
# A biophysical basis for the emergence of the genetic code in protocells

Stuart A. Harrison, Raquel Nunes Palmeira, Aaron Halpern, Nick Lane

PII: S0005-2728(22)00066-4

DOI: https://doi.org/10.1016/j.bbabio.2022.148597

Reference: BBABIO 148597

To appear in: *BBA - Bioenergetics*

Received date: 12 April 2022

Revised date: 27 June 2022

Accepted date: 13 July 2022

Please cite this article as: S.A. Harrison, R.N. Palmeira, A. Halpern, et al., A biophysical basis for the emergence of the genetic code in protocells, *BBA - Bioenergetics* (2022), https://doi.org/10.1016/j.bbabio.2022.148597

# A biophysical basis for the emergence of the genetic code in protocells

Stuart A Harrison, Raquel Nunes Palmeira, Aaron Halpern, Nick Lane.

Department of Genetics, Evolution and Environment

University College London

Darwin Building, Gower Street, London WC1E 6BT

## Abstract

The origin of the genetic code is an abiding mystery in biology. Hints of a 'code within the codons' suggest biophysical interactions, but these patterns have resisted interpretation. Here, we present a new framework, grounded in the autotrophic growth of protocells from $CO_2$ and $H_2$. Recent work suggests that the universal core of metabolism recapitulates a thermodynamically favoured protometabolism right up to nucleotide synthesis. Considering the genetic code in relation to an extended protometabolism allows us to predict most codon assignments. We show that the first letter of the codon corresponds to the distance from $CO_2$ fixation, with amino acids encoded by the purines (G followed by A) being closest to $CO_2$ fixation. These associations suggest a purine-rich early metabolism with a restricted pool of amino acids. The second position of the anticodon corresponds to the hydrophobicity of the amino acid encoded. We combine multiple measures of hydrophobicity to show that this correlation holds strongly for early amino acids but is weaker for later species. Finally, we demonstrate that redundancy at the third position is not randomly distributed around the code: non-redundant amino acids can be assigned based on size, specifically length. We attribute this to additional stereochemical interactions at the anticodon. These rules imply an iterative expansion of the genetic code over time with codon assignments depending on both distance from $CO_2$ and biophysical interactions between nucleotide sequences and amino acids. In this way the earliest RNA polymers could produce non-random peptide sequences with selectable functions in autotrophic protocells.

## Introduction

The genetic code offers tantalising clues to its own origins - a ‘code within the codons’ [1]. The first base in the triplet codon links the amino acid encoded with a shared precursor, implying a relationship between codons and biosynthetic pathways [1–3]. The second position of the anticodon is associated with the hydrophobicity of the amino acid encoded [4,5]. These associations have been identified both in biological systems with anticodon enrichments in aptamers [6,7] and through experimental methods [8–11]. Experimental identification of interactions however is difficult to definitively prove. The third position is agreed to have less coding information but has been linked to amino acid ‘complexity’ and molecular size before [1,12]. These patterns imply that direct biophysical interactions between amino acids and their cognate codons or anticodons fashioned at least some of the genetic code. Both the co-evolution and stereochemical hypotheses have complex histories in the literature, with proponents for and against the proposals [13–16]. The code within the codons remains deeply enigmatic.

The coevolution of metabolism and the code plainly depends on the actual structure of metabolism. Most hypotheses either assume heterotrophic origins or avoid specifying a particular autotrophic metabolism [1–3,17–19]. Di Giulio [18] has noted that a heterotrophic origin of amino acids is not easy to reconcile with the coevolution hypothesis, as it implies the earliest ten amino acids were simply taken from the environment. If so, then it is hard to see how their remote synthesis could correspond to coevolution patterns found in the code. But autotrophic origins are also difficult to reconcile with coevolution if the earliest catalysts of amino-acid biosynthesis were peptidated tRNAs, as normally assumed [18]. If that were the case, then where did the amino acids required for peptidation come from? One possible reconciliation is that simple dinucleotides bound the precursor of the amino acid coded and catalysed its transformation to the cognate amino acid [19]. While this ingenious notion has some theoretical (quantum mechanical) support [20] these ideas all tend to conflict with life as a guide to the origins of metabolism [21]. Neither the deeply conserved chemistry that underpins the mechanism of translation nor the universally conserved pathways of amino-

acid biosynthesis operate in this way. Instead, we favour a possible link with autotrophic protometabolism, which has gained strong experimental support in recent years.

Phylogenetics indicates that the last universal common ancestor (LUCA) was likely autotrophic and chemiosmotic, growing from $H_2$ and $CO_2$ via the acetyl CoA pathway and reverse incomplete Krebs cycle, the gateway into the universally conserved core of metabolism [22–29]. Recent experiments are confirming what has long been predicted by theoretical thermodynamics [30–33]: the chemistry of the universal core of metabolism can indeed occur spontaneously in the absence of genes and enzymes. Starting with $CO_2$ and $H_2$ (or reducing equivalents), intermediates in the acetyl CoA pathway and reverse Krebs cycle form spontaneously [34–36]. Non-enzymatic equivalents of glycolysis, the pentose phosphate pathway [37,38] and gluconeogenesis [39] have been identified as well. Multiple syntheses of amino acids from α-keto acids by direct reductive amination [40,41], and by transamination reactions [42] can also take place. Long-chain fatty acids can be formed by hydrothermal Fischer-Tropsch-type synthesis which chemically resembles the process of fatty acid elongation [43–45]. Recent work suggests that nucleobases might also be formed following the universally conserved biosynthetic pathways, using metal ions as catalysts [46].

The endergonic barrier to the first steps of $CO_2$ reduction can be lowered by proton gradients across Fe(Ni)S barriers [47], driving $CO_2$ fixation through vectorial electrochemistry in a similar way to cells [23,29,48–50]. Mathematical modelling shows that, in principle, dynamic proton gradients could drive the autotrophic growth of protocells in a 'monomer world' (i.e. lacking RNA and peptides; [24]) in which $CO_2$ fixation is catalysed by organically chelated FeS clusters associated with the membrane [51,52]. Experimental work confirms that protocells with mixed-amphiphile membranes are stable under the necessary conditions [53,54], and that cysteine at pH 9 acts as a ligand to form redox-active 4Fe4S clusters equivalent to those in ferredoxin [55]. Together, these findings suggest that the coupling of proton gradients to autotrophic protometabolism could drive protocell growth from $H_2$ and $CO_2$ in far-from-equilibrium environments such as alkaline hydrothermal vents

[23,25,28,29,50]. The key point is that amino-acid biosynthesis would take place via a biomimetic protometabolism in autotrophically growing protocells.

Here we rationalise the possible emergence of the genetic code in the context of autotrophically growing protocells. Protometabolic reactions mimicking the core of extant metabolism and nucleotide polymerization are assumed to take place and can generate short sequences of RNA, which would initially have only possessed ribozymatic activity and lacked genetic information. Beginning with the universal core of autotrophy, we show a close relationship between the first position of the codon and the distance from $CO_2$ fixation; between purines at the second position and the hydrophobicity of amino acid; and between the properties of the codon and the length of the amino acids encoded. These simple rules allow us to assign the cognate amino acid to the large majority of codons, corroborating the idea that the code coevolved with autotrophic metabolism, with direct biophysical relationships between amino acids and random RNA strings.

**The first codon position corresponds to distance from $CO_2$ fixation**

The standard genetic code specifies 20 different amino acids which are universal across all domains of life. Yet abiotic syntheses of amino acids both in laboratory and environmental settings show substantially greater structural variation than in life, raising the question of why specific amino acids were incorporated into a genetic code. A protometabolic perspective on the genetic code neatly resolves this issue. By assuming that the chemistry that constitutes the conserved core of metabolism was the predominant prebiotic chemistry, the amino acid complement available is specified by this network. Further selection may depend on the self-reactivity of chemical species such as that proposed by Hendrickson, Wood and Rathnayake (2021) discussed further in the SI.

The protometabolism hypothesis places a new emphasis on the importance of extant metabolism for understanding the emergence of the genetic code. **Figure 1** shows a metabolic map, beginning with $CO_2$ and proceeding to the 20 standard amino acids. Rather than reconstructing the pathway through phylogenetic analyses, as in earlier work [57], we

instead emphasised the chemistry underlying the reactions. Amino-acid synthesis pathways were manually curated from the MetaCyc database [58] on the basis of their conservation between archaea and bacteria. Structuring metabolism around $CO_2$ fixation in this way generates a striking pattern. Amino acids encoded by G at the first position of the codon are clearly seen to be the first amino acids produced by the flux of carbon from the acetyl-CoA pathway into the reverse incomplete Krebs cycle. Amino acids encoded by A at the first base of the codon are predominantly the second set of amino acids produced from $CO_2$ and $H_2$, followed by C-encoded amino acids; U-encoded species are largely the furthest from $CO_2$ fixation.

This map makes an important distinction from the code coevolution hypothesis in its classical forms, which as noted earlier assume biosynthetic transformations occurring on peptidated tRNA molecules. Rather than the first codon position reflecting families of amino acids with similar biosynthetic relationships, our map points to a temporal patterning of amino acid recruitment into the genetic code. This has the potential to explain why some amino acids with distinct biosynthetic origins still have the same nucleotide at the first position in the codon, e.g. histidine and the glutamate derived amino acids. Initial weak $CO_2$ flux (in the context of growing protocells [51,52]) would mean that the highest concentration of amino acids would be those encoded by G at the first position. Note that we have limited the three hexacodonic amino acids (serine, arginine and leucine) to single first codon groups that best reflect their biosynthetic and temporal groups (A, C, C at the first codon, respectively). This is justified in the SI. To exclude bias in visual interpretation we have totalled the number of reaction steps from $CO_2$ for each amino acid. **Figure 2** shows that, while there is overlap between the species, there is a clear rise in the number of reaction steps required for the synthesis of each set of amino acids.

Given that each set of amino acids is associated with a specific first-position codon, we looked to see if a corresponding temporal ordering could be achieved for the cognate nucleotides. A direct relationship between the biosynthetic emergence of amino acids and nucleotides could explain why the earliest set of amino acids emerged at the same time as

guanosine, for example, but we could not identify any temporal ordering with nucleotides. Guanosine and adenosine nucleotides are derived through the same number of steps, and just one step less than uridine. Worse, cytidine species are derived secondarily from uridine, so their recruitment of an earlier set of amino acids makes little sense.

In reality, the number of steps is a very loose guide to their likelihood. We are currently modelling the kinetic and thermodynamic probability of each step in the core pathways of metabolism, to determine the most likely patterns of flux. Glycine is the key amino acid for purine synthesis in the *de novo* pathway. The earliest synthesis pathway for glycine was probably direct assimilation from a methenyl-pterin species [59,60]. The $CO_2$ fixation pathway is rich in guanosine derived cofactors, notably methanopterins and folates, the $C_1$ carriers in methanogenesis and acetogenesis respectively, as well as flavins and molybdenum cofactors [59,61,62]. Together, these point to a positive feedback loop in which guanosine synthesis feeds back on $CO_2$ fixation and so glycine synthesis. Greater flux through any $C_1H_2X$ species could increase glycine concentration, while in turn improving the yield of purine nucleotides.

Given that the temporal pattern predicts that purine-encoded amino acids were the earliest amino acids to be recruited to the genetic code (**Figure 1**), combined with the identified positive feedback loop between guanosine cofactors, glycine, and purine synthesis, it seems possible that in the earliest stages of metabolic and code evolution there was an unequal purine to pyrimidine ratio (as others have postulated previously [5,12]). How the specific ordering of G – A – C – U came to exist remains unclear from a purely metabolic point of view. A possible explanation is that the nucleotides which form the strongest binding interactions (G and C both form three hydrogen bonds via classic Watson-Crick base paring) recruit before those with weaker interactions (A and U form two hydrogen bonds). If so, then these binding interactions could reflect early (chemical) selection for codon:anticodon interactions with strong binding interactions between small protometabolic equivalents of mRNA and tRNA.

**The second codon position reflects amino acid hydrophobicity**

Building on the assumption of a temporal phenomenon at the first position, we found a repeating pattern, in which the hydrophobicity of the amino acid encoded corresponds to the hydrophobicity of the base at the second position of the anticodon. This pattern further specifies codon assignments. We initially observed the pattern using the hydropathy scale of Kyte and Doolittle (1982). Specifically, the more hydrophilic the amino acid within the set of amino acids, the more likely it was to be associated with U. Conversely, the more hydrophobic the amino acid, the more likely it was to be associated with A. This broad association corresponds to earlier observations on stereochemical interactions within the code [1,10,19]. What is new here is that the pattern repeats across the four time periods corresponding to the bases at the first position of the codon (**Figure S1**). This means that considering the timing of amino acid emergence with its hydrophobicity serves to specify two nucleotides in the codon-anticodon pair.

The repeating pattern is not perfect. In particular, the amino acids cysteine and arginine deviate from their expected trends. This may be addressed by considering that the Kyte-Doolittle hydropathy scale is based on the folding behaviour of modern proteins. Cysteine forms internal disulphide bridges and reactive thiols are often buried in the hydrophobic core of proteins to protect from oxidation [64]. Arginine's guanidinium group, remains charged across all pH ranges ≤ 10 [65] meaning that it is undoubtedly hydrophilic, yet its capacity to form cation-π interactions suggests some degree of hydrophobic interaction. Given these uncertainties relating to protein context, we investigated the validity of our hypothesis by considering a range of other hydrophobicity scales. Specifically, we compared the scales proposed by Hopp [66], Janin [67], LogP values determined by Tayar [68], Engleman [69] and Eisenberg [70]. **Figure S1** shows that the pattern, while still present, is obfuscated by a lack of uniform units of hydrophobicity, and variation in hydrophobicity assignments for some amino-acid species.

To mitigate these variations in hydrophobicity, we took inspiration from Trinquier and Sanejouand (1998), who attempted to determine what effective property was best preserved

in the genetic code. While we do not agree with all their reasoning, their determination of a mean hydrophobicity ranking from 43 discrete scales is an inspired choice. **Figure 3** shows violin plots based on these data. Three features stand out. First, when considering the mean values (red points) for each amino acid, the pattern of hydrophobic assignment for each 2' nucleotide is consistent across all four first codon positions, with the only deviation being histidine (which we return to later). Second, the pattern of amino acid to nucleotide assignment is robust for the earliest two groups of amino acids, encoded by G and A at the first codon position (upper panels), but is less obvious for the later C and U groups (lower panels). Third, the hydrophobicity rankings exhibit substantial variation, confirming that hydrophobicity scales are highly variable, and so could easily misrepresent the actual biophysical interactions. Whether hydrophobicity itself or some related property such as partition energy [72] is reflected in the genetic code is therefore unclear; but it is nonetheless sufficient to explain some codon assignments. Regardless of whether hydrophobicity or partition energy provides the best measurement of the physicochemical properties of amino acids, we are specifically referring to direct interactions between amino acids and the second position of the anticodon, rather than to a later stage of enzyme catalysis as proposed by Caldararo and Di Giulio [71,72].

**The third codon position corresponds to size in non-redundant codons**

With our focus on the distinction between purines and pyrimidines, we noticed that the length of the amino acid can specify its allocation between non-redundant codons. Specifically, where sister codons NNR and NNY encode different amino acids (e.g. GAY = aspartate, GAR = glutamate) the identity of the nucleotide at the third position corresponds to the length of the amino acid encoded. Length in this context is the computed distance of the amino acid in its extended linear form [73] and does not consider alternative structural states. **Figure 4** shows this simple metric of amino acid length (maximal distance from carboxylic acid to the end of the R group) is consistent in six of these seven cases: the longer amino acid is assigned the NNR codon. This size dependency is direct for the codon

(large amino acid, large purine base at the third position) but is inverted for the anticodon (large amino acid, small pyrimidine base at the third position of the codon). That implies an extension of the stereochemical effect at the third position of the anticodon, which could specify the code assignment for an amino acid based on how well it fits into some unknown pocket.

Fascinatingly, this pattern appears to hold for the proteinogenic but non-standard amino acids. Pyrrolysine is encoded at the amber stop codon UAG. Tyrosine is encoded at the UAY codons. Pyrrolysine is longer than tyrosine, and in line with this pattern, should have a Y at the third position of the anti-codon. Selenocysteine is encoded by UGA, cysteine by UGY codons. Whilst they have the same structure, the selenium atom is larger than the sulfur and this may be a sufficient difference to dictate a size-based discrimination. Further to this, these amino acids obey the expected rules of stereochemistry at the second position.

Examining this size-based relationship, we looked to see if there was any set of parameters in the structure of codons that could explain a binding pocket for amino acids. We failed to identify any patterns, but instead noticed that the position of redundant and non-redundant codons in the genetic code is not random. **Figure 5** shows an inverted code wheel, in which the third position is in the centre and the third position outermost. Clearly, redundancy is not randomly distributed around the genetic code and obeys specific rules which are entirely dependent on nucleotide sequence. This pattern allowed us to identify three rules which specify the locations of redundancy in the genetic code:

1. If there is a C in the second position of the codon (G in the second position of the anticodon) all codons **must** exhibit third position redundancy.
2. If there is an A in the second position of the codon (U in the second position of the anticodon) all codons **cannot** exhibit third position redundancy.
3. In instances where rules 1 or rule 2 do not apply (i.e. U or G in the second position of the codon), if there is a strong nucleotide (C or G) in the first codon position (S in the anticodon too) then all codons **must** exhibit third position redundancy.

Near identical rules have been reported before [74,75] in the context of translational optimisation, but our interpretation is subtly different as it links to both amino acid selection and code structure.

**Predicting the full genetic code through temporal and stereochemical rules**

If all 64 possible codons were to encode a separate species, there would be 64 different amino acids; as such, degeneracy is built into the code. The rules outlined above reduce the informational density of the genetic code to 24 units (**Figure 5A** and **B**). Given that the standard code consists of 20 amino acids, three of which are hexacodonic (with two locations in the code, taking us up to a coding capacity of 23), and that there is a need to specify stop locations (getting up to 24), this reasoning constrains the 64 possible nucleotide permutations to the level that is observed throughout all life. This in turn suggests that the structure of the genetic code is in part determined by the nucleotide triplets themselves, perhaps due to conformational states they can form and how they interact with amino acids.

These codon sets can be used as template to assign specific amino acids. If we assume that the temporal pattern holds in its entirety, then following the order of G-A-C-U specified by the first position of the codon (**Figure 2),** we can assign the four groups of amino acids to the quadrants shown in **Figure 6C**, where the amino acids in each of the four groups are placed around the codon wheel. We assigned the hexacodonic amino acids (Ser, Arg and Leu) to the single codon group most in line with their biosynthetic origins. Next, we used the hydrophobic interactions at the second position of the anticodon to assign the amino acids in each quadrant, so that the most hydrophobic amino acid was assigned to the most hydrophobic nucleotide (A), and conversely, the most hydrophilic amino acid to most hydrophilic base (U) using the averaged hydrophobicity ranks from **Figure 3**.

These assignments lead to a ‘first draft’ code, which nonetheless bears considerable similarity to the standard genetic code (**Figure 6D**). The main issue here is that histidine (circled in red) is in the wrong place, as noted earlier for its hydrophobicity, which biases the

assignments in the entire C codon quadrant. Correcting for the hydrophobicity of histidine is sufficient to correctly assign all C group amino acids (**Figure 6E**). The reason that histidine specifically deviates is unknown. The question marks remaining in **Figure 6E** are all in the U quadrant. These U-encoded amino acids are difficult to assign, as there are too many gaps in this section of the code.

Where there are two amino acids of similar hydrophobicity, which differ only at the third position, our assignments are based on size, with the longer amino acid assigned to NNR codons and the shorter one to NNY (**Figure 6E**). Together, these assignments generate a code that has clear similarity to the extant genetic code. The remaining gaps distributed around the codon wheel are conveniently the codons utilised by the hexacodonic amino acids. The emergence of hexacodonic nucleotides could therefore be a simple space filling phenomenon (**Figure 6F**), in which the amino acids filling these gaps are assigned to the code as appropriate, still following the stereochemical rules. Given that the AGR codons for arginine are commonly mutable in variant genetic codes [76] and that the reassignments are typically to stop codons or glycine or serine, both of which fulfil the same hydrophobicity position in alternative quadrants, this seems reasonable. There are a number of other factors not elucidated here that we discuss briefly in the SI, such as why methionine and tryptophan become single codon restricted, and why the biological non-coded amino acids generated in metabolism did not become incorporated into the genetic code. Applying these simple rules generates a codon table that is remarkably close to the standard genetic code. In fact, barring the assignment of AGR, our predicted code looks like several of the mitochondrial codes [77].

These observations are all drawn from patterns in the modern genetic code but are here interpreted through an autotrophic origins-of-life lens. Notwithstanding the remaining questions, our consideration of the code from an autotrophic point of view clearly makes sense of the emergence of translation from protometabolism. While we acknowledge some circularity in deriving the code from patterns found within the code, the simplest and arguably the only sensible interpretation of these patterns is that (i) they are real and strong enough to

shine through billions of years of evolution; (ii) they reflect direct biophysical interactions between cognate amino acids and specific nucleotide sequences that developed over time as protometabolism itself expanded; and (iii) the precise interactions may not be between a naked anticodon and its amino acid but could be between short RNA aptamers with binding pockets and amino acids (as both codon and anticodon are involved in the patterns, and Watson-Crick base-pairing may play some role).

**Conclusion**

We have presented here a hypothesis on the origins of the genetic code, which is grounded in a reinterpretation of longstanding observations of patterns within the genetic code [1,2,19]. The rules we have identified effectively recapitulate the extant genetic code on the basis of spontaneous protometabolism from $H_2$ and $CO_2$ at the origin of life and some form of stereochemical interactions between nucleotides and amino acids. We predict that these would involve direct non-covalent binding interactions between short aptamers of RNA and amino acids. Further work to confirm each rule in this specific context is underway, using NMR and molecular dynamic simulations. But the fact that longstanding patterns in the code are amplified and clarified by the assumption of strictly autotrophic protocellular origins lends credence to our interpretation.

Some complexities arise from the rules identified. In particular how could a temporal pattern on the *codon*, and a stereochemical pattern on the *anti-codon* co-emerge? The fact that these patterns are observed on both the codon and the anticodon may point to a three-component mechanism of early translation, as observed in life (i.e. codon, anticodon, amino acid). Other important elements that are missing from this hypothesis concern the RNA species involved (proto-mRNA, proto-tRNA, proto-ribosome) which constitute the mechanistic components of the genetic code. We envision that only RNA species with the complexity of small interacting aptamers could ensure that amino acids are incorporated into the code in the temporal ordering observed or could underpin the weak stereochemical interactions that still shine through the code. While these specific interactions remain

enigmatic, we note that they require no more than short polymers of RNA that lack any information content. Our scenario therefore offers a new perspective on the origins of information in biology. Any form of direct interactions, in which random RNA sequences interact in a non-random way with amino acids means that the sequence of amino acids in a nascent peptide would be templated by the RNA sequence. RNA aptamers with no intrinsic information content formed within growing protocells would be expected to have functions relating to protocell growth, such as $CO_2$ fixation, interactions with nucleotide cofactors, or copying the RNA sequences themselves (as the RNA polymerase is a $Mg^{2+}$-dependent protein, it is feasible that a short $Mg^{2+}$-binding peptide could partially mimic its function). The assumption of a spontaneous protometabolism in growing protocells therefore makes sense of the code within the codons, and simultaneously offers a framework that enables the transition from deterministic chemistry to genetic information at the origin of life.

**Acknowledgements**

This work was supported by funding from the Biotechnology and Biological Sciences Research Council to NL (BB/V003542/1) and from the Natural Environment Research Council to AH and NL (2236041).

**References**


[1] F.J.R. Taylor, D. Coates, The code within the codons, Biosystems. 22 (1989) 177–187. https://doi.org/10.1016/0303-2647(89)90059-2.

[2] J.T.-F. Wong, A Co-Evolution Theory of the Genetic Code, Proceedings of the National Academy of Sciences. 72 (1975) 1909–1912. https://doi.org/10.1073/pnas.72.5.1909.

[3] J.T.F. Wong, S.K. Ng, W.K. Mat, T. Hu, H. Xue, Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life, Life 2016, Vol. 6, Page 12. 6 (2016) 12. https://doi.org/10.3390/LIFE6010012.

[4] C.R. Woese, The Genetic Code. The Molecular Basis for Genetic Expression. Carl R. Woese, The Quarterly Review of Biology. 43 (1968) 327–327. https://doi.org/10.1086/405846.

[5] L.E. Orgel, Evolution of the genetic apparatus, Journal of Molecular Biology. 38 (1968) 381–393. https://doi.org/10.1016/0022-2836(68)90393-8.

[6] A.D. Ellington, M. Khrapov, C.A. Shaw, The scene of a frozen accident, RNA. 6 (2000) 485–498. https://doi.org/10.1017/S1355838200000224.

[7] M. Yarus, J.G. Caporaso, R. Knight, Origins of the genetic code: The Escaped Triplet Theory, Annual Review of Biochemistry. 74 (2005) 179–198. https://doi.org/10.1146/annurev.biochem.74.082803.133119.

[8] M.A. Khaled, D.W. Mullins, M. Swindle, J.C. Lacey, Complexes of polyadenylic acid and the methyl esters of amino acids, Origins of Life. 13 (1983) 87–96. https://doi.org/10.1007/BF00928886.

[9] J. Reuben, F.E. Polk, Nucleotide-amino acid interactions and their relation to the genetic code, Journal of Molecular Evolution. 15 (1980) 103–112. https://doi.org/10.1007/BF01732664.

[10] M.K. Hobish, N.S.M.D. Wickramasinghe, C. Ponnamperuma, Direct interaction between amino acids and nucleotides as a possible physicochemical basis for the origin of the genetic code, Adv Space Res. 15 (1995) 365–382. https://doi.org/10.1016/S0273-1177(99)80108-2.

[11] A.L. Weber, J.C. Lacey, Journal of Molecular Evolution Genetic Code Correlations: Amino Acids and Their Anticodon Nucleotides, 1978.

[12] U. Baumann, J. Oro, Three stages in the evolution of the genetic code, BioSystems. 29 (1993) 133–141. https://doi.org/10.1016/0303-2647(93)90089-U.

[13] J.C. Fontecilla-Camps, The stereochemical basis of the genetic code and the (mostly) autotrophic origin of life, Life. 4 (2014) 1013–1025. https://doi.org/10.3390/LIFE4041013.

[14] Á. Kun, Á. Radványi, The evolution of the genetic code: Impasses and challenges, Biosystems. 164 (2018) 217–225. https://doi.org/10.1016/J.BIOSYSTEMS.2017.10.006.

[15] E. V. Koonin, A.S. Novozhilov, Origin and evolution of the genetic code: the universal enigma, Iubmb Life. 61 (2009) 99. https://doi.org/10.1002/IUB.146.

[16] E. V. Koonin, Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code, Life 2017, Vol. 7, Page 22. 7 (2017) 22. https://doi.org/10.3390/LIFE7020022.

[17] M. Di Giulio, The coevolution theory of the origin of the genetic code, Physics of Life Reviews. 1 (2004) 128–137. https://doi.org/10.1016/j.plrev.2004.05.001.

[18] M. di Giulio, An Autotrophic Origin for the Coded Amino Acids is Concordant with the Coevolution Theory of the Genetic Code, Journal of Molecular Evolution 2016 83:3. 83 (2016) 93–96. https://doi.org/10.1007/S00239-016-9760-X.

[19] S.D. Copley, E. Smith, H.J. Morowitz, A mechanism for the association of amino acids with their codons and the origin of the genetic code, Proceedings of the National Academy of Sciences. 102 (2005) 4442–4447. https://doi.org/10.1073/pnas.0501049102.

[20] T. Yaman, J.N. Harvey, Computational Analysis of a Prebiotic Amino Acid Synthesis with Reference to Extant Codon-Amino Acid Relationships, (2021). https://doi.org/10.3390/life11121343.

[21] S.A. Harrison, N. Lane, Life as a guide to prebiotic nucleotide synthesis, Nature Communications. 9 (2018) 5176. https://doi.org/10.1038/s41467-018-07220-y.

[22] G. Fuchs, Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life?, Annual Review of Microbiology. 65 (2011) 631–658. https://doi.org/10.1146/annurev-micro-090110-102801.

[23] N. Lane, J.F. Allen, W. Martin, How did LUCA make a living? Chemiosmosis in the origin of life, BioEssays. 32 (2010) 271–280. https://doi.org/10.1002/bies.200900131.

[24] S.D. Copley, E. Smith, H.J. Morowitz, The origin of the RNA world: Co-evolution of genes and metabolism, Bioorganic Chemistry. 35 (2007) 430–443. https://doi.org/10.1016/J.BIOORG.2007.08.001.

[25] M.J. Russell, A.J. Hall, The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front, J Geol Soc London. 154 (1997) 377–402. https://doi.org/10.1144/gsjgs.154.3.0377.

[26] G. Wächtershäuser, Evolution of the first metabolic cycles., Proceedings of the National Academy of Sciences. 87 (1990) 200–204. https://doi.org/10.1073/pnas.87.1.200.

[27] E. Smith, H.J. Morowitz, Universality in intermediary metabolism., Proc Natl Acad Sci U S A. 101 (2004) 13168–73. https://doi.org/10.1073/pnas.0404922101.

[28] M.J. Russell, W. Martin, The rocky roots of the acetyl-CoA pathway, Trends Biochem Sci. 29 (2004) 358–363. https://doi.org/10.1016/J.TIBS.2004.05.007.

[29] W. Martin, M.J. Russell, On the origin of biochemistry at an alkaline hydrothermal vent, Philosophical Transactions of the Royal Society B: Biological Sciences. 362 (2007) 1887–1926. https://doi.org/10.1098/rstb.2006.1881.

[30] J.P. Amend, T.M. McCollom, Energetics of Biomolecule Synthesis on Early Earth, in: 2010: pp. 63–94. https://doi.org/10.1021/bk-2009-1025.ch004.

[31] J.P. Amend, D.E. LaRowe, T.M. McCollom, E.L. Shock, The energetics of organic synthesis inside and outside the cell., Philos Trans R Soc Lond B Biol Sci. 368 (2013) 20120255. https://doi.org/10.1098/rstb.2012.0255.

[32] E.L. Shock, M.D. Schulte, Organic synthesis during fluid mixing in hydrothermal systems, Journal of Geophysical Research: Planets. 103 (1998) 28513–28527. https://doi.org/10.1029/98JE02142.

[33] J.P. Amend, E.L. Shock, Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and Bacteria, FEMS Microbiology Reviews. 25 (2001) 175–243. https://doi.org/10.1111/j.1574-6976.2001.tb00576.x.

[34] M. Preiner, K. Igarashi, K.B. Muchowska, M. Yu, S.J. Varma, K. Kleinermanns, M.K. Nobu, Y. Kamagata, H. Tüysüz, J. Moran, W.F. Martin, A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism, Nature Ecology & Evolution. 4 (2020) 534–542. https://doi.org/10.1038/s41559-020-1125-6.

[35] S.J. Varma, K.B. Muchowska, P. Chatelain, J. Moran, Native iron reduces CO2 to intermediates and end-products of the acetyl-CoA pathway, Nature Ecology & Evolution. (2018) 1. https://doi.org/10.1038/s41559-018-0542-2.

[36] K.B. Muchowska, S.J. Varma, E. Chevallot-Beroux, L. Lethuillier-Karl, G. Li, J. Moran, Metals promote sequences of the reverse Krebs cycle., Nat Ecol Evol. 1 (2017) 1716–1721. https://doi.org/10.1038/s41559-017-0311-7.

[37] M.A. Keller, A. V Turchyn, M. Ralser, Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean., Mol Syst Biol. 10 (2014) 725. https://doi.org/10.1002/MSB.20145228.

[38] G. Piedrafita, S.J. Varma, C. Castro, C.B. Messner, L. Szyrwiel, J.L. Griffin, M. Ralser, Cysteine and iron accelerate the formation of ribose-5-phosphate, providing insights into the evolutionary origins of the metabolic network structure, PLoS Biology. 19 (2021). https://doi.org/10.1371/JOURNAL.PBIO.3001468.

[39] C.B. Messner, P.C. Driscoll, G. Piedrafita, M.F.L. De Volder, M. Ralser, Nonenzymatic gluconeogenesis-like formation of fructose 1,6-bisphosphate in ice, Proceedings of the National Academy of Sciences. 114 (2017) 7403–7407. https://doi.org/10.1073/pnas.1702274114.

[40] C. Huber, G. Wächtershäuser, Primordial reductive amination revisited, Tetrahedron Letters. 44 (2003) 1695–1697. https://doi.org/10.1016/S0040-4039(02)02863-0.

[41] L.M. Barge, E. Flores, M.M. Baum, D.G. VanderVelde, M.J. Russell, Redox and pH gradients drive amino acid synthesis in iron oxyhydroxide mineral systems, Proceedings of the National Academy of Sciences. 116 (2019) 4828–4833. https://doi.org/10.1073/pnas.1812098116.

[42] R.J. Mayer, H. Kaur, S.A. Rauscher, J. Moran, Mechanistic Insight into Metal Ion-Catalyzed Transamination, J. Am. Chem. Soc. 143 (2021) 42. https://doi.org/10.1021/jacs.1c08535.

[43] T.M. Mccollom, G. Ritter, B.R.T. Simoneit, Lipid synthesis under hydrothermal conditions by Fischer-Tropsch-type reactions, Origins of Life and Evolution of the Biosphere. 29 (1999) 153–166. https://doi.org/10.1023/A:1006592502746.

[44] D. He, X. Wang, Y. Yang, R. He, H. Zhong, Y. Wang, B. Han, F. Jin, Hydrothermal synthesis of long-chain hydrocarbons up to C 24 with NaHCO 3 -assisted stabilizing cobalt, Proceedings of the National Academy of Sciences. 118 (2021). https://doi.org/10.1073/pnas.2115059118.

[45] M.J. Russell, Cobalt: A must-have element for life and livelihood, Proc Natl Acad Sci U S A. 119 (2022). https://doi.org/10.1073/PNAS.2121307119.

[46] J. Yi, H. Kaur, W. Kazöne, S.A. Rauscher, L.-A. Gravillier, K.B. Muchowska, J. Moran, A Nonenzymatic Analog of Pyrimidine Nucleobase Biosynthesis, Angewandte Chemie International Edition. (2022) e202117211. https://doi.org/10.1002/ANIE.202117211.

[47] R. Hudson, R. de Graaf, M. Strandoo Rodin, A. Ohno, N. Lane, S.E. McGlynn, Y.M.A. Yamada, R. Nakamura, L.M. Barge, D. Braun, V. Sojo, $CO_2$ reduction driven by a pH gradient, Proceedings of the National Academy of Sciences. 117 (2020) 22873–22879. https://doi.org/10.1073/pnas.2002659117.

[48] W. Martin, M.J. Russell, On the origin of biochemistry at an alkaline hydrothermal vent, Philosophical Transactions of the Royal Society B: Biological Sciences. 362 (2007) 1887–1926. https://doi.org/10.1098/rstb.2006.1881.

[49] N. Lane, W.F. Martin, The Origin of Membrane Bioenergetics, Cell. 151 (2012) 1406–1416. https://doi.org/10.1016/J.CELL.2012.11.050.

[50] R. Vasiliadou, N. Dimov, N. Szita, S.F. Jordan, N. Lane, Possible mechanisms of CO2 reduction by H2 via prebiotic vectorial electrochemistry, Interface Focus. 9 (2019). https://doi.org/10.1098/RSFS.2019.0073.

[51] T. West, V. Sojo, A. Pomiankowski, N. Lane, The origin of heredity in protocells, Philosophical Transactions of the Royal Society B: Biological Sciences. 372 (2017). https://doi.org/10.1098/RSTB.2016.0419.

[52] R.N. Palmeira, M. Colnaghi, S.A. Harrison, A. Pomiankowski, N. Lane, The limits of metabolic heredity in protocells, BioRxiv. (2022). https://doi.org/10.1101/2022.01.28.477904.

[53] S.F. Jordan, H. Rammu, I.N. Zheludev, A.M. Hartley, A. Maréchal, N. Lane, Promotion of protocell self-assembly from mixed amphiphiles at the origin of life, Nature Ecology & Evolution. (2019). https://doi.org/10.1038/s41559-019-1015-y.

[54] S.F. Jordan, E. Nee, N. Lane, Isoprenoids enhance the stability of fatty acid membranes at the emergence of life potentially leading to an early lipid divide, (2019). https://doi.org/10.1098/rsfs.2019.0067.

[55] S.F. Jordan, I. Ioannou, H. Rammu, A. Halpern, L.K. Bogart, M. Ahn, R. Vasiliadou, J. Christodoulou, A. Maréchal, N. Lane, Spontaneous assembly of redox-active iron-sulfur clusters at low concentrations of cysteine, Nature Communications 2021 12:1. 12 (2021) 1–14. https://doi.org/10.1038/s41467-021-26158-2.

[56] T.L. Hendrickson, W.N. Wood, U.M. Rathnayake, Did Amino Acid Side Chain Reactivity Dictate the Composition and Timing of Aminoacyl-tRNA Synthetase Evolution?, (2021). https://doi.org/10.3390/genes.

[57] M.C. Weiss, F.L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W.F. Martin, The physiology and habitat of the last universal common ancestor, Nature Microbiology. 1 (2016) 16116. https://doi.org/10.1038/nmicrobiol.2016.116.

[58] R. Caspi, R. Billington, I.M. Keseler, A. Kothari, M. Krummenacker, P.E. Midford, W.K. Ong, S. Paley, P. Subhraveti, P.D. Karp, The MetaCyc database of metabolic pathways and enzymes - a 2019 update, Nucleic Acids Res. 48 (2020) D455–D453. https://doi.org/10.1093/NAR/GKZ862.

[59] R. Braakman, E. Smith, The Emergence and Early Evolution of Biological Carbon-Fixation, PLoS Computational Biology. 8 (2012) e1002455. https://doi.org/10.1371/journal.pcbi.1002455.

[60] G. Ceren Tok, A. Teresa Sophie Freiberg, L. Reinschlüssel, I. Emahi, P.R. Gruenke, R. Braakman, E. Smith, The compositional and evolutionary logic of metabolism, Physical Biology. 10 (2012) 011001. https://doi.org/10.1088/1478-3975/10/1/011001.

[61] B.E.H. Maden, Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism, Biochemical Journal. 350 (2000) 609–629. https://doi.org/10.1042/bj3500609.

[62] H.B. White, Journal of Molecular Evolution Coenzymes as Fossils of an Earlier Metabolic State, 1976.

[63] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology. 157 (1982) 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

[64] J.J.A.G. Kamps, R.J. Hopkinson, C.J. Schofield, T.D.W. Claridge, How formaldehyde reacts with amino acids, Communications Chemistry 2019 2:1. 2 (2019) 1–14. https://doi.org/10.1038/s42004-019-0224-2.

[65] M.J. Harms, J.L. Schlessman, G.R. Sue, B. García-Moreno E., Arginine residues at internal positions in a protein are always charged, Proceedings of the National Academy of Sciences. 108 (2011) 18954–18959. https://doi.org/10.1073/pnas.1104808108.

[66] T.P. Hopp, K.R. Woods, A computer program for predicting protein antigenic determinants, Molecular Immunology. 20 (1983) 483–489. https://doi.org/10.1016/0161-5890(83)90029-9.

[67] J. Janin, Surface and inside volumes in globular proteins, Nature. 277 (1979) 491–492. https://doi.org/10.1038/277491a0.

[68] N. El Tayar, R.-S. Tsai, P.-A. Carrupt, B. Testa, Octan-I-01-Water Partition Coefficients of Zwitterionic a-Amino Acids. Determination by Centrifugal Partition Chromatography and Factorization into Steric/Hydrophobic and Polar Components, 1992.

[69] D.M. Engelman, T.A. Steitz, A. Goldman, Identifiying nonpolar transbilayer helicines in amino acid seuqneces of membrane proteins, Annual Review of Biophysics and Biophysical Chemistry. 15 (1986) 321–353. https://doi.org/10.1146/annurev.bb.15.060186.001541.

[70] D. Eisenberg, E. Schwarz, M. Komaromy, R. Wall, Analysis of membrane and surface protein sequences with the hydrophobic moment plot, Journal of Molecular Biology. 179 (1984) 125–142. https://doi.org/10.1016/0022-2836(84)90309-7.

[71] G. Trinquier, Y.H. Sanejouand, Which effective property of amino acids is best preserved by the genetic code?, Protein Engineering Design and Selection. 11 (1998) 153–169. https://doi.org/10.1093/protein/11.3.153.

[72] F. Caldararo, M. di Giulio, The genetic code is very close to a global optimum in a model of its origin taking into account both the partition energy of amino acids and their biosynthetic relationships, Biosystems. 214 (2022) 104613. https://doi.org/10.1016/j.biosystems.2022.104613.

[73] C.B. Ching, K. Hidajat, M.S. Uddin, Evaluation of Equilibrium and Kinetic Parameters of Smaller Molecular Size Amino Acids on KX Zeolite Crystals via Liquid Chromatographic Techniques, Separation Science and Technology. 24 (1989) 581–597. https://doi.org/10.1080/01496398908049793.

[74] U. Lagerkvist, "Two out of three": an alternative method for codon reading., Proceedings of the National Academy of Sciences. 75 (1978) 1759–1762. https://doi.org/10.1073/pnas.75.4.1759.

[75] U. Lagerkvist, Unorthodox codon reading and the evolution of the genetic code, Cell. 23 (1981) 305–306. https://doi.org/10.1016/0092-8674(81)90124-0.

[76] R.D. Knight, S.J. Freeland, L.F. Landweber, Rewiring the keyboard: evolvability of the genetic code, Nature Reviews Genetics 2001 2.1 2 (2001) 49–58. https://doi.org/10.1038/35047500.

[77] A. Elzanowski, J. Ostell, The Genetic Codes, National Centre for Biotechnology Information. (2019). https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes (accessed March 2, 2022).

[78] C. Bonfio, L. Valer, S. Scintilla, S. Shah, D.J. Evans, L. Jin, J.W. Szostak, D.D. Sasselov, J.D. Sutherland, S.S. Mansy, UV-light-driven prebiotic synthesis of iron–sulfur clusters, Nature Chemistry 2017 9:12. 9 (2017) 1229–1234. https://doi.org/10.1038/nchem.2817.

[79] A. Di Martino, C. Galli, P. Gargano, L. Mandolini, Ring-closure reactions. Part 23. Kinetics of formation of three- to seven-membered-ring N-tosylazacycloalkanes. The role of ring strain in small- and common-sized-ring formation, Journal of the Chemical Society, Perkin Transactions 2. 106 (1985) 1345. https://doi.org/10.1039/p29850001345.

## Figures

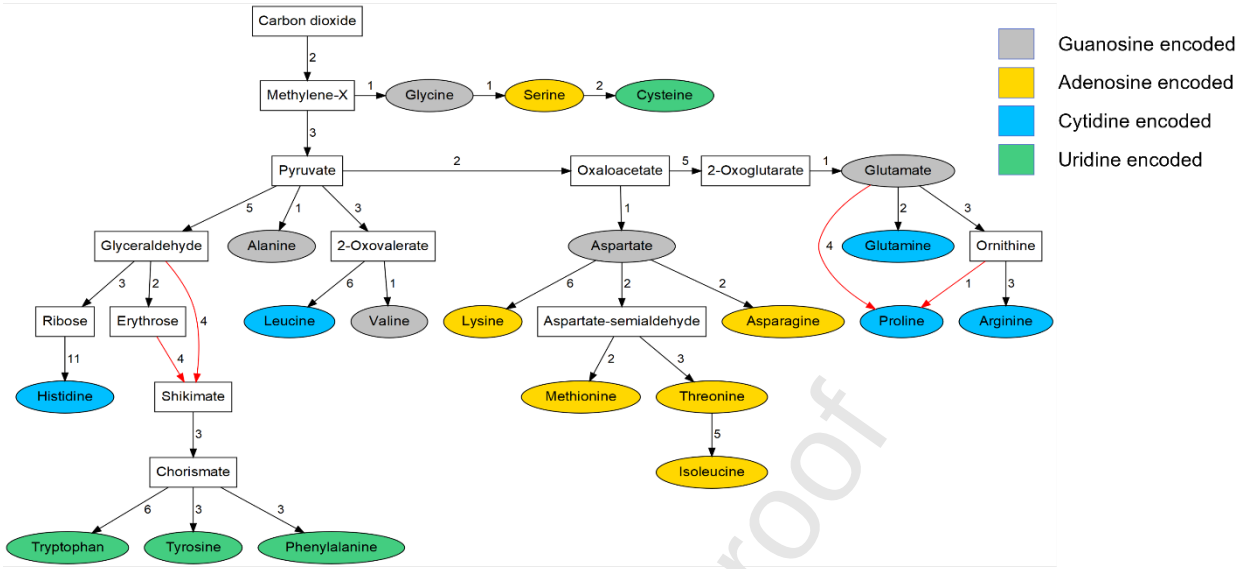


**Figure 1 Metabolic map showing reconstructed ancestral amino acid synthesis pathways**

Metabolism oriented around autotrophic $CO_2$ fixation via the acetyl-CoA pathway. Rectangular nodes represent branch point metabolites; oval-coloured nodes indicate amino acids. Edges (arrows) indicate reaction pathways between nodes. Black arrows indicate chemical mechanisms conserved between archaea and bacteria; red arrows indicate alternative pathways employed in each domain. Connections are labelled with the number of reaction steps between species; spontaneous reactions have been omitted. The amino acids are coloured according to the nucleotide at the first codon position (G in gray, A in yellow, C in blue and U in green). Pathway information was manually mined from the MetaCyc database [58]. Amino acids with multiple first codons are coloured in one of their respective nucleotides as discussed in the main text.

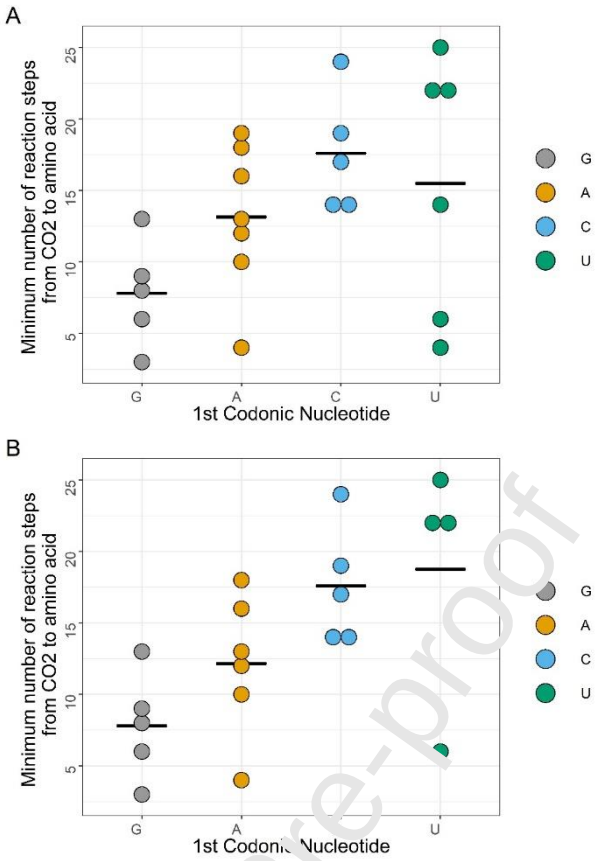


**Figure 2: Minimum number of reactions required for amino-acid synthesis from $CO_2$**

Amino acids are grouped by the nucleotide at the first position of the codon. The mean number of reaction steps is represented as a horizontal black bar. Panel A includes the dual assignment of the three hexacodonic nucleotides. Panel B shows these amino acids assigned to the single nucleotide group most consistent with their biosynthetic family and temporal ordering. The outlier amino acid with U at the first position of the codon is cysteine.

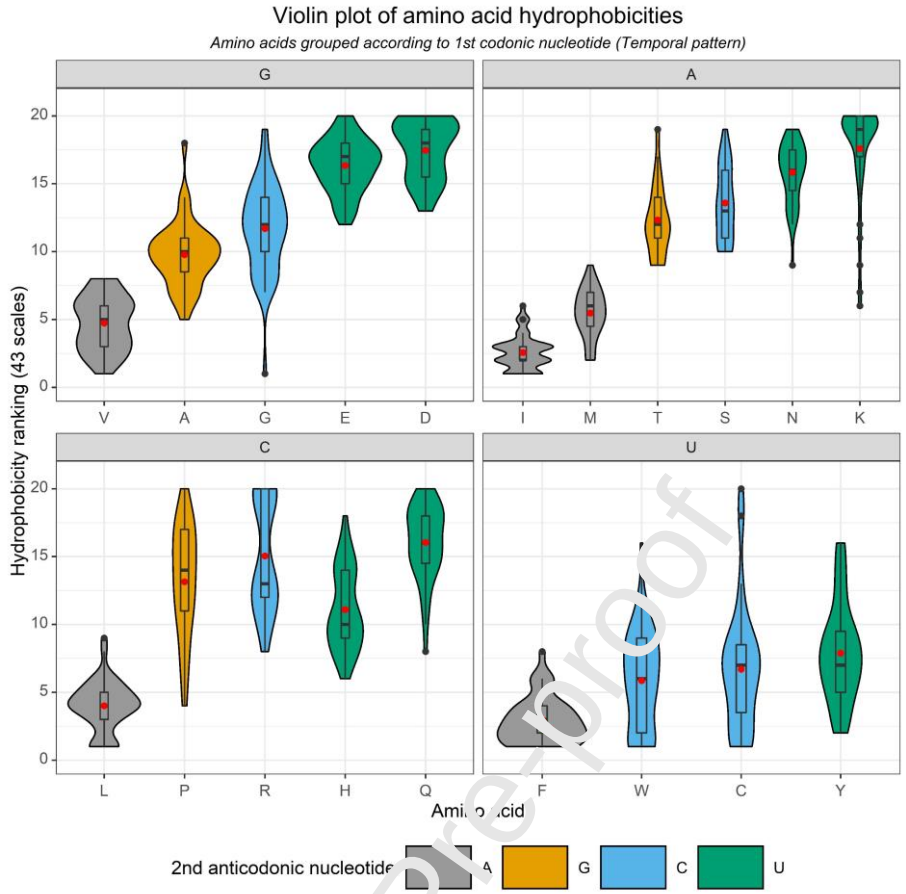


**Figure 3: Violin plots of amino acid mean and range hydrophobicity**

Amino acids are grouped according to the nucleotide at the first codon position, which we propose corresponds to a temporal grouping during the emergence of life. The violin plots are colour-coded according to the nucleotide at the second position of the anticodon, which is presumed to relate to stereochemical interactions. The hydrophobicity of the amino acid corresponds approximately to the hydrophobicity of the base at the second position. The red points show the mean ranking for each amino acid based on the method of Trinquier and Sanejouand (1998) which ranks amino acids based on 43 separate scales of hydrophobicity. The boxplots show the median values (bold black bar) and the upper and lower quartiles, with the whiskers showing the upper or lower quartiles ± 1.5*interquartile range.

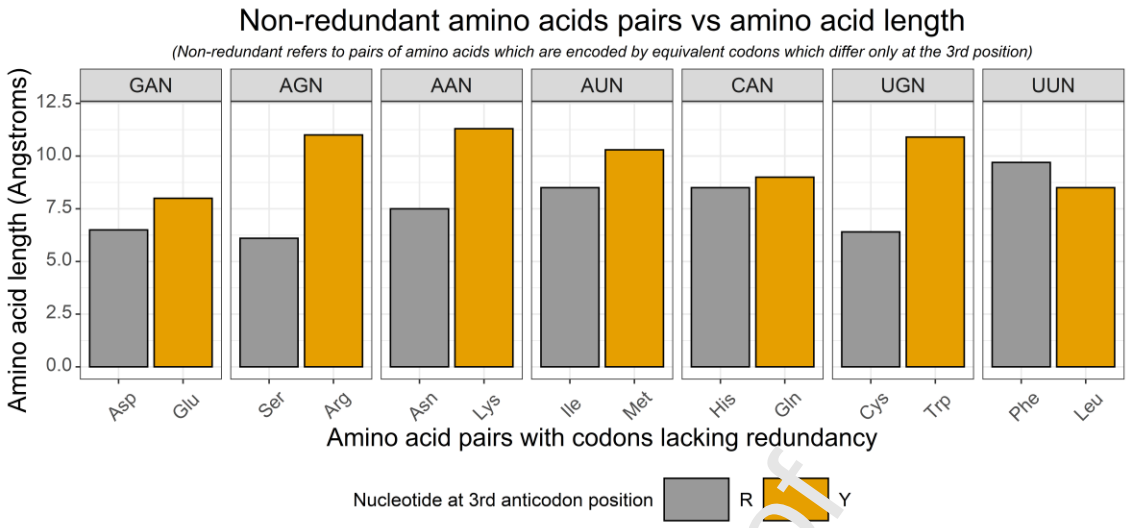


**Figure 4: Lengths of amino-acid pairs encoded by codons non-redundant at the third position.**

Non-redundant pairs of amino acids have been grouped according to their codons and colour coded according to the nucleotide species present at their 3$^{rd}$ codon position. These pairs seem to show an overall trend that the smaller amino acid is associated with a smaller nucleotide at the third position of the codon, with pyrimidines (Y) being smaller than purines (R). Conversely, in the anticodon, the smaller amino acid is associated with the larger base, suggesting spatial constraints

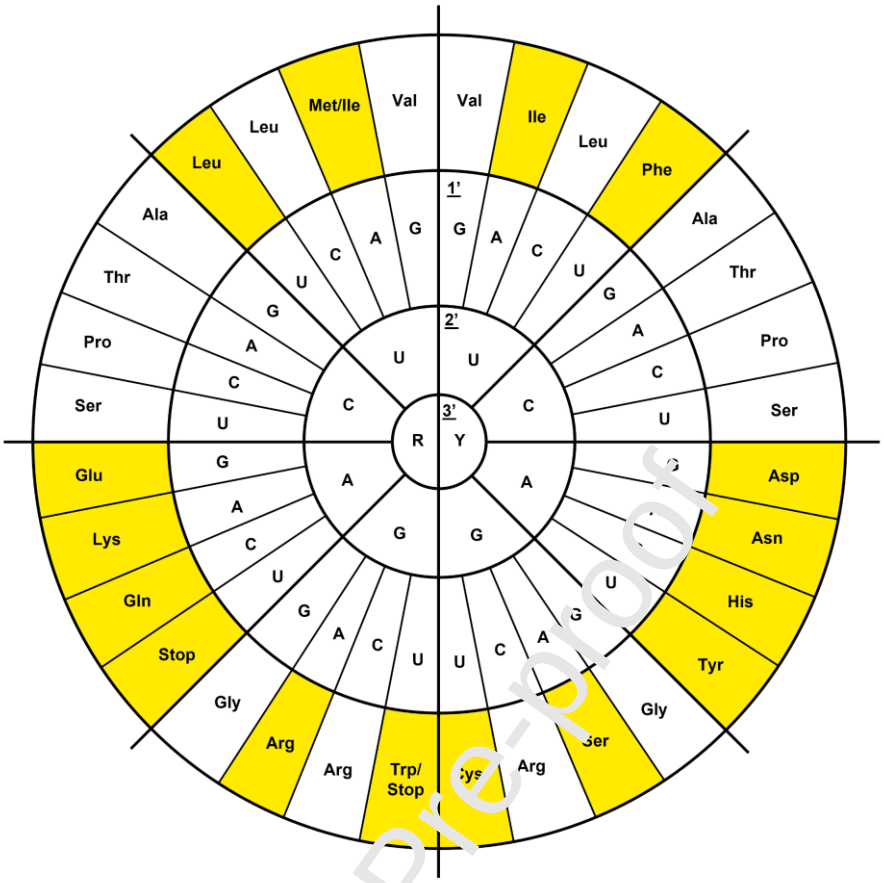


**Figure 5: Reversed codon table with redundant codons coloured**

The genetic code in wheel format, with the normal ring order for the codon sequence reversed. The innermost ring corresponds to the 3' end of the codon, while the outermost ring corresponds to the 5' end of the codon. The amino acids encoded by codon segments are labelled in the outermost ring. Non-redundant amino acids are coloured in yellow. The codon table, inverted in this way shows symmetry in the vertical axis; this is a product of the way the table is constructed. The pseudosymmetry which can be observed in the horizontal axis however indicates that redundancy in the genetic code is a non-random phenomenon that can be predicted by nucleotide sequence.

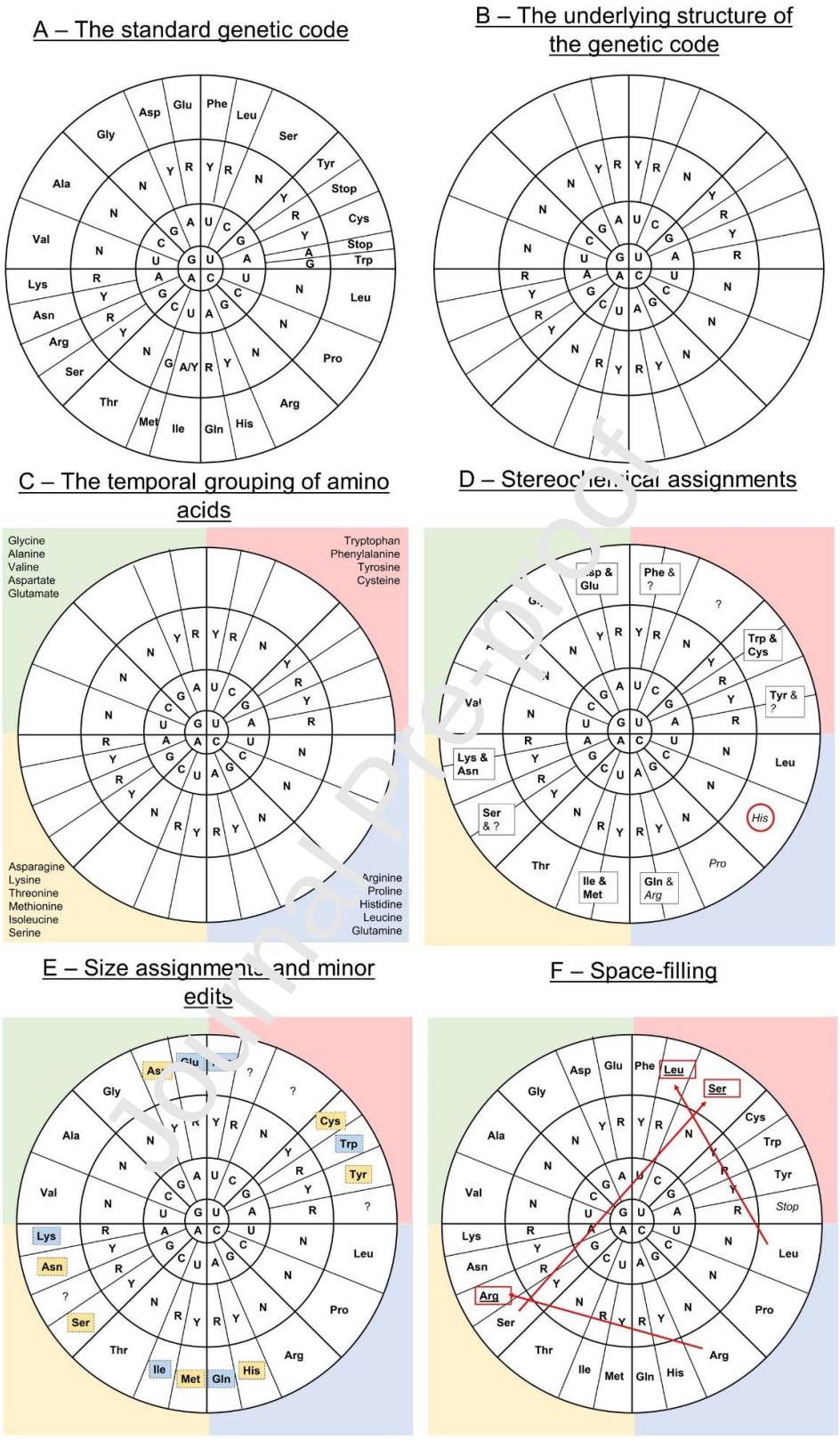


**Figure 6: Using identified rules to assign the genetic code**

(A) The standard genetic code, included for reference purposes. (B) The underlying structure of the genetic code obtained by following the rules of redundancy. (C) Temporal grouping of amino acids determined by mapping early metabolism from $CO_2$ fixation. (D) Stereochemical

assignment for positions in each group (Most hydrophobic amino acid to most hydrophobic base and vice versa) using the averaged values in Figure 3. Correct assignments in bold, incorrect assignments in italics. Histidine has been ringed in red as a major outlier which alters the assignment in the third temporal group. (E) Adding a further layer of assignment using the 3$^{rd}$ position size rules – longer amino acids (blue boxes) become associated with codons with a 3$^{rd}$ position R, smaller amino acids (yellow boxes) become associated with 3$^{rd}$ position Y, in instances where there is no competing amino acid the amino acid is also assigned 3rd position Y. In addition, we have corrected histidine's position. F) Final space filling phenomenon to obtain a complete version of the genetic code.

## Supplementary Materials

### Specifying codonic amino acids

A protometabolic hypothesis on the origins of life assumes that prebiotic chemistry has a direct continuity with the universal core of extant metabolism. This assumption indicates that the amino acids that are produced in metabolism would also have been available in a prebiotic system. Further, if this protometabolism was the most reliable, or the simplest, or the most efficient chemistry available then the standard amino acids would have been the most readily available for incorporation into a genetic code.

Deselection of biological non-proteinogenic amino acids have been discussed by Hendrickson, Wood and Rathnayake, 2021. We agree with a substantial portion of their work, particularly their arguments for the deselection of ornithine, homoserine and homocysteine. Deselection of citrulline may also be similar. We do not have an explanation for why some α-keto acids that are found in metabolism such as 2-oxobutanoate which could form homoalanine were never transaminated nor incorporated.

### Single codon restrictions for hexacodonic amino acids

The temporal pattern we observe, and indeed the original code coevolution hypotheses are complicated by the hexacodonic amino acids – serine, arginine and leucine. These amino acids have 6 codons, a set of 4 cognate codons, and a second location of 2 cognate codons. There is always an alternate 1$^{st}$ codon nucleotide. This is somewhat at odds with a temporal pattern as biosynthetic pathways mostly remain fixed. In order to deconvolute this pattern, we restrict each of the 3 amino acids to a single 1$^{st}$ nucleotide group which best reflects their biosynthetic origins. Arginine is derived from glutamate; all other glutamate derived amino acids are in the C group, so the group that makes the most sense is C. Serine is secondarily derived from glycine, the amino acid closest to carbon dioxide; we therefore assigned it to its earliest group, A. Leucine is in both group C and U. There are no distinctive details that we can used to assess which group is most likely; as such we restrict it to the earlier group, C.

**Metabolic Pathway curation for protometabolic map**

Metabolic pathways were manually curated from the MetaCyc metabolomics database [58] with appraisal from supporting literature detailed on the site for some specific pathways. Each metabolite of interest was appraised individually, and synthetic pathways were appraised on specific criteria. First, only pathways that were reported to operate in the autotrophic direction were selected, eliminating degradation pathways; some pathways were reported in mixed direction and were considered as potentially autotrophic. In addition, these pathways were required to connect to an autotrophic system starting with $CO_2$ fixation and the Krebs cycle.

Second, the pathways must have been reported to be present in both bacterial and archaeal domains of life – pathways restricted to eukaryotes were discounted. Further to this, pathways were appraised on the degree in which they map between domains. Pathways which are found in a few clades of bacteria and/or archaea were discounted if pathways with greater conservation were present.

Third, pathways were appraised on their chemistry. In some instances, metabolites lack a conserved synthesis pathway between bacteria and archaea. In these instances, the chemical nature of the pathways was assessed for their degree of similarity. If the sequence of chemical reactions was conserved (even if the catalysts themselves were not) those pathways were considered to be equivalent. Examples include the acetogenesis and methanogenesis pathways of $CO_2$ fixation, which differ in the use of reductant and C1 carriers. In addition, if there are multiple pathways with functionally identical chemical species and similar interconversions, these pathways were considered identical i.e., the succinylase vs acetylase variants of lysine synthesis.

Fourth, the effective simplicity of the pathway was also taken into consideration. Pathways which utilise inorganic species were selected over pathways that degrade existing biomolecules. This was most of relevance for thiol incorporation into cysteine and methionine, where sulfide assimilation pathways were prioritised over thiol transfers between such species.

Finally, in the few instances where there was no rationalising the ancestral pathway between bacteria and archaea these pathways were considered to be separate. The shikimic acid synthesis pathway and the formation of proline are the primary examples here. A summary of selected pathways can be found in **Table S1**.

This table was used to determine minimum number of steps from $CO_2$ to each amino acid and nucleotide of relevance. The number of steps is the number of enzymatic steps from each species assuming relevant other species are present i.e., in the formation of histidine, once AICAR is formed, it is assumed ATP is available to react with it. In instances where multiple products converge the longest more representative pathway was selected i.e., the synthesis of tryptophan was calculated via the indole group rather than the considerably shorter route via serine.

| **Amino acid** | **Ancient metabolic pathway** | **Annotation** |
|---|---|---|
| Glycine | Direct glycine synthesis pathway | Identified by phylogenetics as the most ancient pathway. [59] |
| Alanine | Transamination of pyruvate | Near universal |
| Valine | Aliphatic synthesis pathway | Near universal |
| Aspartate | Transamination of oxaloacetate | Near universal. |
| Glutamate | Transamination of 2-oxoglutarate | Near universal. |

| | | |
|---|---|---|
| Serine | Serine hydroxy methyltransferase | Glycine synthesis is reliant on the same cofactors and this is one additional step, so synthesis from glycine makes greater sense than from sugar synthesis. [59] |
| Lysine | Diaminopimelate pathway | 4 of the 6 pathways in the MetaCyc database are near identical in chemistry differing only in the protecting groups utilised. The other two, (variants of the homocitrate pathway) requires dedicated carrier proteins, and whilst found in some archaea does not appear widely distributed. |
| Asparagine | Activation and amination | All pathways of asparagine synthesis are chemically identical. They differ only in what does the activation of the carboxylic acid. |
| Threonine | Homoserine pathway | Near universal |
| Methionine | Direct homocysteine synthesis route | All synthetic routes are similar starting from aspartyl semialdehyde (AspSA) The methanogenic archaeal route which directly thiolates and reduces AspSA was selected due to simplicity, and it maps between archaea and bacteria. The homoserine transsulfuraltion route is a viable alternative that extends methionine synthesis by 2 reactions. |
| Isoleucine | Threonine pathway | Isoleucine synthesis was the most difficult to evaluate with the threonine pathway the |

| | | |
|---|---|---|
| | | most common in bacteria and the citramalate pathway in archaea. Both connect well to the previously constructed sections of the network. They both converge on 2-oxobutanoate so are related species. The threonine route was selected due to the existence of spontaneous reactions in the pathway. Citramalate is a viable synthetic route due to it similarity to the oxidative Krebs however. |
| Glutamine | Activation and animation | All pathways chemically identical. |
| Histidine | Histidine synthesis pathway | Invariant autotrophic pathway. |
| Proline | Alternative pathways | Poor overlay between bacteria and archaea. Synthetic routes from ornithine and from glutamate semialdehyde. Net difference in chemistry is negligible though both are net reduction and cyclisation reactions. |
| Leucine | Leucine synthesis pathway | Invariant. Extension of the aliphatic amino acid synthetic route. Chemistry is like the oxidative Krebs. |
| Cysteine | Via activated serine intermediate | Chemically invariant between multiple pathways. Hydroxyl activation and then sulfuration. Sulfide donor varies between species. We prefer the direct assimilation of |

| | | |
|---|---|---|
| | | hydrogen sulfide due to the autotrophic nature of the reaction. |
| Phenylalanine | Synthesis from Prephenate | Note for all aromatic amino acids: The initial portion of aromatic group synthesis – the formation of shikimic acid – differs between bacteria and archaea. Because they converge on the same point before the formation of precursors common to all aromatic amino acids however, we feel confident with these assignments. |
| Tyrosine | Synthesis from prephenate | Largely conserved. Synthesis could occur from the *p*-oxidation of phenylalanine. |
| Tryptophan | Indole pathway | Near universal |

***Table S1: Pathway assignments for a map of ancient metabolism connecting $CO_2$ to amino acid synthesis***

*Table detailing the pathways assigned for amino acid syntheses in an ancient metabolism. Amino acids have been annotated with named pathway or chemical description where appropriate. They have been annotated with limited detail on why some specific pathways have been selected.*

**Hydrophobicity assignments for temporal amino acid groups with single hydrophobicity scales**

We wanted to determine whether amino acids in each temporal group could be assigned into their cognate codons on the basis of hydrophobicity. Multiple hydrophobicity scales exist. We chose 6 scales (**Figure S1**). These scales show that hydrophobicity can be used to assign amino acids to relevant second anticodon nucleotides with moderate accuracy. Later

temporal groups show poorer assignment, arginine is a consistent exception. Additionally, these hydrophobicity scales show poor consistency in assignment variation between scales.

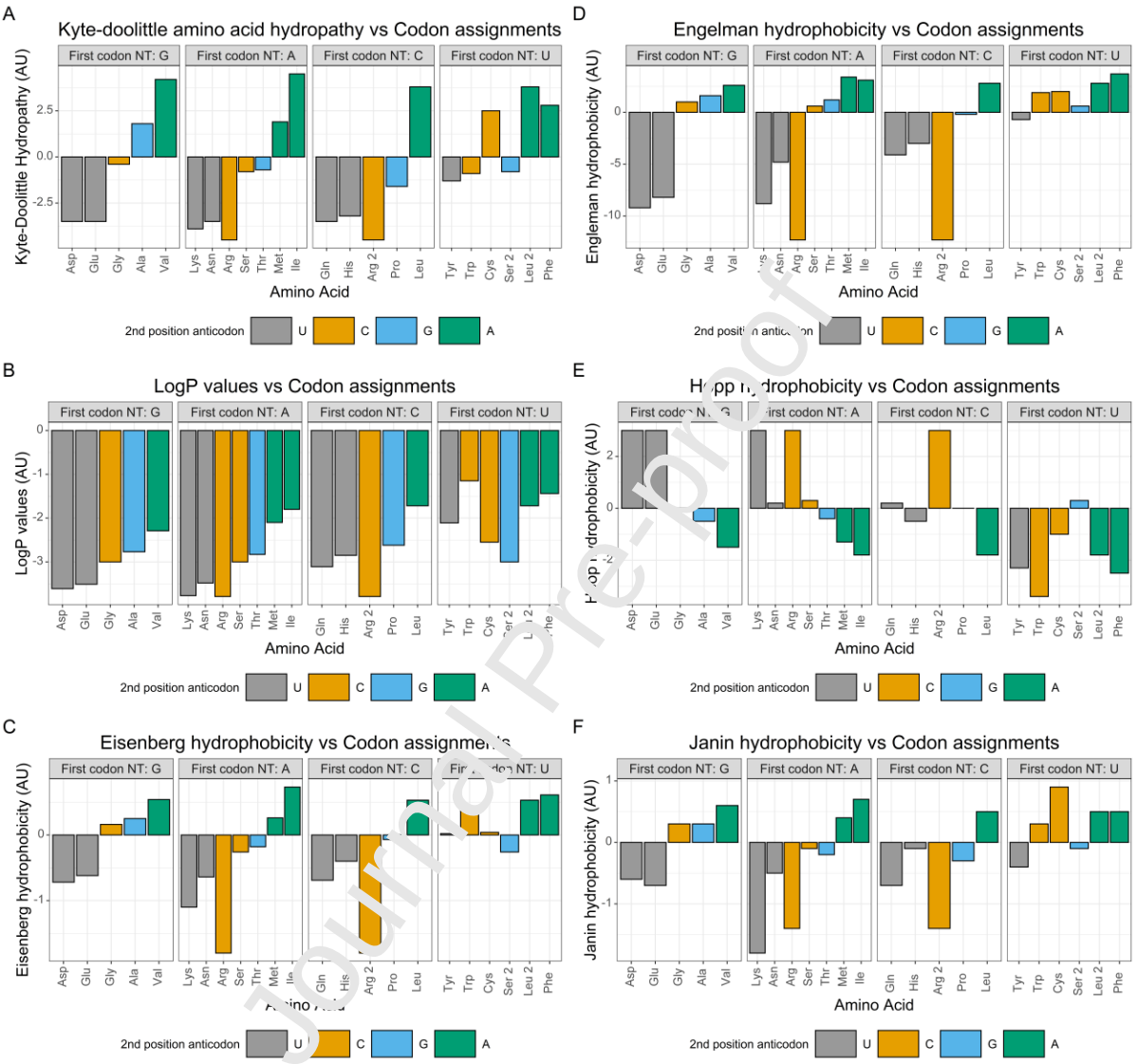


**Figure S1: Hydrophobicity scales compared against 2nd anticodons hydrophobicity and grouped by temporal emergence**

*The relationship between the 1st codon position (groups of amino acids) and the 2nd anticodon position (bar colours) and a set of hydrophobicity scales. Panel A: The Kyte Doolittle hydropathy, Panel B: Octanol-water LogP values for amino acid zwitterions determined by Tayar et al, 1992. Panel C: Eisenberg hydrophobicity scale, Panel D: Engelman hydrophobicity, Panel E: Hopp hydrophobicity scale and Panel F: Janin hydrophobicity scale.*

**Cysteine as a key outlier in the temporal pattern**

Cysteine is an outlier of major relevance to the origins field. By our temporal scheme, U at the first position of the codon implies a relatively late assignment to the genetic code. Yet cysteine is essential for iron-sulfur clusters [55,78], a catalytic species with ancient function in $CO_2$ fixation. As discussed in Hendrickson, Wood and Rathnayake (2021) cysteine can also self-react during adenylation (see **Figure S2**). Like homocysteine, the lower pKa of the thiol group makes it a stronger nucleophile at neutral pH increasing the likelihood of cyclisation but unlike homocysteine forms an unfavoured 4-membered ring [79]. We suggest that the competing effects of cysteine's catalytic utility and its own reactivity delay but do not prevent its incorporation into the code. This means that cysteine could have played an important role in prebiotic iron-sulfur chemistry before it was incorporated into the genetic code.

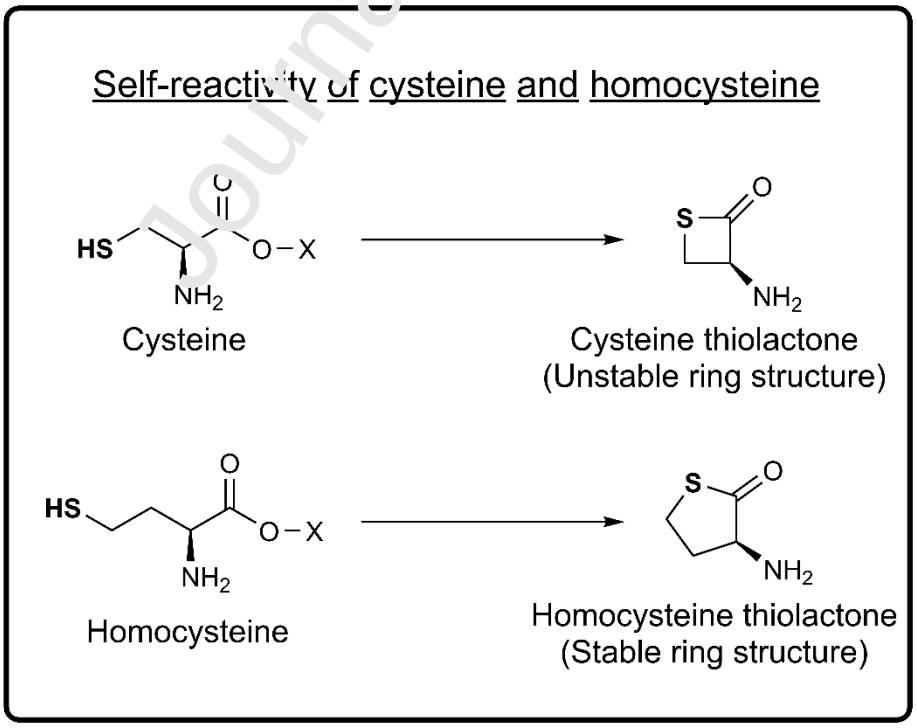

**Figure S2.**

*Structural diagram illustrating how amino acid thiol side chains can form cyclic structures. The 4-member cysteine thio-lactone is more unstable than the homocysteine equivalent due to greater bond strain. Nucleophilic species in bold. X indicates chemical activation of the carboxylic acid required for lactone formation in aqueous solution, this was likely adenylation in the context of the development of the genetic code.*

**Single codon restriction**

Using the rules identified in above specifies a genetic code that is highly similar to the extant genetic code. The major outlier are the single codon restrictions for tryptophan and methionine. Given these amino acids are non-restricted in many mitochondrial variant codes, the single codon restriction may be due to genomic regulation. Alternatively, given these two amino acids enact a strong metabolic expense; Tryptophan is the furthest from $CO_2$ and is also the largest amino acid in terms of fixed carbon, and methionine is the initiation amino acid (for reasons unknown) which comes with its own metabolic implications.

**Declaration of interests**

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

**A biophysical basis for the emergence of the genetic code in protocells**

Stuart A. Harrison, Raquel Nunes Palmeira, Aaron Halpern and Nick Lane

**Highlights:**

We provide a new framework for the origin of the genetic code in protocells growing by $CO_2$ fixation

Using a simple set of rules we are able allocate the large majority of codon assignments

These codon assignments are based on biophysical interactions in an expanding protometabolism

Biophysical interactions between RNA strings and templated amino acids can drive protocell growth

Autotrophic protocell growth gives a new context for the origin of information in biology