The series of genetic codes that preceded our own

Joanna Masel[1], Jordan Douglas[2], Sawsan Wehbi[3], H. Solomon McShea[4]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA

[2]Research School of Biology, Australian National University, Canberra, ACT 2600, Australia

[3]Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona, 85721, USA

[4]Department of Microbiology & Immunology, University of California, San Francisco

Corresponding author: Joanna Masel masel@arizona.edu

Abstract

We produce a working hypothesis for the steps by which the canonical genetic code might have been constructed. We assume amino acids were added in the order inferred from their enrichment or depletion in ancestrally reconstructed sequences dating back to the Last Universal Common Ancestor (LUCA) relative to that in ancient but post-LUCA controls. To convert this ordering into a time series of 64-codon genetic codes, we followed the "2-1-3 rule" that information appeared first at the middle codon position, then mostly at the 1st before the 3rd, with purine vs pyrimidine distinctions tending to precede distinctions between purines or between pyrimidines. Two notable deviations from the 2-1-3 rule suggest an ancestral promiscuity between valine/isoleucine/ methionine, and between alanine/threonine, to produce "statistical proteins". The same promiscuities enable cladograms of Class I and Class II aminoacyl-tRNA synthetase catalytic domains to align perfectly with the amino acid order of appearance. Trifonov's previous amino acid ordering did not produce this consilience, with both a worse fit to the 2-1-3 rule, and discordance of synthetase trees with the order. More support for late resolution of hydrophobic promiscuity comes from post-LUCA differentiation between the synthetases of valine and isoleucine, in a manner dependent on two conserved tryptophans. The observed promiscuous activities of synthetases, and deep mutational scanning data on which amino acid substitutions are best tolerated, also align well with our reconstructed time series of genetic codes. As a working hypothesis, this series can guide research on random peptides, urzymes, and substitution models.

Keywords: early life; astrobiology; enzyme promiscuity; bacteria; archaea

Introduction

The origin of the genetic code, which specifies which of the 64 RNA triplets are translated into which of the 20+ amino acids, has long captivated biologists. By some metrics, the genetic code seems, if not optimal, then at least markedly superior to alternative arrangements (Haig and Hurst 1991; Freeland and Hurst 1998; Freeland *et al.* 2003; Itzkovitz and Alon 2007). In other ways, it seems arbitrary, perhaps a "frozen accident" of contingent evolutionary history (Crick 1968). A number of theories have been put forth regarding non-accidental principles of its construction (see Koonin and Novozhilov (2009) and Kondratyeva et al. (2022) for review).

The most compelling theories begin by presuming that the evolution of the genetic code, like the evolution of anything else, did not involve competition among unrelated random assemblies, but instead proceeded through a series of steps, beginning with fewer amino acids and progressively adding more (Ardell 1998; Cavalcanti *et al.* 2004; Sella and Ardell 2006; Stoltzfus and Yampolsky 2007; Massey 2008; Higgs 2009; Massey 2016). Patterns in these steps would create patterns in the final genetic code. E.g., when new amino acids tend to take codons that were previously assigned to amino acids with similar properties, this creates a code that is robust to error. Reconstructing the steps via which amino acids were added, and identifying the principles guiding them, is key to further progress in understanding the origin of the genetic code.

Reconstructions of such ancient events inevitably entail speculation. We find it useful to orient arguments around empirical "facts" (meaning statements that are undisputed) that seem informative enough to constrain speculation. We are explicit about what we consider to be a fact, versus a proposition supported more indirectly by facts. By combining two such empirically derived propositions, we derive a specific hypothesis regarding the sequence of intermediate codes. We then look for compatibility between this hypothesis and data not included in the original set of facts.

One set of initial facts comes from the current code's layout. Patterns in which codons have more chemically similar amino acids support the proposition that there might once have been a singlet code where only the middle position of each nucleotide triplet bore information, followed later by a doublet code of the first two positions, with information in the third position appearing last (Jukes 1974; Patel 2005; Wu *et al.* 2005; Massey 2006; Travers 2006). Here we accept this "2-1-3" hypothesis as well enough supported to be true as a general tendency, albeit allowing for exceptions whereby some information in the third position might precede full use of information in the first position.

A second set of diverse facts has been used to inform an educated and influential working model as to which amino acids came late vs. early (Trifonov 2000). However, many of the metrics used to make this model pertain to the prebiotic transition rather than to the presumably later origin of the code within a lifeform complex enough to produce a ribosome. Such a lifeform must have used RNA in sophisticated ways, and might have also made considerable use of diverse peptides (Di Giulio 1997; Van der Gulik and Speijer 2015; Fried *et al.* 2022; Müller *et al.* 2022). The first use of biopeptides and the origin of the genetic code are distinct questions, with distinct pertinent facts.

Recently, Wehbi et al. (2024) instead inferred a proposition about the order of recruitment from facts about amino acid usage in sequences that date to the Last Universal Common Ancestor relative to slightly less old controls. This approach avoids confounding with metrics more pertinent to the prebiotic transition, and is agnostic with respect to the chemical properties of the amino acids. Here we first make minor improvements to this inference. We then strictly follow this proposition regarding the order in which the amino acids were added, while also being guided by the 2-1-3 proposition, to articulate an explicit hypothesis for the nature of intermediate codes.

We then turn to aminoacyl-tRNA synthetases (aaRSs, abbreviated throughout as e.g. AlaRS for alanyl-tRNA synthetase; see Rubio Gomez & Ibba (2020) for review). The catalytic domains of aaRSs recognize and activate amino acids in the presence of ATP, and attach them to their cognate tRNA. Sometimes an additional active site, or entire editing domain, rejects a promiscuously activated amino acid (e.g. in LeuRS and ProRS), and in some cases biosynthesis is completed only after a tRNA is charged with a precursor amino acid (e.g. of the amino acids glutamine or arginine). We review facts about which aaRSs have latent promiscuity for which alternative amino acids. Patterns of promiscuity give clues about which aaRS evolved from which. There are two trees for the aaRS catalytic domain, corresponding to unrelated structures (Eriani *et al.* 1990): one for Class I aaRSs, the other for Class II. We combine information about aaRSs, plus our order of recruitment, with previously inferred posterior probability distributions of aaRS catalytic domain phylogenies (Douglas *et al.* 2024), to arrive at a slightly revised Class I phylogeny.

We also refer to a meta-analysis of deep mutational scanning (DMS) experiments (Sun *et al.* 2024) summarizing protein tolerance for different amino acid substitutions. Sun *et al.* (2024) fit a single parameter to describe the distribution of fitness effects for each mutation type. We have organized these in a spreadsheet in Supplementary File 1, where e.g. the alanine tab shows this parameter for all mutation types away from alanine, all mutation types to it, and sorted according to the average of the two, from most permissive to most restrictive.

We look for consilience between our order of recruitment, the table of intermediate codes it suggests under the 2-1-3 rule, aaRS promiscuities and trees, and DMS data. As a control, we repeat this search for consilience using Trifonov's (2000) earlier hypothesized order of recruitment. Finally, from the resulting hypothesized order of events, we derive testable predictions.

Order of amino acid recruitment

Wehbi et al. (2024) compared the inferred ancestral usage of LUCA-era protein "clans" to the ancestral usage of protein clans that were presumed to be somewhat ancient but post-LUCA, and that thus served as methodological controls. A clan is a set of Pfam domains that are believed to be homologous but whose sequences are too divergent to be summarized by the same Hidden Markov Model. Wehbi et al. (2024) was careful in ensuring the antiquity of the LUCA-era clans, but did not take similar precautions to exclude the antiquity of the presumed post-LUCA clans that served as the control sequences. We therefore repeated their analysis with a more stringent denominator control to produce revised LUCA usages, and hence order of recruitment, in Table 1.

Amino Acid	LUCA usage	SE
G	1.1231	0.0303
V	1.1141	0.0257
T	1.1136	0.0340
M	1.0886	0.0413
I	1.0867	0.0337
H	1.0604	0.0445
A	1.0560	0.0335
E	1.0289	0.0377
P	1.0042	0.0298
K	0.9871	0.0416
S	0.9846	0.0282
D	0.9637	0.0287
N	0.9625	0.0407
C	0.9617	0.0732
L	0.9224	0.0263
R	0.9212	0.0287
F	0.8778	0.0329
Q	0.8623	0.0343
Y	0.8554	0.0358
W	0.6295	0.0480

Table 1. Revised order of recruitment of amino acids into the genetic code. Methods are the same as in Wehbi et al. (2024), except that we excluded from the post-LUCA denominator 1) clans that had already diversified into multiple Pfams by the time of the Last Archaeal Common Ancestor (LACA) or the Last Bacterial Common Ancestor (LBCA), and 2) clans with at least one Pfam, annotated as a LACA or LBCA candidate, that had duplicated prior to LACA or LBCA, respectively. Briefly, LUCA usage is equal to the reconstructed amino acid frequency at confidently inferred sites for clans inferred to be present in one copy in LUCA, divided by similarly reconstructed frequencies in clans inferred to be present in one copy in LACA or LBCA.

The relationship with amino acid size reported in Wehbi et al. (2024) becomes stronger following this revision (Figure 1A). The relationship grows even stronger again when considered separately for amino acids activated by Class I vs Class II

aaRSs, focusing only on amino acids directly activated by aaRSs rather than produced post-activation (Figure 1B).

The coevolution theory hypothesizes that amino acids were added to the genetic code as metabolic pathways expanded to synthesize them (Wong 1975). The coevolution theory makes the prediction that amino acid precursors in current metabolic networks were added to the code before the amino acids that are synthesized downstream of them. This prediction is not made by alternative theories of a pre-code nucleopeptide world, nor of dramatically different early metabolic pathways. Referring to the diagram of current metabolism given in Klipcan & Safro (2004), our ordering is compatible with the coevolution theory for 6 pairs: putting serine before cysteine and methionine, putting glutamate before glutamine, proline, and arginine, and putting threonine before isoleucine. However, it contradicts the coevolution theory for 4 pairs: putting aspartate after threonine, methionine, and lysine (with asparagine a tie in our data), and putting glycine after serine.

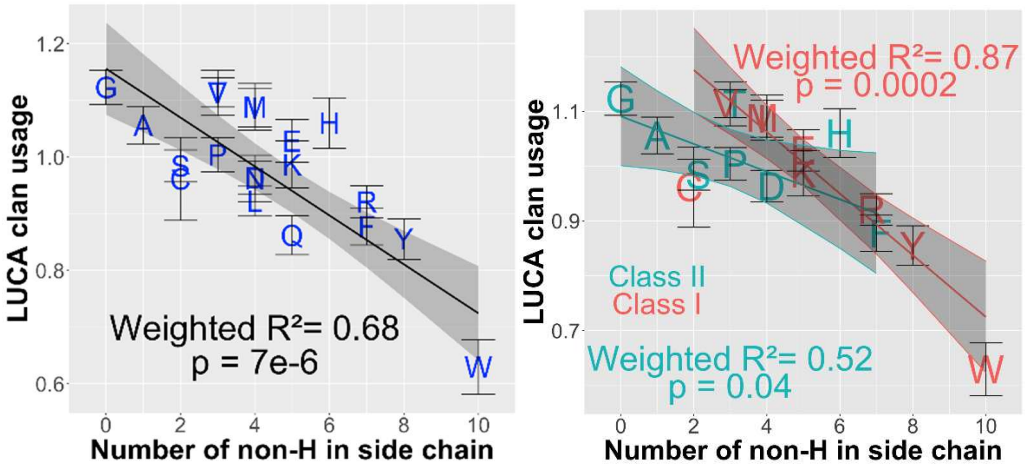


Figure 1. Smaller amino acids were added earlier to the genetic code. LUCA clan usage is from Table 1. In b), glutamine (Q) and asparagine (N) are omitted because the tRNA is charged with glutamate and aspartate, respectively, from which Q or N are produced while bound to tRNA. LysRS-I is presumed to precede LysRS-II, and CysRS to precede SepRS. The squared inverse of the LUCA usage SE values (shown as error bars from Table 1) were used for the 'weights' argument in the lm() function within the "stats" package in R, to produce the R^2 and p-values for weighted model 1 regression lines, with 95% CI shading.

Intermediate genetic codes

To construct a series of genetic codes, our first rule is to add amino acids strictly in the order shown in Table 1, albeit sometimes more than one at a time when LUCA usages from Table 1 are almost equal. Our second guideline is that when we place each amino acid in turn in its position in the modern genetic code, we expand the set of codons that encode it in a manner governed by the 2-1-3 rule. I.e., we assume that the genetic code tended to first use information from the middle nucleotide, then from the first, and only lastly from the third. This second consideration is a guideline not a rule, and we exercise discretion as to when to invoke promiscuous coding to avoid serious deviations from this guideline. Our second guideline also assumes that at all three positions, distinguishing purines from pyrimidines preceded distinguishing among purines or among pyrimidines. Indeed, in the modern genetic code, the third position continues to distinguish very little among purines or among pyrimidines.

In a "shift" rather than an addition, codons change identity among already-existing amino acids. Given limited data to inform shifts, we mostly neglect them in the interests of parsimony (although see arginine and methionine below).

We largely neglect the evolutionary history of stop codons. We assume that tRNAs and release factors compete to bind to every codon, with specificity likely increasing during the evolution of the genetic code. This means that all 64 codons were used at each stage, any of which, in the absence of sufficient binding of charged tRNA, could have functioned as a stop codon (Sonneborn 1965). Given uncertainties in the potentially rapid evolution of stop codon identity, we assign all 64 codons as sense in all but the current genetic code and its immediate predecessor. Indeed, some of today's stop codons can encode pyrrolysine (Polycarpo *et al.* 2004; Shalvarjian *et al.* 2025) or selenocysteine (Lee *et al.* 1989), in a relatively inefficient fashion in which appropriately charged tRNA competes with release factors; this may offer a glimpse into how stop codons operated prior to completion of the canonical genetic code.

The view of tRNAs competing to bind an anticodon shares an underlying thermodynamic principle with the view that when a newly added amino acid claims codons that previously coded for something else, it proceeds via a stage of ambiguous coding, during which the new amino acid achieves a smaller dissociation constant than the old. This process of codon reassignment differs from the theory of "codon capture" (Osawa and Jukes 1989; Kollmar and Mühlhausen 2017), where codons must first fall into disuse (presumably becoming stop codons) before being regained in a new form.

We next apply our rules and guidelines to produce the set of intermediate codes shown in Fig. 2. The first three amino acids in Table 1 – glycine, valine, and threonine – are in different columns of the contemporary genetic code; we expand them in Fig. 2.i to fill whole columns, representing information in the middle position. We further expand glycine to both columns with a purine in the middle position. Note that these three amino acids provide a highly complementary hydrophobic-polar-glycine (HPG) combination that has been explored by toy models of protein folding (Bhattacherjee and Wallin 2012), where polar regions appear on the surface and hydrophobic regions in buried positions, driven by the hydrophobic effect (Frank and Evans 1945; Dill 1990). Because glycine's side chain is so small, it makes the backbone of the polypeptide (which is polar (Honig and Cohen)) more available for hydrogen bonding; glycine backbones are the most common backbones found in active sites (Ribeiro *et al.* 2020). The backbone is notably important for ATP binding in Class I aaRSs (Kaiser *et al.* 2018). Hydrophobic-polar systems are believed to be sufficient for protein folding (Dill 1985; Kamtekar *et al.* 1993).

Hydrophobic isoleucine and methionine come next in Fig. 2.ii. If we strictly used parsimony to hypothesize that ancient codes distinguished between isoleucine and methionine in the same way that the contemporary code does, this would severely violate our second guideline, by using not just first position information, but even third position information. To avoid such a severe deviation from the 2-1-3 rule, we posit that the first column was non-specific, indeed perhaps accommodating any hydrophobic amino acid.

Other evidence also supports the hypothesis of a promiscuous ancestor (Fournier *et al.* 2011). Early life is thought to use relatively non-specific "statistical proteins" or "protein quasispecies" (Woese 1965; Fitch and Upper 1987; Wills *et al.* 2015) that incorporated whichever amino acids were available and recognized by an aaRS. An inferred ancestral aaRS sequence for ValRS, IleRS, LeuRS, and MetRS (which we abbreviate as VILM-RS) lacks residues required for specificity in modern aaRSs (Jabłońska *et al.* 2022). Lower LUCA enrichment for isoleucine than for valine might reflect differences in the timing of amino acid availability, rather than a shift in VILM-RS specificity. Different organisms might have used different ensembles of hydrophobic amino acids, with horizontal gene transfer between them nevertheless transferring proteins of interest. These earliest stages of the code's evolution are of course the hardest to reconstruct, and our indication of promiscuity, especially in later panels of Fig. 2, can also be read as uncertainty regarding exactly which hydrophobic amino acid(s) were more specifically encoded by which codons.

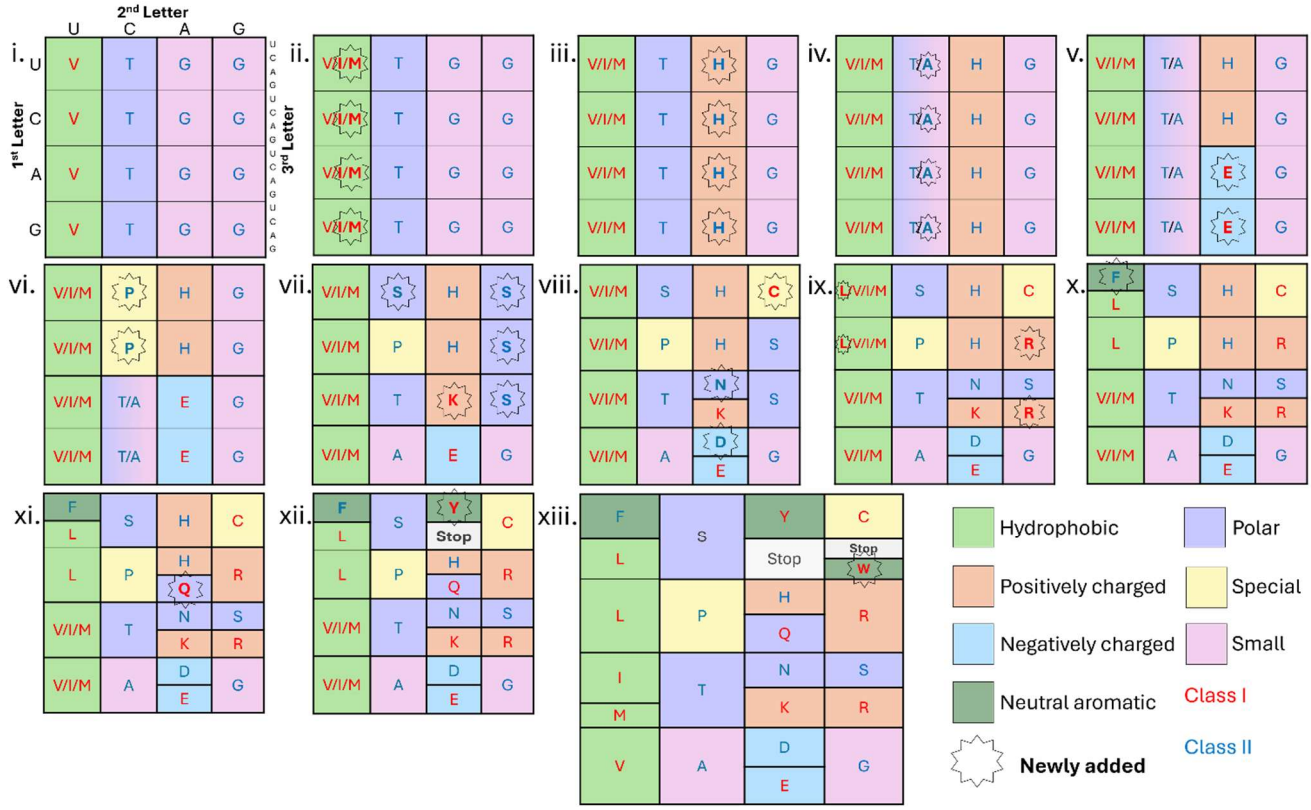


Figure 2. Hypothesized intermediate genetic codes with amino acid ordering from Table 1 LUCA clan usage. Each square indicates four codons spanning all nucleotides in the third position. Past locations of start and stop codons are left as unknown; any codon is expected to act as a stop codon when too little charged tRNA is bound. Listing multiple amino acids for the same codon can indicate either promiscuity or unknown status. "Small" amino acids glycine and alanine are deliberately assigned a color close to polar, in recognition of the polar nature of the backbone. "Neutral aromatics" include amino acids with 6-carbon rings (histidine is not always neutral). We assigned these a color close to other hydrophobic amino acids, emphasizing their row rather than column pattern. Other possible categories are amides (i.e. hydrogen bond acceptors) NQ, found in the third column and nucleophiles (i.e. hydrogen bond donors) SCKUT, found in the first and third rows. HYRDE can also serve as nucleophiles in some circumstances.

The addition of histidine in Fig. 2.iii completes a singlet code. This makes available all the amino acids for the HxGH motif of Class I aaRSs (where x can be V, I, L or M; see Figure 1 of Chaliotis *et al.* (2016)).

As with the hydrophobic amino acids, we add alanine in Fig. 2.iv as part of non-specific coding, to continue to avoid distinguishing among the two purines in the first position so early. Conventionally, threonine is classed as polar and alanine as hydrophobic, making ambiguous coding of that combination surprising. However, alanine is only a little larger

than glycine, and may similarly allow the polar backbone to play a role. Across the meta-analysis of deep mutational scanning (DMS) experiments summarized in Supplementary File 1, substitutions between alanine and threonine are well tolerated, better tolerated than between alanine or threonine and any other amino acid except the also-polar serine (Supplementary File 1), making this ambiguity plausible. Alanine and threonine have very similar effects on protein folds (Tsuboyama *et al.* 2023, Fig. 3a). Background colors in Fig. 2 correspondingly indicate both alanine and glycine as "small", deliberately using a color similar to that for polar residues.

Next, glutamate in Fig. 2.v adds a negatively charged amino acid, which might help balance histidine's positive charge, an important consideration in protein stability. Next comes proline in Fig. 2.vi.

It is unclear when the ambiguity between alanine and threonine resolves; to obtain their position in the modern code requires distinguishing between purines in the first position. This is also required when adding serine and lysine in Fig. 2.vii, and for this reason we tentatively mark the resolution of the A/T ambiguity here.

The fact that serine's codons today are not all connected by a path of single mutations is puzzling (Bernhardt 2016; Inouye *et al.* 2020; Lei and Burton 2021). We assign additional glycine codons to serine in such a way as to mutationally connect today's serine's codons, and to set up for arginine to later take all of its codons from serine. Unlike most aaRSs, SerRS (Cusack *et al.* 1998) and SerRS-arch (Bilokapic *et al.* 2006b) recognize their tRNA substrates mainly through structural and sequence features in the tRNA body rather than via the anticodon, making reassignments easier. SerRS later took back codons from arginine in a number of different mitochondrial codes (Elzanowski and Ostell 2024).

We place LysRS-I before LysRS-II (Ribas de Pouplana *et al.* 1998), noting that LysRS-II seems more prone to displace LysRS-I than vice versa, e.g. being found in both eukaryotic cytosol and mitochondria. This ordering is also more compatible with our aaRS trees below.

Given similar LUCA usage, we add cysteine, aspartate, and asparagine together in Fig. 2.viii. Aspartate and asparagine mark the first use of information (other than possibly methionine) in the third codon position. While this precedes a complete doublet code (i.e. not all information in the first position is used) this is to some extent still true even in the canonical code with respect to leucine and arginine.

In Fig. 2.ix, we introduce leucine as promiscuous/ambiguous with other hydrophobic amino acids, resolving it only with the subsequent addition of phenylalanine in Fig. 2.x. We keep valine, isoleucine, and methionine as ambiguous for even longer, for reasons discussed together with aaRS trees below. Resolution could have occurred earlier; we

have little information to go on as to when and how it occurred, and ambiguity in earlier panels should be read as an expression of ignorance. Methionine is unusual in using only one codon – the only other amino acid for which this is true, tryptophan, is the last of the standard 20 to be added. Methionine also uses a second adjacent codon in a number of mitochondrial codes (Elzanowski and Ostell 2024).

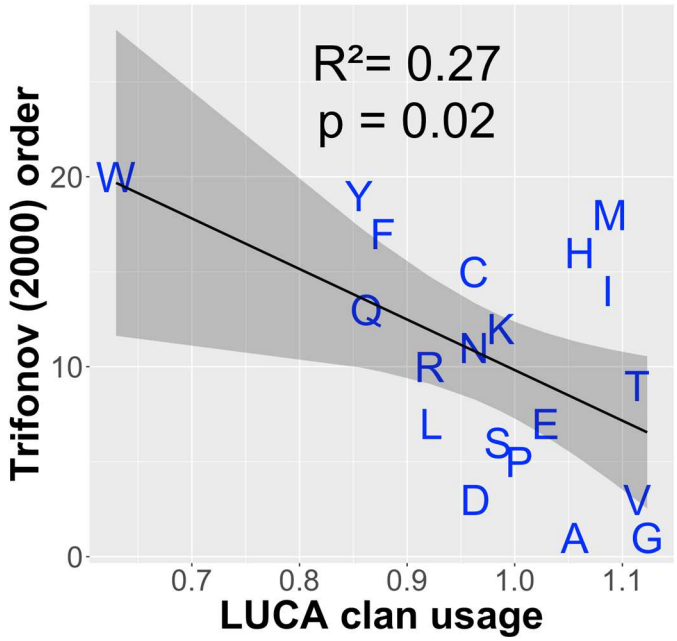
The modern position of arginine, also added in Fig. 2.ix, is unusual in that it spans a purine and a pyrimidine in the first position. In our reconstruction, all of these codons are taken from serine.

We next add glutamine, tyrosine, and tryptophan, in Fig. 2.xi-xiii. We add stop codons when adjacent positions change to using tyrosine and tryptophan. Stop codons could have once resided at quite different locations, in ways we do not know how to reconstruct.

Controls

In the section above, strictly following the order of amino acid recruitment from LUCA enrichment produced reasonable adherence to the 2-1-3 rule. The two notable exceptions suggested plausible ambiguous coding states for V/I/M and A/T. To assess whether these two observations constitute supporting evidence (consilience) for the hypothesized Fig. 2 order of events, one approach is to apply our rules to random alternative orders of amino acid recruitment. The challenge with this approach is that we exercised discretion in the assignment of early ambiguities to the genetic code. It was not clear to us how to automate the process of spotting compelling ambiguities across a large ensemble of random codes.

Instead, we apply the same manual inspection approach to just a single alternative order of amino acid recruitment: that proposed by Trifonov (2000), namely G/A, V/D, P, S, E/L, T, R, N, K, Q, I, C, H, F, M, Y, W, where "/" indicates simultaneous introduction. We note that this order correlates with ours (Supplementary Fig. 1) such that some compelling patterns might therefore be shared due to overlap. Some of the criteria Trifonov (2000) used to construct the order might also be circular with respect to the patterns in question, e.g. the relatively recent divergence of TrpRS and TyrRS was one of the criteria for assigning them as late, making aaRS trees not an independent source of support. These biases make our test more stringent than comparison to a random ordering.



316

317 **Supplementary figure 1.** Amino acid usage in LUCA correlates with Trifonov's final rank
318 order of recruitment (2000). Model 1 regression lines are shown in black with 95% CI gray
319 shading. Error in Trifonov's order is presumed to be larger, motivating its placement as the
320 dependent variable. Spearman's ρ=-0.48, p =0.03.

321

322 Despite considerable effort, we found no satisfyingly plausible ambiguities while
323 producing the series of codes shown in Figure 3. As in Figure 2, a singlet code is completed
324 prior to use of the first codon position. However, information from the third position
325 appears when the 7th and 8th amino acids are added, earlier than when the 12th-14th amino
326 acids are added in Figure 2.viii. This inferior adherence to the 2-1-3 rule for the control
327 provides some support for our Figure 2 hypothesis. A distinction between purines and
328 pyrimidines in the 1st position begins when the 6th amino acid is added in Fig. 3.iv, earlier
329 than when the 10th and 11th amino acids are added in Fig. 2.vii, again supporting our Figure
330 2 hypothesis.

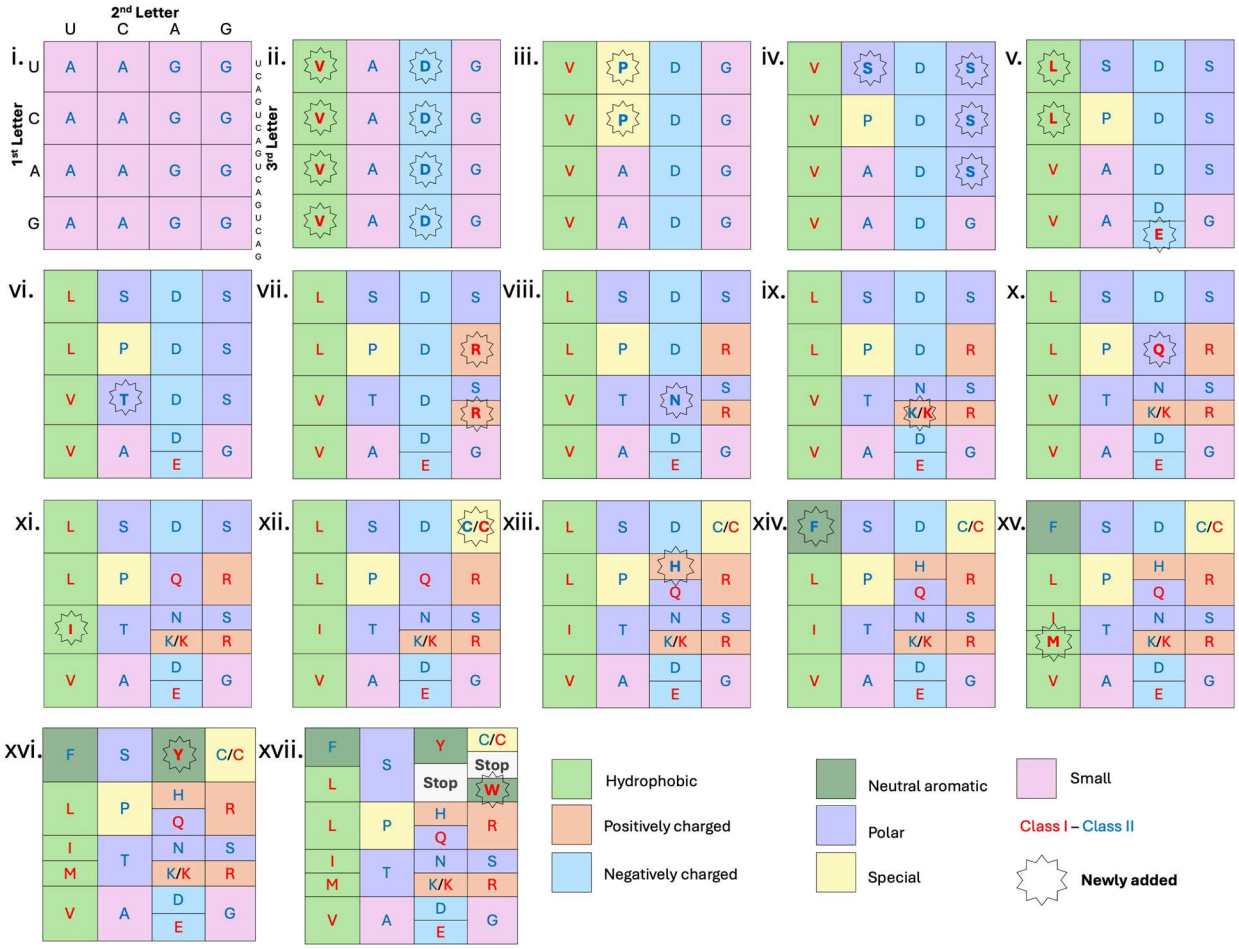


Figure 3. Intermediate genetic codes generated using Trifonov's ordering, as a control analysis. The same scheme is used as in Figure 2.

Synthetase trees

With our Figure 2 hypothesis articulated, and even supported by a control analysis, we next turn to another set of facts, those concerning the aaRSs. We also draw from information on which amino acid substitutions are well tolerated in DMS data, using the meta-analysis of Sun et al. (2024) summarized in Supplementary File 1, which was so far used only to support the plausibility of ambiguous coding of alanine and threonine.

Class I and Class II aaRSs, which have different structures (Eriani *et al.* 1990), are shown in red and blue font, respectively, in Figures 2 and 3. We expect gene duplication and divergence events within the two corresponding aaRS trees to reflect the events by which new amino acids were recruited. Trees inferred for the amino-acid binding, catalytic

domains of aaRSs (Douglas *et al.* 2024) have poor support for the deeper branches that describe the relationships between "subclasses", even as the subclass groupings and sometimes the relationships within subclasses are often well resolved. We therefore use additional information, as described in the section below, to modify low confidence branches of the Class I tree.

It seems likely that aaRS duplication and divergence events would tend to draw from "moonlighting" or promiscuous activities of the ancestral aaRS. An important but underused source of information as to which synthetases might have evolved into which therefore comes from promiscuous activities of modern aaRSs to activate the wrong amino acid. There are two arguments for using promiscuity information to infer aaRS evolution. First, promiscuities reflect amino acid similarity from the perspective of an aaRS, which we expect to correlate with the evolutionary steps that took place during the construction of the code. Second, promiscuities might also have a contingent evolutionary basis, i.e. an aaRS may be more likely to have promiscuous activity with respect to an amino acid that it used to target earlier in its evolutionary history.

Some promiscuity is suppressed by editing functions (Perona and Gruic-Sovulj 2014), which can be knocked out to reveal masked promiscuities. Specificity can also be perturbed by mutation to reveal evolutionarily adjacent promiscuities, allowing experiments to probe what might have happened during evolution. Supplementary File 2 summarizes aaRS promiscuities, some of them latent, that we collated through a literature search – each of them is also described below.

Some aaRS duplication/divergence events might split a set of codons into two, especially when the ancestral aaRS was promiscuous and its descendants specialized. Other newly evolved aaRSs might "take" codons from elsewhere; in this case we expect this shift to involve tRNAs that are related at the time (even if that tRNA phylogenetic signal is lost today), and hence for the aaRS duplication and divergence event to connect mutationally adjacent codons.

Revised tree of the Class I synthetase catalytical domain

We modify the point-estimate Class I aaRS catalytic domain tree of Douglas et al. (2024), within the range of posterior uncertainty. Most aaRSs first activate an amino acid with ATP, and then bind tRNA, but LysRS-I, ArgRS, and GluRS/GlnRS complete these steps in the opposite order (Ibba *et al.* 1999). This suggests that these aaRSs might be related. However, Douglas et al. (2024) made ArgRS sister, with 0.43 posterior support, to a subclass containing CysRS, MetRS, LeuRS, ValRS, and IleRS, and made LysRS-I sister, with 0.59 support, to a subclass containing TyrRS and TrpRS. We instead make ArgRS and

LysRS-I sisters to each other, and related next to the clade including GlxRS (which charges glutamate without discriminating between glutamate's and glutamine's tRNAs) and the discriminating GluRS and GlnRS. An ArgRS/LysRS-I clade received 0.13 support in the posterior distribution of Douglas et al. (2024). Putting ArgRS and LysRS-I together makes sense in terms of the biochemical similarity (positive charge) between arginine and lysine, and putting the pair with Glx/Gln/GluRS make sense under the presumption that aaRS differentiation tended to proceed via small steps in anticodon specificity, with glutamate codons today adjacent to lysine codons, which are in turn adjacent to arginine codons, with that order of duplication suggested by Table 1. In DMS data, mutations between glutamate and lysine are quite well tolerated, and those between lysine and arginine are extremely well tolerated (Supplementary File 1).

A key driver of LysRS-I's phylogenetic placement by Douglas et al. (2024) as sister to TyrRS/TrpRS was the shared absence of a short three-stranded Z-shaped antiparallel sheet (known simply as "Z"). In their model, the presence/absence of insertion modules in different proteins were augmented with amino acid sequence to inform the phylogeny. Each insertion module was assumed to be born once, and only once, after which patterns of its loss contribute to likelihood. Incorporating insertion modules into the analysis required judgement calls to be made about which modules were homologous and which originated independently. These modules included various domains and loops involved in aminoacylation, tRNA recognition, and editing. Here we suggest that the Z insertion module found in several Class I aaRSs is homologous to the longer LysRS-I insert (Figure 4), rather than two being unrelated as Douglas et al. (2024) proposed. We therefore performed a new Bayesian phylogenetic analysis on the Class I aaRS, identical to Douglas et al. (2024) except assuming that LysRS-I contained the Z insertion module, prior to an additional unique insertion module. As TrpRS and TyrRS are the only Class I aaRS families that lack the Z-fold, this lowers support for LysRS being sister to these two.

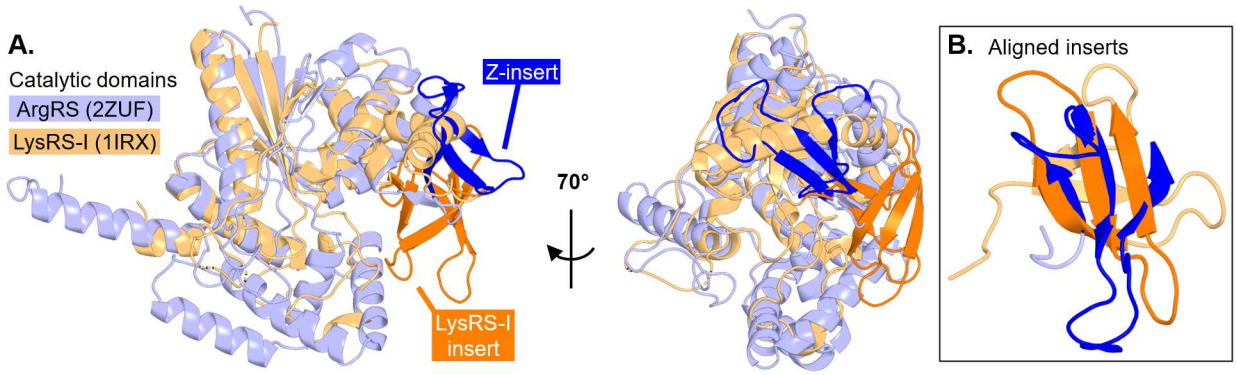


Figure 4: The LysRS-I insertion module might be derived from the Z-fold insertion module, rather than having evolved independently. The three-stranded beta sheet (Z-fold) found in many Class I aaRS (including ArgRS) resides within the same position of the catalytic domain as the beta-rich loop insertion found only in LysRS-II (Douglas *et al.* 2024). The insertion modules occupy different poses in aaRS tertiary structure (A) but share the same beta sheet topology as is apparent when manually aligned (B). The two pale beta sheets behind in LysR-I (B) are scored as a second insertion module that built on the Z-fold. Structural alignment in (A) generated using FoldMason (Gilchrist *et al.* 2026) on two crystal structures sourced from *Pyrococcus horikoshii* (Terada *et al.* 2002; Konno *et al.* 2009).

With this constraint modified, posterior support for an ArgRS/LysRS-I clade rose from 0.13 to 0.30. There was 0.16 support for our hypothesis of having them both within a clade with GlxRS (vs. 0.02 in Douglas et al. (2024)), and slightly higher 0.22 support for including both of them in a clade with the CVILM-RSs. We stick to the slightly less preferred relationship with GlxRS for the reasons given above: LysRS-I, ArgRS, and GlxRS share a mechanism in which tRNA binding precedes amino acid activation, glutamate codons are adjacent to lysine codons, and DMS data shows higher exchangeability between glutamate and lysine than between CVILM and lysine. A full list of clade posterior supports is in Supplementary File 3 for the revised Class I aaRS tree, and in Supplementary File 4 for the unchanged Class II tree from Douglas et al. (2024).

Class I synthetase tree in light of our hypothesis

We next assess how our Figure 2 hypothesis about intermediate codes aligns (i.e. is consilient) with facts that were not used to construct it. Specifically, we examine facts about aaRS trees (Figure 5), aaRS promiscuities (Supplementary File 2), and which amino acid substitutions are best tolerated across a meta-analysis of DMS experiments (Supplementary File 1 summarizing Sun *et al.* 2024).

Figure 5 shows chronograms whose branch lengths are based on both sequence and insertion modules, following the method of Douglas et al. (2024). Shading indicates LUCA as the time of divergence of archaeal vs. bacterial forms of GlxRS and LeuRS (Class I, left) and AspRS, ProRS, and PheRS (Class II, right). Agreement in each case gives some support to the use of chronograms. Given the expected antiquity of LUCA (Moody *et al.* 2024), its position about half-way in time indicates that aaRSs evolved more rapidly pre-LUCA, which is not surprising. Our use of chronograms does not assume uniform rates of evolution across time, but only that rates of evolution changed in a concerted fashion across branches under an uncorrelated relaxed clock model.

If we allow for promiscuous VIM, CVIM, and LVIM ancestors, then marking the addition of each amino acid in order (omitting glutamine and asparagine, because these amino acids were first synthesized from precursors charged to tRNA by aaRSs different from GlnRS and AsnRS) aligns well with Class I aaRS tree branch lengths in Figure 5a. Cysteine addition converting VIM-RS into CVIM-RS briefly precedes subfunctionalization of CVIM-RS into CysRS and LVIM-RS, the latter event adding leucine at the same time as arginine in Fig. 2.ix.

The chronogram suggests that the promiscuity between valine, isoleucine, and leucine did not resolve until after LUCA. ValRS and IleRS share two conserved W residues in the binding pocket that interacts with valine or isoleucine (Fukai *et al.* 2000), despite W being the last addition to the code; this would be extraordinary in the absence of ancestral promiscuity. However, it is consistent with promiscuity that resolved only after W was added to the code, which itself occurred post-LUCA (Wehbi *et al.* 2026).

The VILMC clade shows a significant evolutionary history of editing non-canonical hydrophobic amino acids, amino acids that might have been lost from the code at the time that promiscuities were lost. Even CysRS, which displays high specificity at the activation step, possesses a superfluous editing function. This may be a carryover from its ancestry, resembling the mechanism by which MetRS edits misactivated homocysteine (Jakubowski and Fersht 1981; Jakubowski 1990; Jakubowski 1994). IleRS can activate norvaline and homocysteine in addition to valine (Schmidt and Schimmel 1994; Jakubowski 1996; Bilus *et al.* 2019). Both bacterial-like LeuRS-bact (Chen *et al.* 2000; Cusack *et al.* 2000; Cvetesic *et al.* 2014) and archaeal/eukaryotic-like LeuRS-arch (Anderson and Fowden 1970; Chen *et al.* 2011) have editing functions that prevent the mischarging of isoleucine, methionine, homocysteine, norvaline, norleucine, and homoalanine. LeuRS-bact can also activate methionine. ValRS has an editing domain to prevent its promiscuous charging of non-canonical homoalanine as well as isoleucine, threonine, alanine, serine, and cysteine (Jakubowski 1980; Fukai *et al.* 2000).

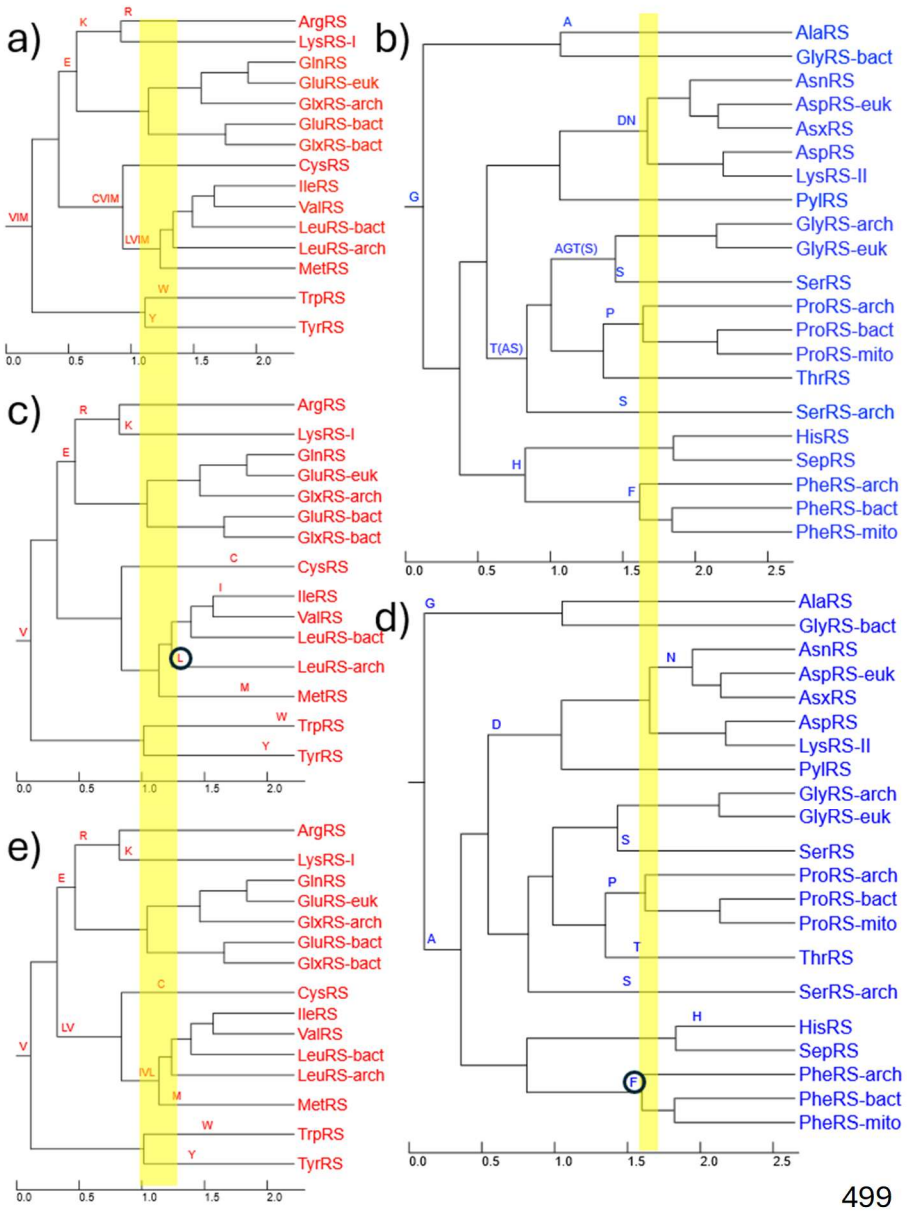


Figure 5. AaRS catalytic domain trees, annotated with amino acid order of appearance from this paper (a-b) vs. from Trifonov (c-e). Out of order appearances of L and F are circled in c) and d), with L avoided in e) through the inclusion of a promiscuity inferred from our ordering. LUCA is shown with yellow shading, on the basis of the common ancestors of archaeal and bacterial GlxRS, LeuRS, AsxRS, ProRS, and PheRS. We omit the appearance of glutamine, which began as a modification of tRNA charged with glutamine; specialized GlnRS evolved later, in eukaryotes. We similarly assume that asparagine used AsxRS prior to AsnRS. We assume that LysRS-I preceded LysRS-II, and that cysteine was first added via Class I CysRS rather than Class II SepRS; in each case the later aaRS branch arises post-LUCA. The topologies of these trees were fixed during Bayesian phylogenetic inference, while the branch lengths were estimated using the same insertion module method described by (Douglas *et al.* 2024), where the phylogeny is inferred using both amino acid sequence and the presence/absence of these modules. Branch lengths are a weighted average of births and death per insertion module, and substitutions per site.

A LeuRS urzyme activates, in order of efficiency: isoleucine, methionine, alanine, lysine, and only then leucine (Carter *et al.* 2014). Interestingly, across a set of solubility-optimized Class I urzymes, while all of the variants were highly promiscuous, there were some variants that had low activity for isoleucine, valine, and methionine, but no variants with low activity for leucine (Patra *et al.* 2025). Together, these results suggest that leucine might have been present at some frequency prior to having a specialized aaRS. Some of leucine's codons are taken by threonine in the yeast mitochondrial code, by serine in a yeast alternative nuclear code, and by alanine in the *Pachysolen tannophilus* nuclear code (Elzanowski and Ostell 2024).

Another source of data on ancestral usage of codons for these hydrophobic amino acids is a tree of the anticodon binding domain (rather than the catalytic domain as in Figure 5) of Class I and II aaRSs (Douglas and Bromham 2025). For the anticodon binding domain, there is deep divergence between LeuRS and MetRS on one side and IleRS and ValRS on the other side of an ancient node. The four squares of the first column of the code might thus have once encoded, from top to bottom, for: 1) an unknown mixture of leucine, methionine, and non-canonical amino acids such as homocysteine, 2) leucine as is still the case today, 3) isoleucine as is still the case for three of the four codons today, and 4) valine as is still the case today. In this scheme, leucine and methionine both once corresponded to a pyrimidine in the first position, and isoleucine and valine to a purine. Methionine's shift from the top square in the code (where F and L are now found) to its current position (adjacent to I) using only a single codon might then have occurred quite late, compatible with the increased information required to use only a single codon. This shift might have occurred in conjunction with methionine's codon becoming the start codon.

Figure 5a implies duplications of a promiscuous ancestor to separately spin off an ancestral form of GlxRS (taking codons from histidine), CysRS (taking codons from serine), and TyrRS (taking codons from histidine). In DMS data, histidine is very exchangeable with tyrosine and somewhat exchangeable with glutamate (Supplementary File 1). Serine is one of the amino acids that cysteine is most exchangeable with (Supplementary File 1). Tyrosine is not very exchangeable with the LVIM group of hydrophobic amino acids, but is very exchangeable with phenylalanine, which is in turn exchangeable with leucine and methionine. If one of the lost hydrophobic ancestors included phenylalanine (as well as leucine and methionine) within its potentially promiscuous range, this would make it a more plausible ancestor of TyrRS. However, while this hypothetical Class I promiscuous ancestor could have charged phenylalanine prior to the evolution of today's Class II PheRS, phenylalanine's depletion from LUCA suggests phenylalanine was little used in the early codes of LUCA's ancestors.

GlxRS duplicated and diverged to produce GlnRS (taking codons from histidine). While GlnRS lacks an editing domain, mutations reveal a latent ability to activate glutamate (Hong *et al.* 1998; Bullock *et al.* 2008). In DMS data, mutations between glutamate and glutamine are well tolerated, as are those between glutamine and histidine (Supplementary File 1).

Finally, TrpRS duplicated from TyrRS, taking codons from cysteine. Tryptophan and tyrosine are exchangeable with one another, but tryptophan is not very exchangeable with cysteine. These codons need not have transitioned directly from cysteine to tryptophan; SepRS charges tRNAs corresponding to these codons with O-phosphoserine and SecRS with selenocysteine; exchangeabilities with these amino acids are unknown.

Class II synthetase tree in light of our hypothesis

The order of amino acids aligns well with Figure 5b if we allow for promiscuous aaRS(s) capable of activating not just T and A (consilient with the 2-1-3 rule in Figure 2) but also S and G. Serine is biochemically similar to threonine, making it plausible that a latent promiscuity for serine could have pre-dated the availability of serine (or alanine). We introduce T(AS)-RS in Figure 5b to represent threonine appearing (at least via this aaRS) second only to glycine, with parentheses indicating that alanine and serine activities were latent.

We assume that the root of the Class II tree was GlyRS, given that glycine is the smallest amino acid, even though G and T are equally enriched in LUCA in Table 1. Both contemporary types of GlyRS (related only at the root of the tree) are highly specific for glycine, despite lacking an editing domain (Valencia-Sánchez *et al.* 2016). However, the evolution of ThrRS likely proceeded through a promiscuous ancestor, and this promiscuity might have persisted for some time, especially given that GlyRS-arch specificity eventually re-evolved from this TA(S)-RS ancestor. With the addition of alanine, we show AGT(S)-RS as the ancestor of both GlyRS-arch and SerRS (a clade with 85% posterior support). We show ambiguity ending in Fig. 2.vii with the appearance of serine, although there is little to inform this. There is only 61% posterior support for making SerRS-arch an outlier to a SerRS/ProRS/ThrRS/GlyRS clade, a topology that produces the least consilient branch within Figure 5, with serine arriving mid-branch rather than at a node as in the rest of Figures 5a and 5b. This makes the order of events in this clade particularly uncertain.

A reconstructed GlyRS-bact urzyme is less promiscuous than a reconstructed ancestor of LVIM-RS, but nevertheless has some promiscuous activity, in decreasing order of efficiency, for cysteine, alanine, histidine, leucine, aspartate, serine and threonine (Patra *et al.* 2024). Our Figure 5B tree suggests that GlyRS duplicated and diverged into new

specificities five times, transitioning to the promiscuous T(AS)-RS (perhaps displacing a promiscuous hydrophobic aaRS), HisRS (taking codons from glycine), AlaRS (sharing codons with threonine), PylRS (taking codons from histidine, and suggesting ancient pyrrolysine usage), and AsxRS (taking codons from glutamate, with which aspartate is highly exchangeable (Supplementary File 1), and noting that AsxRS and AspRS-bact have latent activity to bind glutamate (Rathnayake and Hendrickson 2019)). While we have no suitable data to assess pyrrolysine transitions, the other transitions agree with the ancestral promiscuities of the GlyRS-bact urzyme.

SerRS-arch can activate threonine (Bilokapic *et al.* 2006a), and ThrRS can activate serine (Dock-Bregeon *et al.* 2004), consistent with their sharing a promiscuous ancestor. The more common SerRS can promiscuously activate alanine (Schimmel 2011). This is compatible with our hypothesis that T(AS)-RS first evolved into a form of AGST-RS, before subfunctionalizing into SerRS and GlyRS-arch. Alanine and glycine are the two smallest amino acids, and serine-alanine and glycine-alanine substitutions are very well tolerated in DMS data (Supplementary File 1). The alanine activity of AGST-RS could have been displaced by the ancestor of the contemporary AlaRS, which was independently derived directly from GlyRS-bact. AlaRS can promiscuously activate glycine (Tsui and Fersht 1981), compatible with their phylogenetic relationship and with DMS data.

To assess the evolution of ProRS, we note that while proline substitutions are generally poorly tolerated, this is least true, in order, for serine, alanine, glutamine, and threonine, with transitions to proline better tolerated than transitions away, especially alanine to proline (Supplementary File 1). ProRS can also activate alanine (Beuning and Musier-Forsyth 2000). In Figure 5B, both ProRS and ThrRS evolve from a descendent of T(AS)-RS, with 91% posterior support for the clade.

AsnRS and AspRS do not have specialized editing, but mutations reveal a latent propensity to misactivate each others' targets (Sharma *et al.* 2022), and DMS data shows tolerance of mutations between them. More evidence of a connection is the existence of a "non-discriminating" (and from Figure 5B, ancestral) AsxRS, found in some bacteria and organelles, which charges aspartate to asparagine's tRNA; instead of using AsnRS, the aspartate is then converted into asparagine by an amidotransferase (Curnow *et al.* 1996). Asparagine took codons from lysine, with which it is only somewhat exchangeable (Supplementary File 1), however we note that it has since taken a third codon from lysine in some mitochondrial codes (Elzanowski and Ostell 2024).

LysRS-II duplicated and diverged from AsxRS; this does not contradict our Table 1 order, because we hypothesize LysRS-I as earlier. A broad range of amino acids escape editing after misactivation by LysRS-II, whose editing appears targeted to the non-canonical

ornithine, homocysteine, and homoserine, as well as threonine, cysteine, and alanine (Jakubowski 1997; Jakubowski 1999) . This suggests that these non-canonical amino acids might have persisted even until this relatively late, post-LUCA stage of the code's evolution. This is compatible with post-LUCA resolution of VILM promiscuity.

HisRS duplicated and diverged into SepRS post-LUCA. SepRS charges tRNA with O-phosphoserine, which is then converted to cysteine. Cysteine took codons from serine in Fig. 2, which might at one point have been O-phosphoserine before it became cysteine, and which were adjacent to HisRS's codons at the time of this duplication. One of the four codons in this square sometimes codes for selenocysteine today. SerRS attaches serine to selenocysteine's tRNA, and the serine is subsequently converted to selenocysteine. Interestingly, in eukaryotes and in many archaea, this occurs via an O-phosphoserine intermediate. SepRS has high specificity without an editing domain (Hauenstein *et al.* 2008).

The tree, Table 1, and shared aromaticity of their targets together suggest that PheRS is also a duplicate of HisRS, taking codons from leucine. Phenylalanine is quite exchangeable with leucine in DMS data and somewhat exchangeable with histidine. Interestingly, the most ancient protein sequences, those that had significantly diversified prior to LUCA, are surprisingly enriched for the four aromatic amino acids histidine, phenylalanine, tyrosine, and tryptophan, relative to sequences that arose more immediately before LUCA (Wehbi *et al.* 2024). It is possible that other, competing genetic codes incorporated all four aromatic amino acids earlier and used them more, prior to the evolution of codes to become more compatibility with one another (Vetsigian *et al.* 2006). We note that the three late-added, neutral, 6-ring, aromatic amino acids can be found in the top row of the modern genetic code, in contrast to other biochemical properties tending to correlate with column.

During literature review, we also noted more promiscuities (or lack thereof) that we list here for completeness. AlaRS, which is phylogenetically distant from everything except GlyRS-bact, can promiscuously activate serine (Tsui and Fersht 1981; Guo *et al.* 2009; Schimmel 2011). ArgRS does not have editing activity and has high specificity, at least among the canonical amino acids (Igloi and Schiefermayr 2009; Hauth *et al.* 2023). ProRS can activate cysteine (Ahel *et al.* 2002).

Control hypothesis in light of synthetase data

The Trifonov order of recruitment implies that Class II synthetases evolved before Class I, with glycine and alanine as the original two amino acids, with their synthetases presumed to diverge at the root of the tree. A key question is then which of GlyRS-bact vs. GlyRS-arch

evolved as a reversion from AlaRS. A single AlaRS→GlyRS-bact switch is more parsimonious than one GlyRS-bact→AlaRS switch plus all original lineages derived from AlaRS diverging to no longer recognizing alanine, including by reversion to GlyRS-arch. However, in light of which substitutions are tolerated, it is much more plausible that ProRS evolved from AlaRS than from GlyRS, Indeed, we already posited an AGST-RS intermediate above. ProRS (Beuning and Musier-Forsyth 2000) and SerRS (Schimmel 2011) activate alanine.

Figure 5d accommodates Trifonov's order of recruitment with more difficulty than Figure 5c accommodates ours. Without the TAS-RS promiscuity inspired by the combination of our order of recruitment and the 2-1-3 rule, the deep branch separating SerRS-arch makes even less sense. Histidine and perhaps asparagine additions are pushed post-LUCA, with questions about the nature of the common ancestors of HisRS and SepRS that do not arise for our ordering. Assessing the plausibility of duplications is harder because most stem from the lost AlaRS, whose urzyme cannot be constructed to determine its promiscuity.

Similarly, without VILM promiscuity, leucine appears too late in Figure 5c. Isoleucine also appears so late as to place the subsequent appearance of methionine, tyrosine, and tryptophan, as well as CysRS, in odd places. Even when we allow somewhat arbitrarily for promiscuous VI and IVL ancestors in Figure 5e – notably without inspiration from the 2-1-3 rule as was the case for our own order of recruitment – the timing of CysRS, TyrRS and TrpRS remains odd.

The Trifonov ordering places ArgRS before LysRS-I, the opposite order to ours. In DMS data, mutations between glutamate and lysine are much better tolerated than those between glutamate and arginine (Supplementary File 1), making our placement of arginine before lysine a better fit than Trifonov's, by positing that LysRS-I evolved from GlxRS rather than ArgRS.

Apparent retrofunctionalization reflects aaRSs that acquired specificity late

It has been proposed that late-added amino acids tend to recruit older, less specific aaRSs, in a process called retrofunctionalization (Jabłońska *et al.* 2022; Douglas *et al.* 2024). In Figure 6, we repeat the analysis, quantifying the specialization of an aaRS by the number of insertion modules it has gained during evolution (Douglas *et al.* 2024). As well as revising the order of amino acid recruitment, we make a number of other refinements described in the Figure 6 legend. Loss of significance is driven by our choice to use the GlyRS-bact distance rather than the mean of GlyRS-bact (distance 2), GlyRS-arch (distance 6), and GlyRS-euk (distance 7).

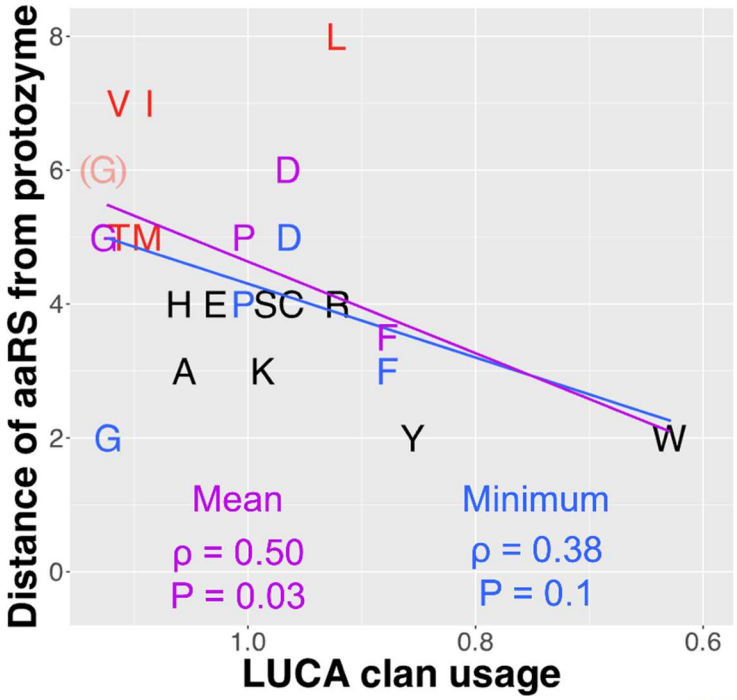


Figure 6. The retro-functionalization signal is driven by the late resolution of promiscuous aaRSs that handle early amino acids. The Class I and II protozymes are catalytically active truncations of the respective aaRS catalytic domain, and are thought to be ancestral to all modern-day aaRS (Martinez-Rodriguez *et al.* 2015; Onodera *et al.* 2021). Distance is the number of insertion modules (defined as a conserved structural element with at least 30 contiguous amino acids) acquired within the catalytic domain since the ancestral structure (the protozyme), as inferred by Douglas et al. (2024) except with one additional insertion in LysRS-I as discussed in Figure 4. A full list of distances is in Supplementary File 5. Some aaRSs come in different forms with different distances. Douglas et al. (2024) took the mean (purple), whereas we take the minimum (blue). Amino acids that we hypothesize to have been added to the code through activation by a promiscuous aaRS are shown in red. We show promiscuous GlyRS-arch in parentheses, with its relative GlyRS-euk having an even larger distance of 7; their only inclusion in statistics is via the purple mean G. Loss of statistical support for the retrofunctionalization signal is driven by our treatment of GlyRS. We are interested in first appearances, and so in both panels, we include LysRS-I but not LysRS-II. We omit Q and N because they were originally added to the code via chemical modification of E and D after the latter amino acids were charged to tRNA. Douglas et al. (2024) included pyrrolysine as the 21st amino acid; we exclude it given the lack of LUCA clan usage data. Pearson's ρ is shown. Using ties in the order of recruitment as represented in Fig. 2, Spearman's ρ=0.37, p=0.1 for the minima and ρ=0.51, p=0.03 for the means.

Our hypothesis of promiscuous ancestral aaRSs posits that specificity evolved late for LeuRS, ValRS, IleRS, MetRS, ThrRS, and the GlyRS-arch/GlyRS-euk clade, long after their target amino acids (with the exception of leucine) were added early to the code by promiscuous aaRS activation. In Figure 6a, we see that these are the aaRSs (marked in red) that were previously responsible for the apparent pattern of retrofunctionalization through

their accumulation of many insertion modules not found in the protozyme. While this is responsible for most of the apparent retrofunctionalization pattern, it is also true that TrpS and TyrRS have few insertion modules.

Conclusions

We assumed an order of amino acid recruitment on the basis of usage in LUCA, then strictly adhered to it while mapping out a series of prior genetic codes guided by the 2-1-3 rule. We achieved closer alignment with the 2-1-3 rule with this amino acid ordering than with a past hypothesized ordering. Notable deviations from the 2-1-3 rule suggested ancestrally promiscuous coding for valine/isoleucine/methionine and for threonine/alanine. The same hypothesized promiscuities were key to mapping the order of amino acid recruitment onto cladograms of the Class I and Class II aaRS catalytic (amino acid binding) domains, a mapping that was not well achieved using the previously hypothesized ordering. Both the aaRS trees and the intermediate genetic code tables align well with data on aaRS promiscuous activities. These highly consilient observations provide preliminary support for our hypothesized series of events in the evolution of the genetic code. They also provide a framework for analyzing alternative orderings that might be proposed in the future.

Testable predictions

Our hypothesized series of events is specific enough to guide the design of future studies to test it. We briefly outline three paths forward.

First, studying the properties of peptides or short proteins that are randomly constructed from limited alphabets (Tanaka *et al.* 2010; Tretyachenko *et al.* 2020; Tretyachenko *et al.* 2022; Makarov *et al.* 2023) has already proved instructive. However, such experiments are expensive, and there is a combinatorially enormous variety of amino acid compositions that could be investigated. Experiments could be made still more instructive if guided by well-motivated hypotheses regarding the likely composition of early proteins. If our working model is correct, then we predict that random proteins constructed from combinations of amino acids that follow Table 2, with stoichiometry following the number of codons, will have particularly advantageous properties. E.g., the early histidine that we propose might be more useful for metal-binding and catalysis, and when positively charged, for binding negatively charged ATP , RNA, or DNA, than the aspartate proposed by Trifonov (2000) as the first (negatively) charged amino acid. Computational approaches could explore a broader variety of compositions, with some evidence that ESMFold might be suitable for such purposes (Dutta *et al.* 2025; Sahakyan *et al.* 2025), or else other

predictive tools might be developed (Buchel *et al.* 2026). Our working model also predicts which promiscuities, forming statistical proteins, will be best tolerated.

Second, our hypothesis can guide both the design and the interpretation of experiments that reconstruct ancestral aaRSs (urzymes), and test for promiscuity. We discuss the interpretation of past experiments along these lines, but our working model helps illustrate which key ancestral nodes remain unexplored.

A third way to test predictions is to fit non-stationary models of amino acid (Douglas *et al.* 2025; McShea *et al.* 2025) or codon substitution, to test whether ancient proteins are still adjusting to changes to the genetic code that occurred after they emerged (Jordan *et al.* 2005). Key differences between Fig. 2 and Fig. 3 are particularly interesting. E.g., Figure 2 predicts that ancient proteins will be enriched for transitions from histidine to glutamine, whereas Figure 3 predicts enrichment in the opposite direction.

Data availability

Data are provided in supplementary files and Table 1. Code is provided at https://github.com/sawsanwehbi/Genetic-Code-Construction.git.

Acknowledgements: We thank Andrew Wheeler for contributions to early attempts at Figure 2. We thank Andrew Wheeler and Aleksandar Radakovic for helpful discussions. We thank NSF [DEB-2333243] for funding J.M., and the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program [80NSSC24K0384] and the Robert & Taylor Peterson Extraterrestrial Research Fund at the Arizona Astrobiology Center for funding S.W. J.D. was supported by the Australian Research Council. H.S.M. was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology.

References

Ahel, I., C. Stathopoulos, A. Ambrogelly, A. Sauerwald, H. Toogood *et al.*, 2002 Cysteine Activation Is an Inherent in Vitro Property of Prolyl-tRNA Synthetases. Journal of Biological Chemistry 277**:** 34743-34748.

Anderson, J. W., and L. Fowden, 1970 Properties and substrate specificity of the leucyl-, the threonyl- and the valyl-transfer-ribonucleic acid synthetases from Aesculus species. Biochemical Journal 119**:** 691-697.

Ardell, D. H., 1998 On Error Minimization in a Sequential Origin of the Standard Genetic Code. Journal of Molecular Evolution 47**:** 1-13.

Bernhardt, H. S., 2016 Clues to tRNA Evolution from the Distribution of Class II tRNAs and Serine Codons in the Genetic Code. Life 6**:** 10.

Beuning, P. J., and K. Musier-Forsyth, 2000 Hydrolytic editing by a class II aminoacyl-tRNA synthetase. Proc. Natl. Acad. Sci. USA 97**:** 8916-8920.

Bhattacherjee, A., and S. Wallin, 2012 Coupled Folding-Binding in a Hydrophobic/Polar Protein Model: Impact of Synergistic Folding and Disordered Flanks. Biophysical Journal 102**:** 569-578.

Bilokapic, S., T. Maier, D. Ahel, I. Gruic-Sovulj, D. Söll *et al.*, 2006a Structure of the unusual seryl-tRNA synthetase reveals a distinct zinc-dependent mode of substrate recognition. The EMBO Journal 25**:** 2498-2509.

Bilokapic, S., T. Maier, D. Ahel, I. Gruic-Sovulj, D. Söll *et al.*, 2006b Structure of the unusual seryl-tRNA synthetase reveals a distinct zinc-dependent mode of substrate recognition. The EMBO Journal 25**:** 2498-2509-2509.

Bilus, M., M. Semanjski, M. Mocibob, I. Zivkovic, N. Cvetesic *et al.*, 2019 On the Mechanism and Origin of Isoleucyl-tRNA Synthetase Editing against Norvaline. J. Mol. Biol. 431**:** 1284-1297.

Buchel, F., T. Neuwirthova, T. Tureckiova, G. Fuertes, A. Benda *et al.*, 2026 Learning the structural diversity in random protein sequence space. bioRxiv**:** 2026.2004.2030.722084.

Bullock, T. L., A. Rodríguez-Hernández, E. M. Corigliano and J. J. Perona, 2008 A rationally engineered misacylating aminoacyl-tRNA synthetase. Proc. Natl. Acad. Sci. USA 105**:** 7428-7433.

Carter, C. W., L. Li, V. Weinreb, M. Collier, K. Gonzalez-Rivera *et al.*, 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.

Cavalcanti, A. R. O., E. Soares Leite, B. B. Neto and R. Ferreira, 2004 On the Classes of Aminoacyl-tRNA Synthetases, Amino Acids and the Genetic Code. Origins of life and evolution of the biosphere 34**:** 407-420.

Chaliotis, A., P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker *et al.*, 2016 The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 45**:** 1059-1068.

Chen, J.-F., N.-N. Guo, T. Li, E.-D. Wang and Y.-L. Wang, 2000 CP1 Domain in Escherichia coli Leucyl-tRNA Synthetase Is Crucial for Its Editing Function. Biochemistry 39**:** 6726-6731.

Chen, X., J.-J. Ma, M. Tan, P. Yao, Q.-H. Hu *et al.*, 2011 Modular pathways for editing non-cognate amino acids by human cytoplasmic leucyl-tRNA synthetase. Nucleic Acids Res. 39**:** 235-247.

Crick, F. H. C., 1968 The origin of the genetic code. J. Mol. Biol. 38**:** 367-379.

Curnow, A. W., M. Ibba and D. Söll, 1996 tRNA-dependent asparagine formation. Nature 382**:** 589-590.

Cusack, S., A. Yaremchuk, I. Krikliviy and M. Tukalo, 1998 $tRNA^{Pro}$ anticodon recognition by *Thermus thermophilus* prolyl-tRNA synthetase. Structure 6**:** 101-108.

Cusack, S., A. Yaremchuk and M. Tukalo, 2000 The 2 Å crystal structure of leucyl-tRNA synthetase and its complex with a leucyl-adenylate analogue. The EMBO Journal 19**:** 2351-2361.

Cvetesic, N., A. Palencia, I. Halasz, S. Cusack and I. Gruic-Sovulj, 2014 The physiological target for LeuRS translational quality control is norvaline. The EMBO Journal 33: 1639-1653-1653.

Di Giulio, M., 1997 On the RNA World: Evidence in Favor of an Early Ribonucleopeptide World. Journal of Molecular Evolution 45: 571-578.

Dill, K. A., 1985 Theory for the folding and stability of globular proteins. Biochemistry 24: 1501-1509.

Dill, K. A., 1990 Dominant forces in protein folding. Biochemistry 29: 7133-7155.

Dock-Bregeon, A.-C., B. Rees, A. Torres-Larios, G. Bey, J. Caillet *et al.*, 2004 Achieving Error-Free Translation: The Mechanism of Proofreading of Threonyl-tRNA Synthetase at Atomic Resolution. Molecular Cell 16: 375-386.

Douglas, J., R. Bouckaert, C. W. Carter, Jr and Peter R. Wills, 2024 Enzymic recognition of amino acids drove the evolution of primordial genetic codes. Nucleic Acids Res. 52: 558-571.

Douglas, J., R. Bouckaert, C. W. Carter and P. R. Wills, 2025 Reduced Amino Acid Substitution Matrices Find Traces of Ancient Coding Alphabets in Modern Day Proteins. Mol. Biol. Evol. 42: msaf197.

Douglas, J., and L. Bromham, 2025 Reconstructing substitution histories on phylogenies, with accuracy, precision, and coverage. bioRxiv: 2025.2012.2021.695861.

Dutta, S., M. Makarov, R. Krystufek, E. Manriquez-Sandoval, S. D. Fried *et al.*, 2025 BPS2025 - A primordial acidic amino acid alphabet plausibly served as a stepping stone to foldable polymers: A computational study. Biophysical Journal 124: 216a.

Elzanowski, A., and J. Ostell, 2024 The Genetic Codes. url=https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

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

Fitch, W. M., and K. Upper, 1987 The Phylogeny of tRNA Sequences Provides Evidence for Ambiguity Reduction in the Origin of the Genetic Code. Cold Spring Harbor Symposia on Quantitative Biology 52: 759-767.

Fournier, G. P., C. P. Andam, E. J. Alm and J. P. Gogarten, 2011 Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life. Origins of Life and Evolution of Biospheres 41: 621-632.

Frank, H. S., and M. W. Evans, 1945 Free Volume and Entropy in Condensed Systems III. Entropy in Binary Liquid Mixtures; Partial Molal Entropy in Dilute Solutions; Structure and Thermodynamics in Aqueous Electrolytes. The Journal of Chemical Physics 13: 507-532.

Freeland, S. J., and L. D. Hurst, 1998 The Genetic Code Is One in a Million. Journal of Molecular Evolution 47: 238-248.

Freeland, S. J., T. Wu and N. Keulmann, 2003 The Case for an Error Minimizing Standard Genetic Code. Origins of life and evolution of the biosphere 33: 457-477.

Fried, S. D., K. Fujishima, M. Makarov, I. Cherepashuk and K. Hlouchova, 2022 Peptides before and during the nucleotide world: an origins story emphasizing cooperation

between proteins and nucleic acids. Journal of The Royal Society Interface 19: 20210641.
Fukai, S., O. Nureki, S.-i. Sekine, A. Shimada, J. Tao *et al.*, 2000 Structural Basis for Double-Sieve Discrimination of L-Valine from L-Isoleucine and L-Threonine by the Complex of tRNAVal and Valyl-tRNA Synthetase. Cell 103: 793-803.
Gilchrist, C. L. M., M. Mirdita and M. Steinegger, 2026 Multiple protein structure alignment at scale with FoldMason. Science 391: 485-488.
Guo, M., Y. E. Chong, R. Shapiro, K. Beebe, X.-L. Yang *et al.*, 2009 Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. Nature 462: 808-812.
Haig, D., and L. D. Hurst, 1991 A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution 33: 412-417.
Hauenstein, S. I., Y.-M. Hou and J. J. Perona, 2008 The Homotetrameric Phosphoseryl-tRNA Synthetase from *Methanosarcina mazei* Exhibits Half-of-the-sites Activity. Journal of Biological Chemistry 283: 21997-22006.
Hauth, F., D. Funck and J. S. Hartig, 2023 A standalone editing protein deacylates mischarged canavanyl-tRNAArg to prevent canavanine incorporation into proteins. Nucleic Acids Res. 51: 2001-2010.
Higgs, P. G., 2009 A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biology Direct 4: 16.
Hong, K.-W., M. Ibba and D. Söll, 1998 Retracing the evolution of amino acid specificity in glutaminyl-tRNA synthetase. FEBS Lett. 434: 149-154.
Honig, B., and F. E. Cohen, 1996 Adding backbone to protein folding: why proteins are polypeptides. Folding and Design 1: R17-R20.
Ibba, M., H. C. Losey, Y. Kawarabayasi, H. Kikuchi, S. Bunjun *et al.*, 1999 Substrate recognition by class I lysyl-tRNA synthetases: A molecular basis for gene displacement. Proc. Natl. Acad. Sci. USA 96: 418-423.
Igloi, G. L., and E. Schiefermayr, 2009 Amino acid discrimination by arginyl-tRNA synthetases as revealed by an examination of natural specificity variants. The FEBS Journal 276: 1307-1318.
Inouye, M., R. Takino, Y. Ishida and K. Inouye, 2020 Evolution of the genetic code; Evidence from serine codon use disparity in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 117: 28572-28575.
Itzkovitz, S., and U. Alon, 2007 The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17: 405-412.
Jabłońska, J., Y. Chun-Chen, L. M. Longo, D. S. Tawfik and I. Gruic-Sovulj, 2022 The evolutionary history of class I aminoacyl-tRNA synthetases indicates early statistical translation. bioRxiv: 2022.2006.2009.495570.
Jakubowski, H., 1980 Valyl-tRNA synthetase from yellow lupin seeds: hydrolysis of the enzyme-bound noncognate aminoacyl adenylate as a possible mechanism of increasing specificity of the aminoacyl-tRNA synthetase. Biochemistry 19: 5071-5078.
Jakubowski, H., 1990 Proofreading in vivo: editing of homocysteine by methionyl-tRNA synthetase in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 87: 4504-4508.

Jakubowski, H., 1994 Editing function of *Escherichia coli* cysteinyl-tRNA synthetase: cyclization of cysteine to cysteine thiolactone. Nucleic Acids Res. 22**:** 1155-1160.
Jakubowski, H., 1996 Proofreading in Trans by an Aminoacyl-tRNA Synthetases Model for Single Site Editing by Isoleucyl-tRNA Synthetase. Nucleic Acids Res. 24**:** 2505-2510.
Jakubowski, H., 1997 Aminoacyl Thioester Chemistry of Class II Aminoacyl-tRNA Synthetases. Biochemistry 36**:** 11077-11085.
Jakubowski, H., 1999 Misacylation of tRNALys with Noncognate Amino Acids by Lysyl-tRNA Synthetase. Biochemistry 38**:** 8088-8093.
Jakubowski, H., and A. R. Fersht, 1981 Alternative pathways for editing non-cognate amino acids by aminoacyl-tRNA synthetases. Nucleic Acids Res. 9**:** 3105-3117.
Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin *et al.*, 2005 A universal trend of amino acid gain and loss in protein evolution. Nature 433**:** 633-638.
Jukes, T. H., 1974 On the possible origin and evolution of the genetic code. Origins of life 5**:** 331-350.
Kaiser, F., S. Bittrich, S. Salentin, C. Leberecht, V. J. Haupt *et al.*, 2018 Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases. PLoS Comput. Biol. 14**:** e1006101.
Kamtekar, S., J. M. Schiffer, H. Xiong, J. M. Babik and M. H. Hecht, 1993 Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. Science 262**:** 1680-1685.
Klipcan, L., and M. Safro, 2004 Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. Journal of Theoretical Biology 228**:** 389-396.
Kollmar, M., and S. Mühlhausen, 2017 How tRNAs dictate nuclear codon reassignments: Only a few can capture non-cognate codons. RNA Biology 14**:** 293-299.
Kondratyeva, L. G., M. S. Dyachkova and A. V. Galchenko, 2022 The Origin of Genetic Code and Translation in the Framework of Current Concepts on the Origin of Life. Biochemistry (Moscow) 87**:** 150-169.
Konno, M., T. Sumida, E. Uchikawa, Y. Mori, T. Yanagisawa *et al.*, 2009 Modeling of tRNA-assisted mechanism of Arg activation based on a structure of Arg-tRNA synthetase, tRNA, and an ATP analog (ANP). The FEBS Journal 276**:** 4763-4779.
Koonin, E. V., and A. S. Novozhilov, 2009 Origin and evolution of the genetic code: The universal enigma. IUBMB Life 61**:** 99-111.
Lee, B. J., P. J. Worland, J. N. Davis, T. C. Stadtman and D. L. Hatfield, 1989 Identification of a selenocysteyl-tRNA^{Ser} in mammalian cells that recognizes the nonsense codon, UGA. Journal of Biological Chemistry 264**:** 9724-9727.
Lei, L., and Z. F. Burton, 2021 Evolution of the genetic code. Transcription 12**:** 28-53.
Makarov, M., A. C. Sanchez Rocha, R. Krystufek, I. Cherepashuk, V. Dzmitruk *et al.*, 2023 Early Selection of the Amino Acid Alphabet Was Adaptively Shaped by Biophysical Constraints of Foldability. Journal of the American Chemical Society 145**:** 5320-5329.
Martinez-Rodriguez, L., O. Erdogan, M. Jimenez-Rodriguez, K. Gonzalez-Rivera, T. Williams *et al.*, 2015 Functional Class I and II Amino Acid-activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. Journal of Biological Chemistry 290**:** 19710-19725.

Massey, S. E., 2006 A Sequential “2-1-3” Model of Genetic Code Evolution That Explains Codon Constraints. Journal of Molecular Evolution 62**:** 809-810.

Massey, S. E., 2008 A Neutral Origin for Error Minimization in the Genetic Code. Journal of Molecular Evolution 67**:** 510-516.

Massey, S. E., 2016 The neutral emergence of error minimized genetic codes superior to the standard genetic code. Journal of Theoretical Biology 408**:** 237-242.

McShea, H. S., A. L. Wheeler, P. Goodman, C. Weibel, S. Wehbi *et al.*, 2025 The effectiveness of selection in a species affects the direction of amino acid frequency evolution. bioRxiv**:** 2023.2002.2001.526552.

Moody, E. R. R., S. Álvarez-Carretero, T. A. Mahendrarajah, J. W. Clark, H. C. Betts *et al.*, 2024 The nature of the last universal common ancestor and its impact on the early Earth system. Nat. Ecol. Evol. 8**:** 1654-1666.

Müller, F., L. Escobar, F. Xu, E. Węgrzyn, M. Nainytė *et al.*, 2022 A prebiotically plausible scenario of an RNA–peptide world. Nature 605**:** 279-284.

Onodera, K., N. Suganuma, H. Takano, Y. Sugita, T. Shoji *et al.*, 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.

Osawa, S., and T. H. Jukes, 1989 Codon reassignment (codon capture) in evolution. Journal of Molecular Evolution 28**:** 271-278.

Patel, A., 2005 The triplet genetic code had a doublet predecessor. Journal of Theoretical Biology 233**:** 527-532.

Patra, Sourav K., J. Douglas, Peter R. Wills, L. Betts, Tang G. Qing *et al.*, 2024 A genomic database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class II urzymes. Nucleic Acids Res. 52**:** 13305-13324.

Patra, S. K., N. Randolph, B. Kuhlman, H. Dieckhaus, L. Betts *et al.*, 2025 Aminoacyl-tRNA synthetase urzymes optimized by deep learning behave as a quasispecies. Structural Dynamics 12**:** 024701.

Perona, J. J., and I. Gruic-Sovulj, 2014 Synthetic and Editing Mechanisms of Aminoacyl-tRNA Synthetases, pp. 1-41 in *Aminoacyl-tRNA Synthetases in Biology and Medicine*, edited by S. Kim. Springer Netherlands, Dordrecht.

Polycarpo, C., A. Ambrogelly, A. Bérubé, S. M. Winbush, J. A. McCloskey *et al.*, 2004 An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. Proc. Natl. Acad. Sci. USA 101**:** 12450-12454.

Rathnayake, U. M., and T. L. Hendrickson, 2019 Bacterial Aspartyl-tRNA Synthetase Has Glutamyl-tRNA Synthetase Activity. Genes 10**:** 262.

Ribas de Pouplana, L., R. J. Turner, B. A. Steer and P. Schimmel, 1998 Genetic code origins: tRNAs older than their synthetases? Proc. Natl. Acad. Sci. USA 95**:** 11295-11300.

Ribeiro, A. J. M., J. D. Tyzack, N. Borkakoti, G. L. Holliday and J. M. Thornton, 2020 A global analysis of function and conservation of catalytic residues in enzymes. Journal of Biological Chemistry 295**:** 314-324.

Rubio Gomez, M. A., and M. Ibba, 2020 Aminoacyl-tRNA synthetases. RNA 26**:** 910-936.

Sahakyan, H., S. G. Babajanyan, Y. I. Wolf and E. V. Koonin, 2025 In silico evolution of globular protein folds from random sequences. Proc. Natl. Acad. Sci. USA 122**:** e2509015122.

Schimmel, P., 2011 Mistranslation and its control by tRNA synthetases. Phil. Trans. R. Soc. B 366**:** 2965-2971.
Schmidt, E., and P. Schimmel, 1994 Mutational Isolation of a Sieve for Editing in a Transfer RNA Synthetase. Science 264**:** 265-267.
Sella, G., and D. H. Ardell, 2006 The Coevolution of Genes and Genetic Codes: Crick's Frozen Accident Revisited. Journal of Molecular Evolution 63**:** 297-313.
Shalvarjian, K. E., G. L. Chadwick, P. I. Pérez, P. H. Woods, V. J. Orphan *et al.*, 2025 Methanogenic archaea encoding Pyrrolysine maintain ambiguous amber codon usage. Proc. Natl. Acad. Sci. USA 122**:** e2517473122.
Sharma, V. K., S. Gupta, J. Chhibber-Goel, M. Yogavel and A. Sharma, 2022 A single amino acid substitution alters activity and specificity in Plasmodium falciparum aspartyl & asparaginyl-tRNA synthetases. Molecular and Biochemical Parasitology 250**:** 111488.
Sonneborn, T. M., 1965 Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications, pp. 377-397 in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel. Academic Press.
Stoltzfus, A., and L. Y. Yampolsky, 2007 Amino Acid Exchangeability and the Adaptive Code Hypothesis. Journal of Molecular Evolution 65**:** 456-462.
Sun, M., A. Stoltzfus and D. M. McCandlish, 2024 A fitness distribution law for amino-acid replacements. bioRxiv**:** 2024.2010.2011.617952.
Tanaka, J., N. Doi, H. Takashima and H. Yanagawa, 2010 Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Protein Science 19**:** 786-795.
Terada, T., O. Nureki, R. Ishitani, A. Ambrogelly, M. Ibba *et al.*, 2002 Functional convergence of two lysyl-tRNA synthetases with unrelated topologies. Nature Structural Biology 9**:** 257-262.
Travers, A., 2006 The Evolution of the Genetic Code Revisited. Origins of Life and Evolution of Biospheres 36**:** 549-555.
Tretyachenko, V., V. Voráček, R. Souček, K. Fujishima and K. Hlouchová, 2020 CoLiDe: Combinatorial Library Design tool for probing protein sequence space. Bioinformatics 37**:** 482-489.
Tretyachenko, V., J. Vymětal, T. Neuwirthová, J. Vondrášek, K. Fujishima *et al.*, 2022 Modern and prebiotic amino acids support distinct structural profiles in proteins. Open Biology 12**:** 220040.
Trifonov, E. N., 2000 Consensus temporal order of amino acids and evolution of the triplet code. Gene 261**:** 139-151.
Tsuboyama, K., J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani *et al.*, 2023 Mega-scale experimental analysis of protein folding stability in biology and design. Nature 620**:** 434-444.
Tsui, W.-C., and A. R. Fersht, 1981 Probing the principles of amino acid selection using the alanyl-tRNA synthetase from Escherichia coli. Nucleic Acids Res. 9**:** 4627-4637.
Valencia-Sánchez, M. I., A. Rodríguez-Hernández, R. Ferreira, H. A. Santamaría-Suárez, M. Arciniega *et al.*, 2016 Structural Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. Journal of Biological Chemistry 291**:** 14430-14446.

Van der Gulik, P. T. S., and D. Speijer, 2015 How Amino Acids and Peptides Shaped the RNA World. Life 5**:** 230-246.

Vetsigian, K., C. Woese and N. Goldenfeld, 2006 Collective evolution and the genetic code. Proc. Natl. Acad. Sci. USA 103**:** 10696-10701.

Wehbi, S., N. Ly-Trong, A. Wheeler, B. Q. Minh, D. Lauretta *et al.*, 2026 Tryptophan became part of the universal genetic code post-LUCA. bioRxiv**:** 2026.2005.2028.728509.

Wehbi, S., A. Wheeler, B. Morel, B. Q. Minh, D. S. Lauretta *et al.*, 2024 Order of amino acid recruitment into the genetic code resolved by Last Universal Common Ancestor's protein domains. Proc. Natl. Acad. Sci. USA 121**:** e2410311121.

Wills, P. R., K. Nieselt and J. S. McCaskill, 2015 Emergence of Coding and its Specificity as a Physico-Informatic Problem. Origins of Life and Evolution of Biospheres 45**:** 249-255.

Woese, C. R., 1965 On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54**:** 1546-1552.

Wong, J. T.-F., 1975 A Co-Evolution Theory of the Genetic Code. Proc. Natl. Acad. Sci. USA 72**:** 1909-1912.

Wu, H.-L., S. Bagby and J. M. H. van den Elsen, 2005 Evolution of the Genetic Triplet Code via Two Types of Doublet Codons. Journal of Molecular Evolution 61**:** 54-64.