6:12.

Yarus M, Widmann J, Knight R. 2009. RNA-amino acid binding: A stereochemical era for the genetic code. Journal of Molecular Evolution 69:406-429.

Zhu TF, Budin I, Szostak JW. 2013. Preparation of Fatty Acid or Phospholipid Vesicles by Thin-film Rehydration. Methods in Enzymology 533:267-274.

965 Zull JE, Smith SK. 1990. Is genetic code redundancy related to retention of structural information in both
966 DNA strands? Trends in Biochemical Sciences 15:257-261.

967

968