INTERFACE

royalsocietypublishing.org/journal/rsif

Review

Cite this article: Fried SD, Fujishima K, Makarov M, Cherepashuk I, Hlouchova K. 2022 Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc. Interface* **19**: 20210641. https://doi.org/10.1098/rsif.2021.0641

Received: 8 August 2021
Accepted: 5 January 2022

Subject Category:
Reviews

Subject Areas:
astrobiology, biochemistry, biophysics

Keywords:
origins of life, protein evolution, prebiotic polymers, early peptides

Authors for correspondence:
Stephen D. Fried
e-mail: sdfried@jhu.edu
Klara Hlouchova
e-mail: klara.hlouchova@natur.cuni.cz

Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids

Stephen D. Fried[1,2], Kosuke Fujishima[3,4], Mikhail Makarov[5], Ivan Cherepashuk[5] and Klara Hlouchova[5,6]

[1]Department of Chemistry, and [2]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21212, USA
[3]Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo 1528550, Japan
[4]Graduate School of Media and Governance, Keio University, Fujisawa 2520882, Japan
[5]Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague 12800, Czech Republic
[6]Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague 16610, Czech Republic

SDF, 0000-0003-2494-2193; KF, 0000-0002-8844-812X; IC, 0000-0002-5681-7242; KH, 0000-0002-5651-4874

Recent developments in Origins of Life research have focused on substantiating the narrative of an abiotic emergence of nucleic acids from organic molecules of low molecular weight, a paradigm that typically sidelines the roles of peptides. Nevertheless, the simple synthesis of amino acids, the facile nature of their activation and condensation, their ability to recognize metals and cofactors and their remarkable capacity to self-assemble make peptides (and their analogues) favourable candidates for one of the earliest functional polymers. In this mini-review, we explore the ramifications of this hypothesis. Diverse lines of research in molecular biology, bioinformatics, geochemistry, biophysics and astrobiology provide clues about the progression and early evolution of proteins, and lend credence to the idea that early peptides served many central prebiotic roles before they were encodable by a polynucleotide template, in a putative 'peptide-polynucleotide stage'. For example, early peptides and mini-proteins could have served as catalysts, compartments and structural hubs. In sum, we shed light on the role of early peptides and small proteins *before* and *during* the nucleotide world, in which nascent life fully grasped the potential of primordial proteins, and which has left an imprint on the idiosyncratic properties of extant proteins.

1. Introduction

Proteins are the macromolecules responsible for performing the vast majority of biological functions in extant life, and yet their importance during the Origin of Life is often underappreciated or even neglected. The RNA world hypothesis has recently become recentred at discussions of life's origins, an epistemic shift which can be attributed to remarkable developments in the abiotic chemical syntheses of the four canonical nucleosides [1,2] as well as perhaps equally impressive demonstrations of RNA catalysis discovered by directed evolution—in particular, in mediating RNA replication [3–6]. Recent work suggesting potentially prebiotic pathways to deoxynucleosides [7,8] have moved some to refer to this early period as a 'nucleotide' world. One of the logical extensions of this corpus of work is that a catalytically active hereditary molecular system could have emerged directly from a system of abiotic molecules with low molecular weight. In this train of thought, proteins' (or peptides') role in early life's emergence is typically not discussed and is construed as occupying

THE ROYAL SOCIETY PUBLISHING

Table 1. Prebiotically relevant properties of polypeptides versus polynucleotides.

consideration	section	polypeptides	polynucleotides
abiotic synthesis of building blocks	2	amino acids—trivial and documented at high concentration without human intervention [9,10,18,19]	nucleosides—possible though non-trivial, and requires changes in reaction conditions, and possibly not compatible with Hadean environment [1,2,7,18]
modularity	3	yes—proteins with smaller/alternative alphabets can fold and be functional [20–28]	partial—each base type requires a base-pairing partner [29–31]
abiotic condensation	2	possible through wet–dry cycling, activation with small molecules (e.g. COS), salt-based deliquescence or catalytic peptide ligation [13,32–34]	possible through wet–dry cycling. Though requires phosphorylated monomers and many branching reactions possible given the various nucleophilic moieties on nucleotides [18,35–37]
functional (catalytic) capacity		diverse. Could have supported early metabolism [38,39]	limited primarily to phosphoryl group transfer chemistry (with the important exception of the ribosome, which catalyses aminolysis of esters) [38]
cofactor utilization	4	diverse [40–42]	limited primarily to Mg^{2+} [43–46]
tolerance to backbone impurity		substitution of amides for esters associated with incremental decreases in stability [13,47,48]	base-pairing possible with diverse backbones (peptides, other sugars [49]), though typically tertiary structures not compatible [50]
pH tolerance		high—stable between 3 and 10	low—stable between 5 and 7—due to both backbone cleavage and depurination
tolerance to high Fe^{2+} levels (and other divalent cations)		high [43,51–53]	low—catalyses hydrolysis of phosphodiesters through 'in-line' and Fenton mechanisms [45,46]
unassisted refoldability	5	generally, yes. Complex proteins may require chaperones or translation, but simple proteins can fold unassisted [54,55]	generally, no. Rough energy landscapes mean that energy input or active processes necessary to fold to a single structure [54,55]

a later phase, potentially after the emergence of translation. Hence, proteins become 'important' once they can be encoded and synthesized in accordance with a nucleic acid template. This model has some intuitive appeal because, in extant biochemistry, proteins do not replicate themselves.

We believe that this way of thinking is over-simplistic at best, and likely incorrect. Named by the Swedish biochemist Jöns Jacob Berzelius after the ancient Greek word *πρώτειος* (meaning 'first'), proteins were indeed most likely the earliest biopolymer. The evidence for this comes from many lines of research, including: (i) the ease with which amino acid building blocks can emerge spontaneously through simple and unsupervised gas-phase chemistry [9,10]; (ii) the prevalence of some canonical amino acids (cAAs) in carbonaceous meteorites [11,12]; and (iii) the facile nature of the condensation reaction between amino acids [13] which can be mediated by wet–dry cycles in 'warm little pond' terrestrial settings, or under high-pressure high-temperature conditions in hydrothermal vents [14–16], and may even be possible in the interstellar medium as well [17]. These aspects—while perhaps less high-profile than some recent works focusing on nucleoside and nucleic acid chemistry—merit our utmost consideration as we try to build more detailed models about the abiotic-to-biotic transition (table 1). The goal of this review is to discuss aspects of early proteins and their potential roles *prior to* and *during* the nucleotide world, which we refer to as a 'peptide-like/nucleoside stage' and a 'peptide-polynucleotide stage'.

Numerous other traits of proteins that have not been widely considered make them an ideal proto-biomolecule. A list of many of these properties is given in table 1 and will be discussed in the following sections. One important property of proteins is that their alphabet is reducible as well as extendable (§3). Several lines of evidence suggest that the protein alphabet existed originally in a reduced form [56–58], such as (i) the ease with which certain cAAs (referred to as 'early' amino acids) are abiotically synthesized and the preponderance of similar amino acids on meteorites [59], (ii) the records of genetic code and metabolic pathway evolution [18] and (iii) the inferred amino acid composition of ancestral genomes [60,61]. Moreover, proteins with reduced prebiotic alphabets are still capable of folding into globular-like structures [20–24], and performing molecular recognition [62–64], suggesting that nature could take advantage of the structural and functional potential of polypeptides even with many cAAs missing. Nucleic acids are also capable of having their alphabets reduced [65]; however, the chemical logic of base-pairing places more restrictions on addition to and removal from the alphabet, as each nucleobase type is greatly diminished without a 'well-chosen' partner with which it preferentially basepairs.

Extant proteins are known to further extend their functional toolbox by using a range of cofactors (§4). While most important in extant oxidoreductases, cofactors could have played many more roles during prebiotic times before

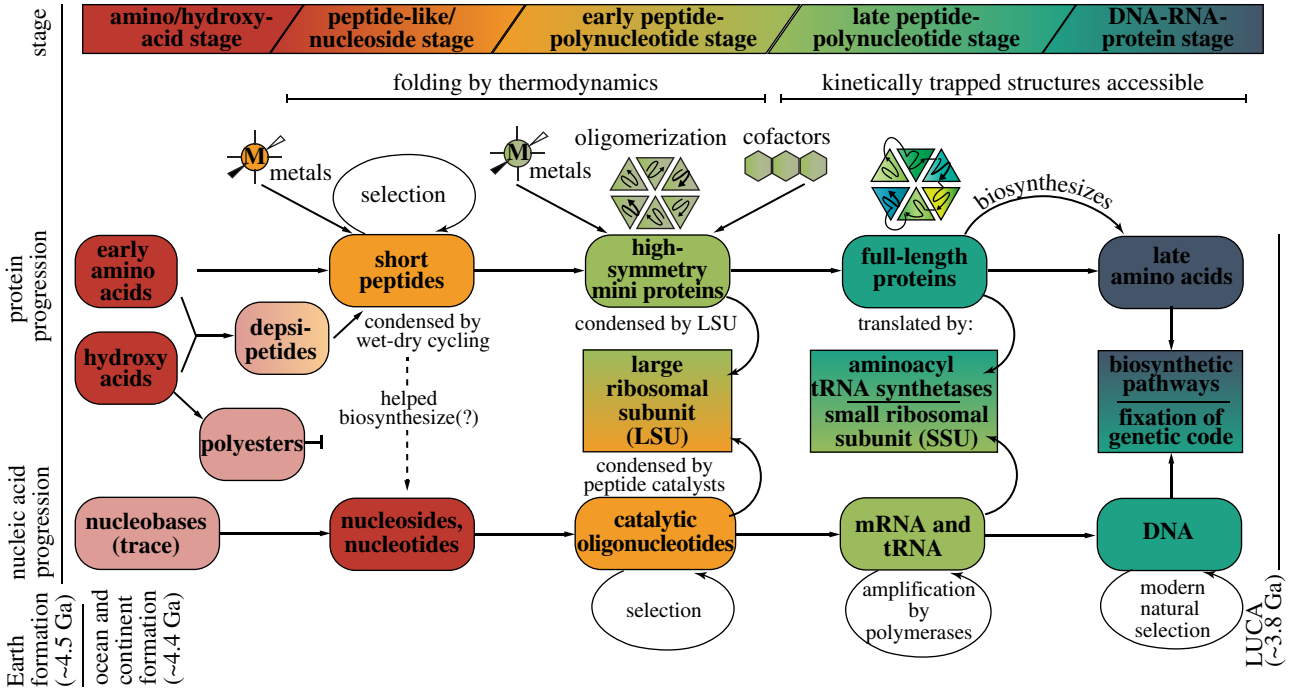

Figure 1. A model for the Origin of Life informed by the early accessibility of various monomeric organic molecules leading to the formation of short peptide-like molecules and emphasizing the various ways in which peptides and nucleic acids coevolved through collaborating at successive stages of sophistication. Section 2 describes these five stages: the amino/hydroxy acid stage, the peptide-like/nucleoside stage, the early peptide-polynucleotide stage, the late peptide-polynucleotide stage and the DNA–RNA–protein stage.

the acquisition of the more sophisticated (late) cAAs, and moreover hark back to a time when prebiotic systems chemistry was very diverse, prior to the establishment of canonical components through a central encoding dogma.

Peptides enjoy one of the simplest and most versatile condensation reactions in organic chemistry, between an amine and a carboxylic acid. The condensation reaction between two amino acids can be mediated through the removal of water without any catalysts [13,34], by activation with a range of prebiotically plausible condensing reagents [66], or potentially catalysed heterogeneously by minerals [19,67]. Thanks to this, a recent network model predicted that iteratively combining products from reactions in seven 'generations' from a starting set of six gases (N_2, CH_4, NH_3, H_2O, H_2S, HCN) would culminate in 27 different peptides [68]. By contrast, the condensation reaction of nucleotides presents several challenges. First, the selective incorporation of 5′–3′ phosphodiester linkages represents a regioselectivity challenge, given the simultaneous presence of 2′ hydroxyls [35,36]. Second, nucleobases possess numerous nucleophilic functional groups, which must compete with the 5′ and 3′ hydroxyl groups on the sugar as the donor to the phosphate group in condensation. Hence, the creation of linear polymers (as opposed to the combinatorially more facile highly branched structures) poses a statistical challenge [37]. Third, the double negative charge present on terminal monophosphates render them quite unreactive without activation or catalysis [69]. These challenges therefore invite the speculation that condensation of nucleotides was *catalysed* [35], with peptides possibly playing a role. This hypothesis—if true—argues for a 'recentring' of prebiotic systems chemistry [39] in which non-encoded peptides played an essential role at the earliest stages [70]. Moreover, it suggests that polynucleotides (and perhaps nucleosides) were themselves the product of early biocatalysis. This scenario provides an alternative to a purely 'organic chemical' emergence of polynucleotides and invites a new way of thinking about the origins of life that combines the 'best' of what peptides and nucleotides had to offer at distinct stages of emergence (figure 1).

2. Protein selection without nucleic acids

In figure 1, we lay out a model for the Origin of Life that integrates a series of emergences in the development of polypeptides and polynucleotides, and emphasizes important ways in which these molecules 'collaborated' with each other at different stages.

2.1. Amino/hydroxy acid stage

The amino/hydroxy acid stage reflects a very early period in which amino acids and alternatives to amino acids (such as hydroxy acids or dicarboxylic acids) accumulated on the Hadean Earth through Miller–Urey reactions in the atmosphere and by delivery from carbonaceous meteorites [18]. Hydroxy acids are highlighted here along with amino acids as their formation of polyesters or depsipeptides (when both ester and amide linkages form in a mixture of amino and hydroxy acids) under wet–dry cycle conditions has been studied as an appealing prebiotically plausible mechanism for peptide bond formation. Depsipeptides have been shown to be enriched with amino acids over time through a combination of ester–amide bond exchange [13]. In our view, the much higher concentration of these building blocks relative to nucleobases (by *ca* 3–4 orders of magnitude [12,71]) and their more facile synthesis from small gases imply an earliest stage where amino acids dominated the portion of the primordial soup destined to become biotic. In essence, the universe's chemical 'preference' for amino acids gave the protein progression (top row of figure 1) a 'leg up'

on the nucleic acid progression (as illustrated by the staggered colour scheme in figure 1) and influenced the state of each of these molecule types' progressions at the subsequent stages.

2.2. Peptide-like/nucleoside stage

In a peptide-like/nucleoside stage, peptides with structural and functional properties expanded by self-templated synthesis without a nucleic acid template. The last few years have witnessed a dizzying number of examples of proteins (oftentimes through short peptide motifs) self-associating in myriads of ways with diverse aggregation numbers and degrees of order. Liquid–liquid phase separated droplets mediated by disordered regions and rigid beta-amyloids represent two ends of this spectrum with respect to order, though remarkably both are often mediated through short peptide-length motifs. These findings bear great significance for the earliest prebiotic stages, when the peptides that were available were probably short and non-globular.

A large body of work has demonstrated the ability of short peptides to self-assemble into a range of morphologies, including fibres, nanotubes, ribbons and vesicles. All of these morphologies are enabled by 'open' self-complementarity which allow stable structures to form with simple polymers (*N* as low as 5 [72] or 6 [73]) and high aggregation numbers. A range of examples have demonstrated how such structures could have propagated, in a prion-like manner, through nucleation, growth and fragmentation, in a manner mimicking self-replication [74–76]. Moreover, there is some evidence that such amyloidogenic sequences can self-replicate by employing the amyloid's 'layered' structure to spatially organize monomers at an open face, thereby biasing condensation reactions to generate peptides of like sequence [77–79]. Self-assembling peptide sequences, organized into cross-beta structures (or other stable morphologies) would also be expected to be more resistant to hydrolysis than other short peptides. Therefore, under wet–dry cycling conditions, structure-forming peptides could be actively selected for at the expense of non-structure-formers, both at the level of synthesis and at the level of hydrolysis. Also, such hydrolytic conditions would have provided a natural chemical evolutionary pressure to purge hydroxy acid constituents from depsipeptides toward an eventual takeover by peptides [13] (figure 1). Nevertheless, the tolerance of hydroxy acids into peptide-like polymers likely provided a helpful stepping stone for functional polymers to emerge at this early stage [47,48].

Therefore, template-based self-assembly of such peptide pools could lead to a type of 'imperfect replication', providing a mechanism to induce amino acids to polymerize into particular sequences over the myriad of alternative possibilities. We note that because of the traditional emphasis paid to the auto-replicative capacities of polynucleotides in the Origin of Life community, more investigation of the self-propagating capabilities of peptides is necessary, though there are a few examples in the literature [66,79,80].

It has been shown that simple amyloids can act as biocatalysts in a number of reactions [66,81–83]. A consideration that has received less attention is that these peptide catalysts may have played important roles in increasing the availability of nucleobases through catalysing elementary chemical reactions, given their paucity in the amino/hydroxy acid stage. Hence, the early emergence of peptides may have supplied the power of biocatalysis to support the synthesis of more challenging building blocks, such as nucleosides and nucleotides. It is also quite plausible that chemical evolution at this stage led to the 'optimized' four nucleobase types that we know today, which have been proposed by several others to be products of natural selection [29–31].

During this time period, ribonucleosides and perhaps deoxyribonucleosides also emerged [1,2,7]—though at concentrations significantly lower than those of amino acids, whose polymers already would have had time to undergo considerable selection for specific sequences with favourable properties. This scenario would explain the nucleoside-recognizing capacity of some highly conserved elementary peptide motifs, such as OB-folds and the Walker A motif. Because longer polynucleotides probably required more sophisticated biocatalytic intervention, we refer to this earlier period as the peptide-like/nucleo*side* stage.

2.3. Early peptide-polynucleotide stage

In the early peptide-polynucleotide stage, longer amino acid sequences arose which can form soluble mini-proteins by taking advantage of homo-oligomerization, metals and organic cofactors to support folding into globular entities. Such mini-proteins differ from the peptides of the earlier stage in that they form 'closed' symmetry homo-oligomers, in contrast with smaller peptides that tend to self-assemble into structures with much higher aggregation number through 'open' symmetry (point-group versus space-group). These globular mini-proteins in association with metals and cofactors may have supported a primitive metabolism capable of harnessing energy from redox gradients [84] and higher standard free energy substrates (in contrast with the previous stage, where 'energy' came primarily from changes in the activity of water inherent in wet–dry cycles). Evidence for this transition can be found in the mutual occurrence of peptide-length sequences in structurally unrelated domains; these elements have variously been referred to as 'supersecondary structures' [85] or 'themes' [86]. It has been previously noted that many of the most elementary protein folds (e.g. TIM barrels, ferredoxins and P-loop NTPases) have an inherent repetitiveness [86–89] which might be traced to short peptide motifs oligomerizing together.

On the other hand, condensing these longer peptide chains required a more efficient catalyst, notwithstanding the fact that homo-oligomeric mini-proteins would not be as amenable to auto-replication as amyloidogenic peptides would. Many paths to 'replication' of such mini-proteins are conceivable. It has been proposed that they could be selected for based on their ability to associate with polynucleotides, finding its apotheosis in the emergence of a large ribosomal subunit (LSU). In the LSU, these two molecule types found a symbiosis in which peptides benefitted from a catalyst that could more efficiently condense amino acids, while RNA benefitted from the protective shell afforded by its peptide binders [44,90–95].

At this early stage, nucleotide polymerization may have been carried out by peptide-assisted ribozyme [43,96] or by the mini-protein catalysts themselves, representing another example of symbiosis. Even though several examples of RNA-directed RNA polymerase ribozymes have been reported [3–6], these systems face several criticisms on their claim to

prebiotic relevance: (i) the requirement of triphosphorylated building blocks, (ii) the length and topological complexity of the ribozyme, (iii) reliance on divalent cation concentrations of the order of 100 mM and (iv) absence of their existence in the biological record (in contrast with the ribosome, which remains to this day). By contrast, the fact that extant RNA polymerization uses protein-based catalysts and the recent discovery that the core fold of RNA polymerases (the double psi beta barrel) [25] can be reconstituted with a limited amino acid alphabet suggest that RNA may have always been polymerized with a protein-based instrument. Hence, these ideas highlight the importance of RNA–peptide coevolution in enabling RNA polymerization as well as peptide polymerization [97,98].

2.4. Late peptide-polynucleotide stage

Without doubt, one of the most important transitions during the prebiotic period was the emergence of a functioning translator in which protein sequences could be specified and encoded from a nucleic acid template. This functionality, which ushered in a 'late' peptide-polynucleotide stage, was not a small order: it required the emergence of the ribosomal small subunit (SSU), a set of aminoacyl-tRNA synthetases (aaRSs) to accurately set a genetic code, and a continuous supply of RTP (R = {A, G}; note that the peptidyl transfer reaction, mediated by the LSU, does not require ATP, and in principle could have taken advantage of a range of activated acyl precursors). The advent of translation fully relieved peptides of the need to self-assemble or self-select to perpetuate particular sequences, and at this stage, polynucleotides exclusively take on the role of informational polymer. We shall discuss the evidence for why and how certain extant protein folds became selected prior to their encoding by nucleic acids and why the LSU probably preceded the SSU (see §5).

At this stage, full-length proteins (rather than miniproteins) can be routinely synthesized, and this obviates the requirement for proteins to fold by oligomerizing small motifs. Polynucleotide segments encoding such motifs stochastically undergo duplication events, resulting in tandem repeats, in which each copy is then free to mutate, resulting in more specialized and less symmetrical proteins. While the parsimony of this type of evolutionary model is appealing, it is always important to point out that convergent evolution cannot be fully ruled out. Specific examples of this mechanism of evolution have been noted in the case of beta propellers [99], the KH domain [100] and Walker-type P-loop NTPases [89,101]. Moreover, the advent of chaperone systems like Hsp60 (GroEL) and Hsp70 (DnaK) further expand the classes of proteins that can be efficiently synthesized [102,103].

2.5. DNA–RNA–protein stage

This stage is characterized by the emergence of a DNA-encoded genome, whose building blocks are supplied through ribonucleotide reductases, and which can be replicated in a high-fidelity, highly processive manner by DNA polymerases. Prior to this stage, DNA and RNA could have been interchangeable and perhaps even co-mingled into polynucleotides with both types of monomers [7,8]. Because of the large size and complexity of modern DNA polymerases and ribonucleotide reductases compared to many other types of proteins, we imagine this at the final stage in our protein–nucleic acid coevolution scheme prior to LUCA, as it presupposes a translational machinery capable of synthesizing large multi-domain proteins. This stage also saw the completion of the canonical genetic code and the stable addition of 'late' cAAs which had to be biosynthesized metabolically (see §3). We envision the fixation of the amino acid alphabet as a late event in this sequence of events, as amino acids such as Trp, Tyr, Lys and Arg involve numerous enzymes in their biosynthesis and therefore likely required a DNA-based genome to retain the suite of catalysts necessary to prepare them. An important question that is worth reflecting upon (as we [20,24,104] and others [21–23,26–28,105] have shown) is how proteins managed to complete fairly sophisticated functions in the previous two stages without the full complement of amino acids (see below). The hallmark of this stage is the end of 'chemical evolution' as Darwinian biological evolution takes root thanks to the intrinsic stability of information encoded in DNA (compared to RNA), and the high (but not perfect) fidelity of polymerases which create the potential for point mutations, insertions and duplications.

3. Amino acid alphabets

Despite its wide variety, all life that we know uses the same alphabet of 20 cAAs (and rarely also selenocysteine and pyrrolysine, as the 21st and 22nd) to code for its proteins. There is ample evidence that the protein alphabet emerged via chemical and biological selection from a set of prebiotically available compounds to its current form [56,57,60,106]. Detailed analyses imply that the current set of 20 (as well as its hypothetical subsets) has unique adaptive properties compared with equal sets of random alternatives [107,108]. Within the chemical space, the canonical alphabet represents unusually optimal spectra of size, charge and hydrophobicity when compared with alternative alphabets [109]. These observations have added weight to a long-standing hypothesis that gradual incorporations of individual amino acids would probably steer the fitness landscape in a similar way, producing near identical sets if amino acid alphabets were to evolve on other Earth-like planets [108,110].

At the same time, the last two decades of biological engineering has informed us that many of the cAAs can be removed or substituted and that proteins can be constructed using amino acids beyond the canonical alpha-amino acids [111,112]. This line of research has been inspiring scientists to search the amino acid chemical space and to define similarly optimal 'xenoalphabets' [113]. Corresponding 'xenoproteins' would represent a great future tool to compare the canonical and alternative alphabets' optimalities with respect to creating polymers with structure and function.

3.1. Early versus late amino acids

During the earliest prebiotic stages, the chemical space that was accessible was probably limited to environmentally available compounds. The sources of these prebiotically plausible amino acids were both endogenous (synthesized on Earth in e.g. hydrothermal vents and atmospheric mixtures) and exogenous (delivered from outside the Earth, e.g. by meteorites). More than 80 and 20 different amino acids have been identified in meteorites and atmospheric spark discharge simulations, respectively, with only half of the 20 cAAs highly represented among these [10,114]. Three independent

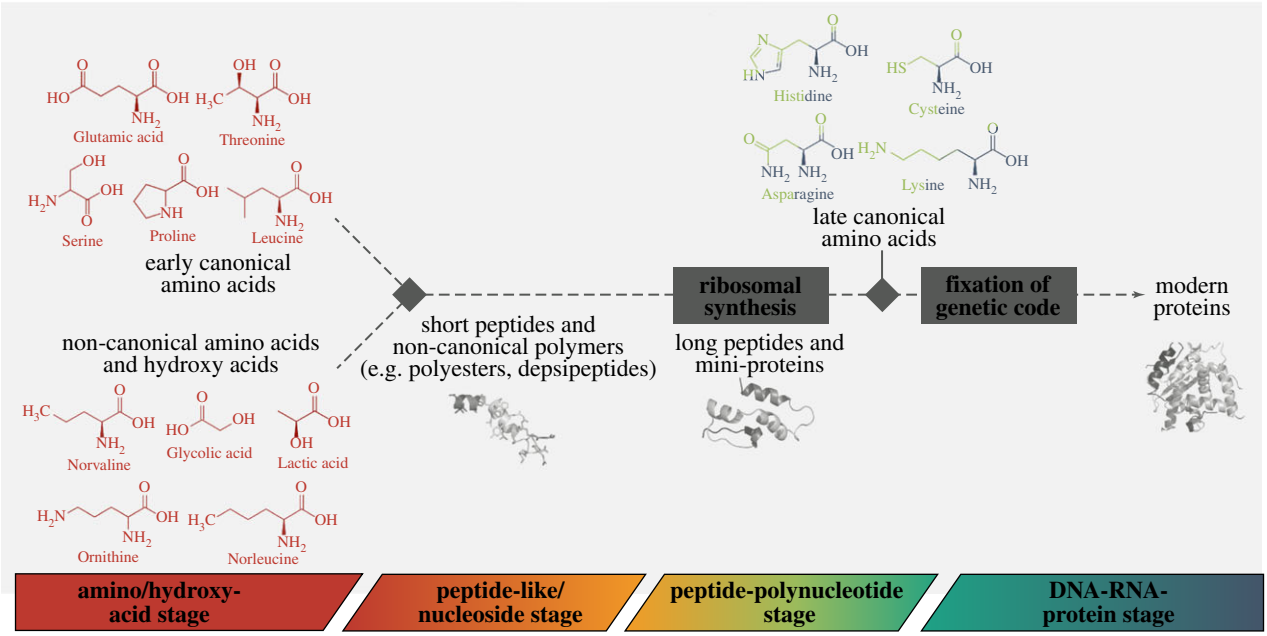

Figure 2. A model for the evolution of the amino acid alphabet.

meta-analyses ranked the prebiotic plausibility of amino acids and their additions to genetic code, agreeing on roughly the same 10 'early' amino acids within the current canonical alphabet [18,56,57] (figure 2). These early cAAs are Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr and Val. The other half of the modern protein alphabet comprises amino acids with higher synthetic costs, more complex structures and higher reactivity, and were most likely products of catalysts and metabolism [56,58]. Cysteine is not included in the list of early amino acids in the available meta-analyses, but it has been pointed out that the majority of the studies that these were based on did not include sulfur in the atomic mix of the experiments. When H_2S was later included in the Miller–Urey experiment, possible cysteine degradation products were detected suggesting that cysteine could be produced in primordial synthesis but was unstable and oxidized [10].

3.2. The early alphabet (smaller versus messy alphabet)

Preceding templated ribosomal proteosynthesis, early peptides were most probably constructed from a more diverse pool of monomers found in the prebiotic environment (figure 1). As mentioned above, a plethora of non-canonical amino acids (ncAAs) and a broader class of amino acid structures, and even some alternatives to amino acids (such as beta- and gamma-amino acids, hydroxy acids or dicarboxylic acids) have been detected in different prebiotic sources and their possible involvement in building early polymers has been considered by several researchers. Some of the alternative linear aliphatic amino acids (such as alpha-aminobutyric acid, norvaline and norleucine) have been detected in various prebiotic settings in similar amounts as the early cAAs [115–117]. These amino acids have been moreover identified as promiscuous targets of some aminoacyl-tRNA synthetases, suggesting that these ncAAs may have had earlier relevance even in the evolving genetic code [118,119]. Similar debates have been rising about the positively charged amino acids as none of the canonical ones (Lys, Arg and His) have been observed among the early set of the alphabet. At the same time, their ncAA analogues with fewer methylene groups (such as ornithine, 2,4-diaminobutyric acid and 2,3-diaminopropionic acid) appear to be more accessible prebiotically [120]. In today's life, cationic amino acids are indispensable and especially key for interaction with nucleic acids [121]. Their absence among the early alphabet represents a barrier to many hypothesized scenarios about protein evolution and hence an important role played by their ncAA analogues during early stages of the genetic code development has been proposed [101,122–124]. Nevertheless, several examples in the literature now show that folding of acidic proteins can be assisted by metal ions or other cationic species to compensate for the lack of positive charges [20,21].

It has been argued that the beta- and gamma-ncAAs (which have been also identified prebiotically, albeit usually in lower yields when compared with alpha-AAs) and polymers built of similarly prebiotically available compounds (such as the aforementioned hydroxy acids or dicarboxylic acids) would be less prone to form secondary and tertiary protein structures than alpha-AAs [106,110]. At the same time, this does not rule such oligomers out of possible prebiotic relevance [106]. Both helical and beta-sheet-like conformations have been observed in beta-AA polymers [125]. Much attention has been recently devoted to polyesters and depsipeptides that have been shown to form during model prebiotic reactions driven by wet–dry cycles from alpha-hydroxy acids or combinations of alpha-AAs and alpha-hydroxy acids, respectively [13,126]. Although such polymers are less stable than peptides, they nevertheless can form secondary structures [127], and it has been argued that they could have served as an important intermediate during chemical evolution. In conclusion, it seems very probable that short peptides incorporating ncAAs and alternative monomers preceded ribosomal synthesis during the peptide-like/nucleoside stage. During the early peptide-polynucleotide stage, the earliest LSU-synthesized peptides also likely incorporated ncAAs and alternative monomers prior to protein synthesis according to an RNA template [128,129]. Their potential role in shaping protein structure/function remains to be better described by the origins of life and

synthetic biology communities. Further evolutionary selection could have produced the canonical genetic code by fixing some early cAAs, purging others, and supplementing the early canonicals with the later structurally and functionally more complex additions.

3.3. Selection of the late amino acids

Although several analyses suggest the probable sequence of the late amino acid incorporation into the genetic code, many debates and questions about the order and factors that influenced these events remain open [56,130,131].

To start with, cysteine is regarded as one of the latest additions to the code according to the Trifonov meta-analysis [130]. At the same time, it is one of the most active and unique amino acids involved particularly in Fe–S clusters (such as in ferredoxin, considered one of the earliest protein domains) and conflicting hypotheses have been proposed as to whether these features were indispensable in early evolution. Powner's group made an argument towards indispensability of Cys in early biological processes, while Moosmann *et al*. suggested that Cys-mediated features could be either ignored or replaced in LUCA [33,131]. One of the main arguments for the late emergence of Cys in the AA alphabet was the absence of its plausible prebiotic synthesis pathway, although this has been recently challenged, albeit in near-neutral pH and low-temperature conditions [33]. However, as mentioned above, many of the prebiotic synthesis experiments did not include sulfur in the source material and when they did, possible degradation products of cysteine were detected probably as a result of its oxidation and therefore implying its conceivable prebiotic synthesis [10]. The ease of Cys degradation was also listed as a factor for its later importance in Wong's theory of amino acid alphabet evolution [132]. Interestingly, the Cys biosynthesis pathway was successfully re-engineered using enzymes lacking cysteine residues, providing an important proof-of-concept that a Cys biosynthetic pathway could be supported by proteins lacking this amino acid [133]. As an aside, it is important to point out that ferredoxin's emergence was also discussed in terms of a simple 'theme' comprised of early amino acids (initially without cysteine) duplicating and later adding cysteine, conforming to the overall model in figure 1, with the addition of late amino acids representing a relatively late stage during pre-LUCA evolution [134].

Another noteworthy amino acid is histidine, which is the most widely employed catalytic residue in enzymes [135]. It was hypothesized that His might have come from a catalytic nucleotide and could be derived from pre-existing purines [118]. At the same time, Shen *et al*. showed the possibility of non-enzymatic His synthesis, which was later disputed as unrealistic under prebiotic conditions [123,136]. Interestingly, Lazcano *et al*. postulated that the extant biosynthetic pathways of His and purine syntheses have evolved separately, with purine synthesis predating that of His [123]. His is therefore regarded as a late amino acid along with the other positively charged cAAs, despite their high relevance in today's protein alphabet. As described above, it has been argued that their important role was substituted by some of the prebiotically plausible basic ncAAs.

Finally, Met, Trp and Tyr are considered the latest additions to the amino acid alphabet [58,130]. A study by Granold *et al*. [58] suggests that they were incorporated into the genetic code during the great oxidation event as they showed antioxidant properties. At the same time, this event would render the employment of Cys toxic to cells, which posed a challenge that early life had to cope with. Intriguingly, the Granold *et al*. study also suggests that Cys might have been an earlier addition to the AA alphabet than previously estimated. An important final consideration to raise is that some of the late cAAs may have arisen transiently or were present in low concentrations in specific environments at earlier stages; we would argue, however, that their 'stable' presence (and certainly their incorporation into a genetic system) would have required metabolism.

3.4. Protein consequences of the evolving alphabet

The earliest peptide/protein-like polymers were probably random (statistical) sequences [137,138]. Using modern tools of synthetic biology, several groups have mimicked random sequences from the canonical alphabet or its reduced subsets, in search of their general properties (summarized in Tong *et al*. [139]). In short, random sequences can inherently form secondary structures similar to their occurrence in biological proteins and between 5 and 20% of random peptides of lengths 80–100 amino acids have been reported capable of undergoing compaction/folding [140–143]. Specific functions have been selected from libraries of random or highly randomized sequences implying that the structural and functional propensities of randomly generated peptides are compatible with an early role in evolution [85,144–146]. Importantly, some of these studies have also informed us that proteins constructed from the limited subset of the early canonical alphabet are in fact more soluble, similarly prone to secondary structure formation and perhaps structurally more compact than if built from the full alphabet [20,147–149]. Hence random peptides likely provided a well-spring of potential, from which chemical evolution could act to select individual species with favourable structural, catalytic or compartmentalizing properties.

Complementary top-down approaches, or 'reverse evolution', have been used to study the effect of the alphabet reduction on protein structure/function of select protein targets. Most of the earlier studies that reduced the amino acid composition (of small selected proteins such as a beta-trefoil fold, SH3 domain and nucleoside kinase) towards the early cAAs reached an alphabet size of 10 to 13 and/or 80–90% early AA composition [22,23,26,27,105]. These studies reported that folding as well as activity can be preserved in these potentially ancient sequences, although decreases in both structural stabilities and catalytic activities were observed. Longo *et al*. pointed out that reduced protein stability can be improved by a halophilic environment when aromatic core packing interactions are missing in the structure [22,27]. The studies from the Akanuma group argue that the early cAAs are sufficient for folding and stability while the late cAAs were recruited to achieve efficient catalysis [23,28]. By contrast, the work by Longo *et al*. [27] suggests that some of the late cAAs are crucial for the evolution of structural stability. A recent mutation study of a dephospho-CoA kinase where all the aromatic amino acids were substituted resulted in the loss of structural stability, though a transition from a molten globule-like structure to a compact functional fold was observed upon ligand binding [21]. The emerging scenario is that while less stable and less functional mini-proteins can still be constructed in the absence of late cAAs, the

deficiencies of these proteins (e.g. lacking positively charged and aromatic amino acids) can be compensated for by high salt concentration, the presence of divalent metal cations or binding to organic cofactors. In agreement, a study by the Tawfik group observed that polyamines and divalent cations can promote folding of highly acidic proteins [21]. We have recently observed that under cell-like conditions, random sequences formed from the 10 early cAAs exhibit similar structure-forming propensity as the full alphabet repertoire despite their very acidic nature. Unlike the full alphabet proteins, they are intrinsically more soluble and exhibit these properties independent of molecular chaperone activities [104]. Importantly, an RNA-binding domain was recently reconstructed using an alphabet of only the 10 early cAAs, uncovering metal cation mediated interaction between the RNA and negatively charged cAAs [20]. Along with a recent reconstruction of an RNA-binding peptide incorporating ornithine as a prebiotically available cationic ncAA, the study by Giacobelli *et al*. provides an important lead to how an early metabolism could function in the absence of late cAAs [20,101]. It is intriguing to speculate that the early preference for acidic AAs over basic AAs (which in turn primarily coordinate metal cations over halogen or chalcogen anions) left a lasting imprint to modern biology in that: (i) virtually all extant proteomes are more acidic than basic [150], (ii) cells maintain metal cations at significantly higher concentrations than elemental anions (the most abundant anion in most cells is in fact glutamate, an early cAA) and (iii) signalling disproportionately uses cations over anions. Therefore, prebiotically plausible ncAAs, metal cations and cofactors have had a lasting impact on extant proteins.

4. Evolutionary significance of cofactors

Cofactors are essential components of many of today's proteins. They stabilize certain protein structures and are required for the catalytic activity of many enzymes. Most of the core cofactors are highly conserved across the three domains of life (with some important exceptions among methanogens, e.g. coenzyme M and factor F430), and they would have played an important role in the earliest evolution of peptides, i.e. during the peptide-like/nucleoside and early peptide-polynucleotide stages according to the model presented above (figure 1). From a chemical point of view, cofactors can be divided into two major classes: inorganic cofactors represented by metal ions (§4.1) and organic cofactors (§4.2).

4.1. Metal cation cofactors

In modern biology, various metal cations (K, Mg, Ca, V, Cr, Mn, Fe, Co, Ni, Cu, Zn and Mo) are involved in over half of the functionally annotated proteins [40], playing an important role in diverse catalytic functions (especially electron transfer) and maintaining the structural integrity of many protein folds. The overall abundance and accessibility of these metal ions has been affected over the evolution of Earth due to changes in the redox state of the ocean and the atmosphere via geochemical and biological processes [151]. While minerals likely played an important role in prebiotic chemical evolution [152] and also could have been part of metal cofactor-associated protein catalysts [153], in this section, we mainly discuss the contribution of metals in their soluble cationic form.

The estimated time interval for the origin of life on Earth ranges from 4.5 Ga to 3.7 Ga (Hadean to early Archaean), according to evidence of earliest habitability and the biosignature boundary [71]. If abiotic peptides existed and contributed to the origin of life and their early evolution, the accessibility of metal ions presumably shaped the types of proto-metalloenzymes that catalysed the key chemical reactions to sustain the primordial biological system. During the Hadean–Archaean period, trace amounts of oxygen favoured highly soluble Fe^{2+} in reducing environments, whereas the accessibilities of Mo, V and Cr were limited [154]. Mn, Co and Ni were present in the Archaean ocean, presumably in the high nM to µM range, but Cu and Zn were believed to be extremely scarce [151].

A recent reconstruction of ancestral metalloenzymes [155] showed, paradoxically, a universal preference for Mo over Fe^{2+} in nitrogen fixation, indicating that the selection of metal elements for proto-metalloenzyme catalysis might not be solely determined by global geochemical abundance, but also actively a consequence of selection for function [156]. Moreover, the local abundances of metals along with their interacting peptides vary between each niche environment and therefore drive different chemical reactions. For example, high concentrations of the transition metals Zn^{2+} and Mn^{2+} (which were globally rare during the Archaean) could be achieved in a geothermal pond where cooled geothermal fluids and condensed vapours resulting from a volcanic activity can enrich certain metal ions based on their difference in boiling temperature [51]. Such vapour condensed environment is favourable to concentrate K^+ ion over Na^+ and thus considered as a plausible environment for protocell evolution to achieve the consistent high K^+/Na^+ ratio that we see in almost all modern cells.

The alkaline earth metal ions Mg^{2+} and Ca^{2+} are deeply involved in diverse biological functions. A model considering continental weathering and hydrothermal alteration of seafloor crust estimates that in the Archaean ocean, Mg^{2+} and Ca^{2+} existed at mM concentrations with higher abundances for Ca^{2+} [157]. Overall, these metal ions were very accessible, both globally or locally, in the Hadean–Archaean ocean and terrestrial sites—an observation that is consistent with their importance in supporting the structure and functions of primitive polypeptides and polynucleotides.

Many translation-related proteins such as tRNA synthetase and translation factors require Mg^{2+} (in some cases Zn^{2+} is also needed) for function and in addition use Mn^{2+} for structural integrity [51]. Mg^{2+} plays many different roles in tRNA synthetases such as ATP binding, amino acid activation and the pyrophosphorolysis of the aminoacyl adenylate [52]. We would like to point out that while the prebiotically available concentrations of Mg^{2+} would have been sufficient for binding to peptides and stabilizing RNA tertiary structures, it would not have been high enough for the activities of many ribozyme replicases that have been developed by laboratory evolution. This discrepancy may suggest an earlier dependence of polynucleotides on peptides, which can relieve the requirements for high Mg^{2+} [43].

Based on the MetalPDB [158], we found that the carboxylate moiety of the two acidic amino acids (Asp or Glu) is strongly associated with Mg^{2+} binding through electrostatic interaction, along with several minor examples of coordinating

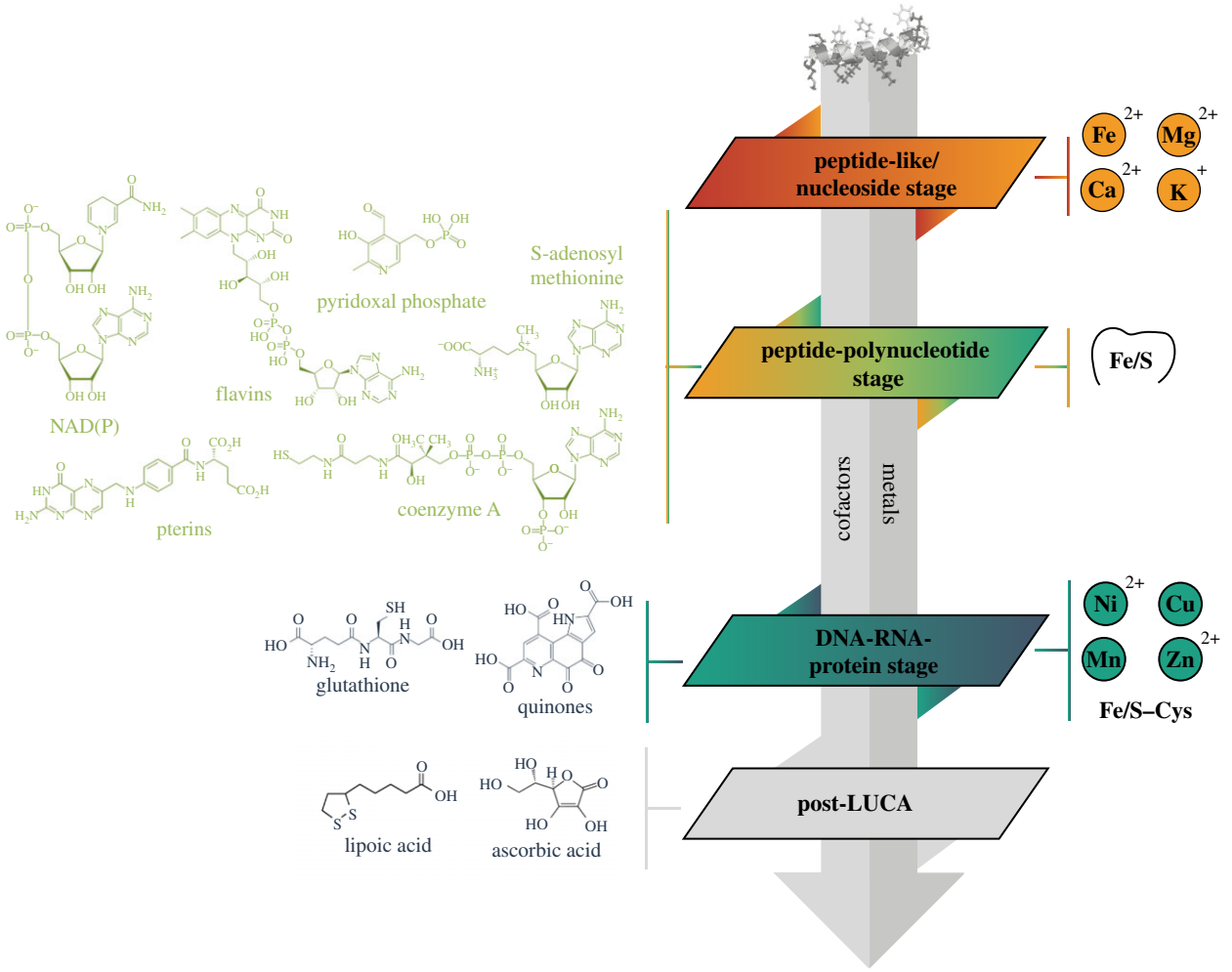

Figure 3. Chronology of the addition of organic cofactors and metal cations at distinct stages.

amino acids such as Asn, Gln and Ser. It is notable that Mg^{2+}, which was an important cofactor during the earliest stages, is typically coordinated by early amino acids, and not by the late amino acids His and Cys, which are more significant for coordinating Cu and Zn. Hence, this observation supports the view that Asp, Glu, Mg^{2+} and Ca^{2+} represent an 'early cohort' of amino acids and metals, while His, Cys, Cu and Zn^{2+} represent a later cohort. Indeed, the carboxylate moiety in minimal metal-binding peptide motif (DXDXD) has been reported to chelate Mg, Mn, Ni and Zn in various modern enzymes and thus has been proposed as one of the earliest metallopeptides [159].

One of the most studied examples of metal coordination within a protein–RNA complex is the ribosome. Mg^{2+} has long been identified for contributing to the overall maintenance of its structure by serving as a counterion to the phosphate moieties and is also concentrated within the peptidyl transferase centre (PTC) [44]. Importantly from a prebiotic perspective, Mg^{2+} in the ribosome can be substituted with Fe^{2+}, a metal that was more abundant in early oceans [53]. A recent study by Rozov and co-workers unveiled the positions of K^{+} ions within the ribosome, indicating the involvement of K^{+} in mediating rRNA–rRNA and rProtein–rRNA interactions, as well as in the PTC by increasing the stability of rRNA and tRNA [160]. It is interesting to note that K^{+} ions were found in pockets formed by the negative ends of the dipoles of carbonyl oxygen atoms from the polypeptide backbone. Similarly, clefts consisting of backbone N–H groups form so-called 'nest' structures [161] which are also important in binding various anionic groups such as phosphates, sulfates, carbonates and iron–sulfur (Fe/S) centres [162]. A good example is the Gly-rich P-loop Walker-A motif that binds ATP or GTP. Because of this simple binding mode with very little side chain involvement, nest motifs of oriented N–H groups and carbonyls within peptide loops might be considered as one of the earliest functional protein motifs [163]. It is appealing to hypothesize that such nest motifs could have occurred in the early Cys-less ferredoxins, employing backbone N–H groups to coordinate Fe/S centre, perhaps in addition to organic sulfides such as methanethiols; indeed even modern ferredoxins show a distinct bias to orient N–H dipoles toward their iron–sulfur centres (figure 3).

The other prebiotically abundant alkaline earth element is Ca^{2+}, which is widely used in intracellular signalling in eukaryotes, and also frequently stabilizes proteins in thermophiles [164]. Ca^{2+} exhibits a high affinity for carboxylates and rapid binding kinetics (100-fold faster than Mg^{2+}) making it a useful metal cofactor [165]. The amino acid chelators for Ca^{2+} are similar to those of Mg^{2+}; namely, predominantly Asp and Glu, followed by several other supporting amino acids (Ser, Thr, Asn and Gly). Unlike Mg^{2+}, cellular Ca^{2+} ion is maintained at extremely low concentrations (10^{-7} M) to prevent Ca^{2+} from precipitating peptides, inorganic phosphates and phosphate-bearing biomolecules [166]. The tendency for Ca^{2+} to precipitate phosphate and polynucleotides could possibly explain biology's preference for Mg^{2+} and why active mechanisms are used to pump it into specific membrane-bound compartments in modern cells. From an early prebiotic

context, however, when ion-impermeable compartments and active efflux mechanisms were likely not available, these observations raise the question as to whether early stages of life sought environmental niches where other anions could have depleted Ca^{2+} from solution.

Finally, considering the emergence and evolution of metabolism, transition metals are essential for the redox chemistry to provide key precursors of biomass from simple inorganic compounds including the essential atoms C, H, O, N, S and P. In view of the global abundance on early Earth and the versatility of chemical reactions, Fe and Fe-bearing minerals clearly stand out due to their non-enzymatic reactions resembling those of ancient cofactors [167] and core metabolic pathways [168,169]. For example, warm acidic Fe^{2+}-rich water can promote a reaction network recapitulating most of the biological TCA and glyoxylate cycle intermediates [170]. Partly electro-reduced FeS ($Fe/S\text{-}Fe^{0}$) is able to catalyse reductive amination leading to the formation of several amino acids from α-ketoacids and ammonia under alkaline condition [171]. These results indicate the versatility and importance of iron-promoted protometabolism, which was eventually taken over by enzymes that harbour metal centres and organic cofactors due to their improved efficacies and specificities [172].

4.2. Organic cofactors

Some organic cofactors are considered evolutionary ancient molecules of prebiotic origin, while others are probably the inventions of early biochemical metabolism [41]. In most cases, they likely originated independently of proteins [173] and the binding of cofactors to primitive polypeptides appears to have been a critical step in protein evolution. The early cofactors might have facilitated protein formation as catalysts (to build amino acids or peptide segments), as molecular chaperones (to facilitate protein folding), and/or as selectors (because of the important function of early cofactors) [173].

Based on the available studies, cofactors can be divided approximately into three categories based on their evolutionary age: (i) ancient cofactors (associated with the peptide-polynucleotide stage) include those that could have been synthesized under prebiotic conditions and therefore existed before the establishment of protometabolism; (ii) early cofactors (associated with the DNA–RNA–protein stage) include chemical moieties that most likely appeared only after the emergence of the first proto-cells but were present in the last common universal ancestor (LUCA); and (iii) late cofactors were developed after the divergence of three domains of life from LUCA (figure 3).

The most ancient cofactors are thought to include nucleotide-derived cofactors (NAD(P), FMN, FAD, coenzyme A, S-adenosylmethionine, pterins and pyridoxal phosphate). Many of these are composed of ribonucleoside or nucleotide units (NAD(P), coenzyme A, S-adenosylmethionine) or they are biosynthetically derived from nucleotides (FMN, FAD and tetrahydrofolic acid). Several cofactors (NAD(P), FAD and coenzyme A) contain AMP as a structural element which is not involved in catalysis but rather serves as a 'handle' for binding to enzymes [174]. Nucleotide-containing cofactors together with inorganic cofactors represent two main groups of cofactors that are assumed to be of prebiotic origin and played the primary role in protein evolution [42]. While metal ions and minerals that resemble metal ion clusters found in the modern proteins (such as Fe/S clusters) should have been widespread in the primordial Earth environment, nucleotide-containing cofactors' provenance in the peptide-polynucleotide stage provides a direct testament to how these two types of molecule types coevolved intimately during early chemical evolution [175–177]. Nucleotide-derived cofactors may have helped facilitate the jump from peptides to mini-proteins (figure 1). Ji *et al*. [173] noted that domains associated with binding nucleotide-derived cofactors are among the most ancient as based on the diverse range of folds associated with binding these cofactors as a core function. For instance, there are 35 folds associated with binding ATP, 27 for NAD(P), 21 for FAD, 16 for FMN, 15 for GTP, 14 for CoA and 13 for SAM. Collectively, this argues that the earliest globular domains were probably selected for their ability to bind cofactors, an activity that was particularly salient in the absence of late amino acids. This observation is in agreement with the current understanding of evolutionary history of protein folds according to which the P-loop NTP hydrolase fold, the adenine nucleotide alpha hydrolase fold (both using ATP), the flavodoxin-like fold (using NADH, NADPH and $FADH_2$), the SAM-dependent methyltransferases fold (using SAM) and the NAD(P)-binding Rossmann fold are among the most ancient protein structures [178–181]. The highly ancestral nature of domains which bind to nucleotide-derived cofactors serves as indirect evidence that such cofactors may have helped stabilize ancient versions of these globular proteins (via 'induced folding'), which were likely divided up into more (and shorter) polypeptide chains, and lacked the greater stability that could be imparted by incorporation of late AAs and longer chain lengths.

5. Refoldability

Sophisticated macromolecules need to be able to fold into well-defined globular structures and maintain those conformations to perform their functions. This capability likely emerged during the early peptide-polynucleotide stage when peptide chains long enough to create a hydrophobic core arose (though it should be noted that due to symmetry and self-assembly, they did not need to be as long as some of today's globular domains, which emerged in the late peptide-polynucleotide stage). That proteins can fold into a single (or small number of) well-defined conformation(s) can be explained by the fact that stable folded structures require hydrophobic residues to form a tightly compacted core, but they have idiosyncratic shapes, and are connected together through a continuous chain that can only bend in specific ways. Satisfying these requirements simultaneously gives protein folding a puzzle-like quality that results in relatively few solutions. Stated another way, the free energy landscape that describes globular polypeptides has a funnel-like architecture with a single minimum (or a small number of minima), ensuring that the native state reflects a thermodynamically stable state, and that there are many possible paths to get there from an arbitrary unfolded state (figure 4).

5.1. The significance of refoldability during the early peptide-polynucleotide stage

Funnel-like energy landscape topologies endow polypeptides with the important feature of reversible refoldability. A consequence of this trait is that when a small protein suffers

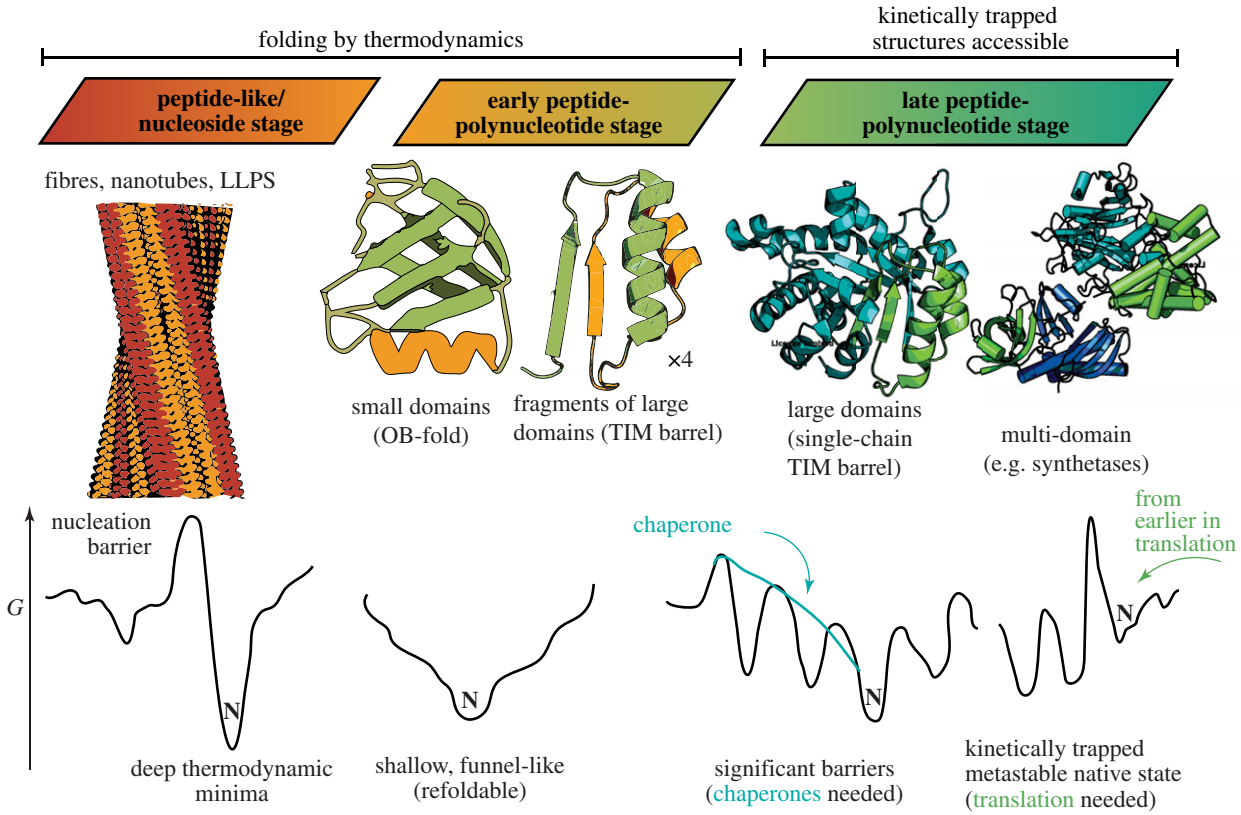

Figure 4. Chronology of peptide and protein topologies available at different stages from the perspective of foldability.

some shock (in temperature, pressure, pH or other condition) that causes it to unfold, it can return to its native structure without any external assistance upon returning to native conditions. The observation that many simple proteins are refoldable implies two features about a protein's energy landscape: first that the native state is indeed the global minimum, and second that it is accessible on a given timescale (i.e. barriers and traps en route to the native state are not too high or deep).

Refoldability was probably an essential feature for early proteins during the early peptide-polynucleotide stages. Because of the absence of a dedicated proteostasis machinery, the only protein quality control system that was available was thermodynamics. A recent work has suggested that LUCA had only one major chaperone, namely Hsp60 (GroEL), as even the essential quasi-universal Hsp70 (DnaK) may have only appeared after the archaea–bacteria divergence [102]. However, GroEL requires a continuous supply of ATP, which necessitates metabolism and hence was probably not available until the *late* peptide-polynucleotide stage. While GroEL would have unlocked protein folds that are harder to reach due to intervening traps and barriers, the biomacromolecules of the early peptide-polynucleotide stage must have had native states that were straightforward to access, unassisted, by following a free energy gradient.

Because refoldability represented an important feature for the earliest proteins, folds that display this canonical biophysical attribute were likely present earlier during the origin of life. A recent study by To *et al*. [182] interrogated the refoldability of the *E. coli* proteome and found that two-thirds out of approximately 1200 proteins were reversibly refoldable on a biologically relevant timescale of 2 h. This group was enriched with monomeric proteins (75% refoldable), single-domain proteins (70% refoldable), small proteins (less than 20 kDa, 80% refoldable) and proteins without any annotated domains (and hence, more likely to be disordered, 87% refoldable).

It was also found that some of the folds that are believed [180,181] to be the most ancient (e.g. OB-folds, 3-helix bundles, ferredoxin-like domains, flavodoxin-like domains, SH3 domains and SAM-dependent methyltransferase domains) were predominantly refoldable (greater than 80%).

The TIM barrel fold is an interesting case study, because it is also often cited as being one of the most primordial fold types, though the study by To *et al*. found it not to be among the more highly refoldable fold types (65% refoldable). This is consistent with the observation that TIM barrels disproportionately require the assistance of GroEL [183]. The relatively larger size of TIM barrels compared to other elementary globular domains, and their eightfold pseudo-symmetry suggest that in the early peptide-polypeptide stage, TIM barrels' antecedents were able to prevail, in the form of a shorter beta-alpha motif that self-assembled into barrels, but that their concatenation into the single-chain TIM barrels that we know today had to wait until a later stage when chaperones (or translation) became available. This same hypothesis probably also holds for the P-loop NTPase fold, also frequently noted as one of the most abundant and ancient protein folds [184]: it probably started as a simple beta-loop-alpha motif (the so-called Walker-A motif), and its expansion and diversification were driven by the availability of chaperones and translation [89].

One of the unexpected findings in To and co-workers' study was the finding that virtually all the ribosomal large subunit proteins were refoldable in a complex mixture under prebiotically plausible conditions. On the other hand, many small subunit proteins as well as translation factors were found *not* to be reversibly refoldable. This finding furthers the notion that a large subunit functioning as a ribozyme peptidyl transferase (a 'proto-PTC') evolved earlier than and independently of the coordinated decoding of tRNAs by a functional small subunit (figure 1) [90,185].

5.2. Chaperones and translation give rise to a late peptide-polynucleotide stage 'explosion'

The advent of chaperones and translation, which we assign to the late peptide-polynucleotide stage, resulted in the elaboration and diversification of larger harder-to-fold domains, as well as multi-domain proteins. Chaperones play a critical role of burning energy to re-extend proteins that are trapped in an intermediate misfolded state, thereby allowing them a fresh chance to fold, in a mechanism referred to as iterative annealing [186,187]. This function was important for the larger domains like TIM barrels and P-loop NTPases, smoothing the transition from self-assembly of smaller motifs to longer chain lengths that emerged through genetic duplication [188].

Translation had major consequences for the types of proteins that could be easily created because it enables proteins to fold co-translationally [189,190]. Co-translational folding facilitates access to kinetically trapped (metastable) native states because it can 'seed' proteins in one region of their energy landscape at an early chain length and then retain them there if synthesis proceeds faster than the egress rate out of that region. It is also generally important for multi-domain protein folding to decouple the folding of individual domains, which would otherwise be prone to generate improper inter-domain contacts [191]. Translation also allows proteins to join together into more diverse complexes by enabling the coordination of protomer folding and subunit assembly. This allows 'obligate complexes' to be routinely synthesized. By contrast, the simpler oligomers that were accessible during the early peptide-polynucleotide stage probably needed to be able to reversibly assemble from independently stable protomer units. At this later stage, protein assembly no longer needs to be dictated exclusively by thermodynamics, and kinetically trapped protein assemblies become accessible (figure 1).

To and co-workers found that virtually all the *E. coli* aminoacyl-tRNA synthetases (aaRSs) are nonrefoldable. On one level, this finding is not surprising, given that these enzymes are in all cases multi-domain proteins, and many use more specialized fold types (e.g. anticodon-binding domains, the class II synthetase fold and the HUP domain). Nevertheless, this finding argues for an intriguing point: just as aaRSs are essential for protein translation from a nucleic acid template, aaRSs themselves also necessitate translation (or chaperones) to properly fold. Moreover, the emergence of aaRSs is a prerequisite for the small subunit of the ribosome to perform its core function in tRNA decoding. In other words, long viewed as among the most ancient protein folds, aaRSs may actually be relatively new in comparison to smaller domains or those which can be split into smaller repetitive themes. To summarize, evidence from refoldability argues that the synthetases, the small subunit and translation all bear hallmarks of a later stage of development, and it is likely that the three emerged together because of their mutual interdependence. In our view, these developments defined the late peptide-polynucleotide stage, and with them, the relaxing of the requirement that proteins' native states be easily locatable.

5.3. Regarding the refoldability of RNA

RNA is often described as having a 'rougher' energy landscape than protein with more near-degenerate minima [192,193]. This character can be attributed to the dominant role of base-pairing, which has an additive quality, and the fact that in RNA, secondary structure is largely decoupled from tertiary structure [55]. As a consequence, many possible conformations with the same (or similar) number of Watson–Crick base pairs are roughly degenerate. This is a major contrast with protein folding, which is characterized by a highly cooperative hydrophobic collapse, and wherein secondary structures are relatively unstable outside the context of a tertiary structure. With these features, single mutations can result in total destabilization of a folded form [194,195].

From a computational perspective, the contrast makes the protein folding problem a more formidable puzzle. But from an Origin of Life perspective, it means simple proteins have a useful trait in their propensity to occupy a single (or small number of) native state(s) that can be reversibly relocated. The intrinsic structural heterogeneity encoded in RNA's energy landscape can be overcome biologically through co-transcriptional folding [196,197], RNA chaperones [198] or suppressed *in vitro* through careful (but arbitrary) annealing schedules or serial dialyses. However, processive RNA polymerases probably only emerged during the late peptide-polynucleotide stage. In the remarkable case of the ribosomal large subunit, which appears to be intrinsically reversibly refoldable, the rRNA refolding process is very likely chaperoned by extensive RNA–protein interactions, wherein rProteins with well-defined tertiary structures induce rRNA to choose specific base-pairing patterns over alternatives [198]. In the other particular case of tRNA (also of ancient provenance, and easily refoldable), refoldability is possible because of high stability and simple topology (i.e. base pairs form across adjacent regions that are separated by short loops).

On the other hand, for intricate ribozymes with topologies more complex than tRNA and fewer protein interactors than ribosomes, inherent refoldability is far from guaranteed. It should give us pause that no ribozyme (aside from the ribosome) is universally distributed or confidently traceable back to LUCA [38,199]. Indeed it has been pointed out that 'There is no conclusive evidence that intron self-splicing and ribozyme-mediated RNA processing are truly primordial activities' [199]. We note that 'strong' RNA world hypotheses that led to the assertion that LUCA was a protoeukaryote (because eukaryotes alone habour the majority of extant catalytic RNA) are inconsistent with current models for the root of the tree of life [200,201]. Finally, it should also give us pause that the remarkable ribozymes discovered in recent decades through directed evolution have all themselves been birthed from sophisticated and processive RNA polymerases, providing the luxury of co-transcriptional folding that was likely not available until at least the late peptide-polynucleotide stage. More research is necessary to elucidate RNA refoldability, as it remains an understudied area. On balance though, there is preliminary evidence to suggest that complex RNAs 'leaned on' peptides and proteins to help tame their rough energy landscapes' proclivity toward structural heterogeneity. Hence, evidence from refoldability argues for another important way in which ancient RNA and proteins needed to cooperate to support key functions during the emergence of life, with the primordial trait of high intrinsic refoldability more generally associated with proteins.

6. Conclusion and future outlook

In this mini-review, we have sought to bring together evidence from molecular biology, bioinformatics, geochemistry

and biophysics to provide insight into the emergence of proteins during the early stages of the origins of life, prior to LUCA. We advocate for this period to be separated into five 'stages' (figure 1): (i) the amino/hydroxy acid stage, (ii) the peptide-like/nucleoside stage, (iii) the early peptide-polynucleotide stage, (iv) the late peptide-polynucleotide stage and (v) the DNA–RNA–protein stage. Through this classification, we seek to highlight the ways in which the antecedents of today's proteins and nucleic acids cooperated and were interdependent on each other at distinct stages of emergence.

Proteins are sometimes viewed as being a later development during the Origin of Life, on account of the fact that they cannot self-replicate in modern biology, and are synthesized in accordance with an RNA template by translation. However, proteins have a number of qualities that make them ideal for supporting early transitions toward biological complexity during the origin of life, including: (i) the abundance of their components from abiotic terrestrial and extraterrestrial sources, (ii) the relative facility of their condensation, (iii) their self-assembly properties into complex morphologies, (iv) their catalytic versatility, (v) the reducibility of their alphabet, (vi) their propensity for intrinsic refoldability and (vii) their capacity to interact with a range of cofactors. Each of these factors played important roles at distinct prebiotic stages. High abiotic abundance of amino acids played an extremely important formative role (during the amino/hydroxy acid stage). Condensation of amides and self-assembly of short peptides (as well as depsipeptides) into large structures were particularly relevant during the peptide-like/nucleoside stage and could have afforded nature some of its earliest molecular scaffolds, compartments and catalysts. The availability of early biocatalysts could have accelerated the availability of other building blocks whose synthesis is more challenging, such as nucleosides and lipids.

Spontaneous folding and refoldability was paramount during an early peptide-polynucleotide stage when quality control mechanisms to maintain biomacromolecules' conformations was not yet available, while meanwhile ancient globular protein folds, composed of smaller self-assembling constituents composed of a smaller palette of amino acids, appeared. These folds had a strong propensity to bind nucleotide-containing cofactors, which expanded catalytic versatility and provided additional stability. Close interactions between such proteins and polynucleotides likely chaperoned polynucleotide folding, while also providing a means for proteins to propagate through association with the more easily replicating polynucleotides. This interdependence resulted in the ribosomal large subunit.

The invention of translation enabled larger proteins to appear that are harder to fold and relieved proteins of any need to propagate themselves. At the same time, protein-based polymerases allowed complex RNA topologies to appear that did not rely on protein binders to help tame their rougher energy landscapes. In our chronology, all of these developments occurred prior to the advent of late amino acids, the fixation of the genetic code and the establishment of DNA-based genomes.

The model we present for the Origins of Life is not without its limitations. The potential of peptides to self-propagate without a nucleic absent template is greatly understudied, and more examples of this behaviour are needed to support the peptide-like/nucleoside stage. Proteomics technologies are still outstripped by nucleic acid sequencing technologies, though are consistently improving, and may help shed much needed light on chemical evolution of peptides. While there are examples of catalytic amyloids, evidence of their catalytic utility being directed toward the synthesis of other prebiotically relevant molecules is lacking and would represent an important discovery. We still have many questions surrounding how specific mini-protein sequences could have been maintained before being directly encoded by replicating genetic material. With the current renaissance underway in Origins of Life research, we are optimistic these currently mysterious aspects can be addressed by future experiments.

At the same time, we hope in this mini-review to motivate a deeper acceptance about the implications of peptide–polynucleotide coevolution. These biomolecule types did not appear in isolation, and this fact should be more reflected in our experiments. For instance, research on prebiotic polynucleotides should consider peptide 'cofactors' rather than use unrealistically high divalent cation concentrations [43]. Research trying to resurrect early proteins composed of fewer amino acids should be more inclusive of nucleotide-based cofactors. And prebiotic chemistry should more strongly consider how selectivity could be afforded by simple catalysts, rather than by high-performance liquid chromatography. In general, models that try to assert the preeminence of one biomolecule type over others during the Origin of Life will probably prove incorrect in the long run. Like people, when different types of biomolecules work together, amazing things can happen.

Data accessibility. This article has no additional data.

Authors' contributions. S.D.F.: conceptualization, investigation, project administration, writing—original draft and writing—review and editing; K.F.: conceptualization, investigation, writing—original draft and writing—review and editing; M.M.: investigation and writing—original draft; I.C.: investigation and writing—original draft; K.H.: conceptualization, investigation, project administration, supervision, writing—original draft and writing—review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the Human Frontier Science Program (grant no. HFSP-RGY0074/2019). In addition, K.F. is supported by ELSI-First Logic Astrobiology Donation Program, I.C. is supported by the Visegrad Fund Scholarship (no. 52110039) and M.M. and I.C. are supported by the project the 'Grant Schemes at CU' (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935), project no. START/SCI/148. S.D.F. thanks the NSF for a CAREER grant (MCB 2045844), supporting work on protein refoldability.

Acknowledgements. The authors wish to thank the three reviewers of this review paper for their helpful comments and suggestions.

References

1. Powner MW, Gerland B, Sutherland JD. 2009 Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242. (doi:10.1038/nature08013)

2. Becker S *et al.* 2019 Unified prebiotically plausible synthesis of pyrimidine and purine RNA

ribonucleotides. *Science* **366**, 76–82. (doi:10.1126/science.aax2747)

3. Wochner A, Attwater J, Coulson A, Holliger P. 2011 Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**, 209–212. (doi:10.1126/science.1200752)
4. Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001 RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**, 1319–1325. (doi:10.1126/science.1060786)
5. Horning DP, Joyce GF. 2016 Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl Acad. Sci. USA* **113**, 9786–9791. (doi:10.1073/pnas.1610103113)
6. Bartel DP, Szostak JW. 1993 Isolation of new ribozymes from a large pool of random sequences. *Science* **261**, 1411–1418. (doi:10.1126/science.7690155)
7. Xu J, Chmela V, Green NJJ, Russell DAA, Janicki MJJ, Góra RWW, Szabla R, Bond A, Sutherland JD. 2020 Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature* **582**, 60–66. (doi:10.1038/s41586-020-2330-9)
8. Bhowmik S, Krishnamurthy R. 2019 The role of sugar-backbone heterogeneity and chimeras in the simultaneous emergence of RNA and DNA. *Nat. Chem.* **11**, 1009–1018. (doi:10.1038/s41557-019-0322-x)
9. Miller SL. 1953 A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529. (doi:10.1126/science.117.3046.528)
10. Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL. 2011 Primordial synthesis of amines and amino acids in a 1958 Miller H_2S-rich spark discharge experiment. *Proc. Natl Acad. Sci. USA* **108**, 5526–5531. (doi:10.1073/pnas.1019191108)
11. Cronin JR, Pizzarello S. 1983 Amino acids in meteorites. *Adv. Sp. Res.* **3**, 5–18. (doi:10.1016/0273-1177(83)90036-4)
12. Glavin DP, Elsila JE, McLain HL, Aponte JC, Parker ET, Dworkin JP, Hill DH, Connolly HC, Lauretta DS. 2021 Extraterrestrial amino acids and L-enantiomeric excesses in the CM2 carbonaceous chondrites Aguas Zarcas and Murchison. *Meteorit. Planet. Sci.* **56**, 148–173. (doi:10.1111/maps.13451)
13. Forsythe JG, Yu SS, Mamajanov I, Grover MA, Krishnamurthy R, Fernández FM, Hud NV. 2015 Ester-mediated amide bond formation driven by wet-dry cycles: a possible path to polypeptides on the prebiotic Earth. *Angew Chem.* **127**, 10 009–10 013. (doi:10.1002/ange.201503792)
14. Imai EI, Honda H, Hatori K, Brack A, Matsuno K. 1999 Elongation of oligopeptides in a simulated submarine hydrothermal system. *Science* **283**, 831–833. (doi:10.1126/science.283.5403.831)
15. Furukawa Y, Otake T, Ishiguro T, Nakazawa H, Kakegawa T. 2012 Abiotic formation of valine peptides under conditions of high temperature and high pressure. *Orig. Life Evol. Biosph.* **42**, 519–531. (doi:10.1007/s11084-012-9295-0)
16. Takahagi W *et al.* 2019 Peptide synthesis under the alkaline hydrothermal conditions on Enceladus. *ACS Earth Sp. Chem.* **3**, 2559–2568. (doi:10.1021/acsearthspacechem.9b00108)
17. McGeoch MW, Dikler S, McGeoch JEM. 2020 Hemolithin: a meteoritic protein containing iron and lithium. (http://arxiv.org/abs/2002.11688)
18. Kitadai N, Maruyama S. 2018 Origins of building blocks of life: a review. *Geosci. Front.* **9**, 1117–1153. (doi:10.1016/j.gsf.2017.07.007)
19. Rode BM. 1999 Peptides and the origin of life. *Peptides* **20**, 773–786. (doi:10.1016/S0196-9781(99)00062-5)
20. Giacobelli VG, Fujishima K, Lepšík M, Tretyachenko V, Kadavá T, Bednárová L, Novák P, Hlouchová K. 2021 In vitro evolution reveals primordial RNA-protein interaction mediated by metal cations. *BioRxiv*. (doi:10.1101/2021.08.01.454623)
21. Despotović D, Longo LM, Aharon E, Kahana A, Scherf T, Gruic-Sovulj I, Tawfik DS. 2020 Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry* **59**, 4456–4462. (doi:10.1021/acs.biochem.0c00800)
22. Longo LM, Lee J, Blaber M. 2013 Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl Acad. Sci. USA* **110**, 2135–2139. (doi:10.1073/pnas.1219530110)
23. Shibue R, Sasamoto T, Shimada M, Zhang B, Yamagishi A, Akanuma S. 2018 Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, 1227. (doi:10.1038/s41598-018-19561-1)
24. Makarov M *et al.* 2021 Enzyme catalysis prior to aromatic residues: reverse engineering of a dephospho-CoA kinase. *Protein Sci.* **30**, 1022–1034. (doi:10.1002/pro.4068)
25. Yagi S, Padhi AK, Vucinic J, Barbe S, Schiex T, Nakagawa R, Simoncini D, Zhang KY, Tagami S. 2021 Seven amino acid types suffice to reconstruct the core fold of RNA polymerase. *bioRxiv*. (https://www.biorxiv.org/content/10.1101/2021.02.22.432383v1)
26. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. 1997 Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809. (doi:10.1038/nsb1097-805)
27. Longo LM, Tenorio CA, Kumru OS, Middaugh CR, Blaber M. 2015 A single aromatic core mutation converts a designed 'primitive' protein from halophile to mesophile folding. *Protein Sci.* **24**, 27–37. (doi:10.1002/pro.2580)
28. Kimura M, Akanuma S. 2020 Reconstruction and characterization of thermally stable and catalytically active proteins comprising an alphabet of ~13 amino acids. *J. Mol. Evol.* **88**, 372–381. (doi:10.1007/s00239-020-09938-0)
29. Eschenmoser A, Krishnamurthy R. 2000 Chemical etiology of nucleic acid structure. *Pure Appl. Chem.* **72**, 343–345. (doi:10.1351/pac200072030343)
30. Gardner PP, Holland BR, Moulton V, Hendy M, Penny D. 2003 Optimal alphabets for an RNA world. *Proc. R. Soc. B* **270**, 1177–1182. (doi:10.1098/rspb.2003.2355)
31. Szathmáry E. 2003 Why are there four letters in the genetic alphabet? *Nat. Rev. Genet.* **4**, 995–1001. (doi:10.1038/nrg1231)
32. Gorlero M, Wieczorek R, Adamala K, Giorgi A, Schininà ME, Stano P, Luisi PL. 2009 Ser-his catalyses the formation of peptides and PNAs. *FEBS Lett.* **583**, 153–156. (doi:10.1016/j.febslet.2008.11.052)
33. Foden CS, Islam S, Fernández-García C, Maugeri L, Sheppard TD, Powner MW. 2020 Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. *Science* **370**, 865–869. (doi:10.1126/science.abd5680)
34. Rodriguez-Garcia M, Surman AJ, Cooper GJT, Suárez-Marina I, Hosni Z, Lee MP, Cronin L. 2015 Formation of oligopeptides in high yield under simple programmable conditions. *Nat. Commun.* **6**, 8385. (doi:10.1038/ncomms9385)
35. Hill AR, Orgel LE, Wu T. 1993 The limits of template-directed synthesis with nucleoside-5′-phosphoro(2-methyl)imidazolides. *Orig. Life Evol. Biosph.* **23**, 285–290. (doi:10.1007/BF01582078)
36. Giurgiu C, Li L, O'Flaherty DK, Tam CP, Szostak JW. 2017 A mechanistic explanation for the regioselectivity of nonenzymatic RNA primer extension. *J. Am. Chem. Soc.* **139**, 16 741–16 747. (doi:10.1021/jacs.7b08784)
37. Sheng J, Li L, Engelhart AE, Gan J, Wang J, Szostak JW. 2014 Structural insights into the effects of 2′-5′ linkages on the RNA duplex. *Proc. Natl Acad. Sci. USA* **111**, 3050–3055. (doi:10.1073/pnas.1317799111)
38. Jeffares DC, Poole AM, Penny D. 1998 Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36. (doi:10.1007/PL00006280)
39. Frenkel-Pinter M, Frenkel-Pinter M, Samanta M, Ashkenasy G, Leman LJ, Leman LJ. 2020 Prebiotic peptides: molecular hubs in the origin of life. *Chem. Rev.* **120**, 4707–4765. (doi:10.1021/acs.chemrev.9b00664)
40. Lu Y, Yeung N, Sieracki N, Marshall NM. 2009 Design of functional metalloproteins. *Nature* **460**, 855–862. (doi:10.1038/nature08304)
41. Holliday GL, Thornton JM, Marquet A, Smith AG, Rébeillé F, Mendel R, Schubert HL, Lawrence AD, Warren MJ. 2007 Evolution of enzymes and pathways for the biosynthesis of cofactors. *Nat. Prod. Rep.* **24**, 972–987. (doi:10.1039/b703107f)
42. Chu XY, Zhang HY. 2020 Cofactors as molecular fossils to trace the origin and evolution of proteins. *ChemBioChem* **21**, 3161–3168. (doi:10.1002/cbic.202000027)
43. Tagami S, Attwater J, Holliger P. 2017 Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat. Chem.* **9**, 325–332. (doi:10.1038/nchem.2739)
44. Hsiao C, Mohan S, Kalahar BK, Williams LD. 2009 Peeling the onion: ribosomes are ancient molecular fossils. *Mol. Biol. Evol.* **26**, 2415–2425. (doi:10.1093/molbev/msp163)
45. Guth-Metzler R *et al.* 2020 Cutting in-line with iron: ribosomal function and non-oxidative RNA cleavage.

Nucleic Acids Res. **48**, 8663–8674. (doi:10.1093/nar/gkaa586)

46. Berens C, Streicher B, Schroeder R, Hillen W. 1998 Visualizing metal-ion-binding sites in group I introns by iron(II)-mediated Fenton reactions. *Chem. Biol.* **5**, 163–175. (doi:10.1016/S1074-5521(98)90061-8)
47. Deechongkit S, Dawson PE, Kelly JW. 2004 Toward assessing the position-dependent contributions of backbone hydrogen bonding to β-sheet folding thermodynamics employing amide-to-ester perturbations. *J. Am. Chem. Soc.* **126**, 16 762–16 771. (doi:10.1021/ja045934s)
48. Deechongkit S, Nguyen H, Powers ET, Dawson PE, Gruebele M, Kelly JW. 2004 Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics. *Nature* **430**, 101–105. (doi:10.1038/nature02611)
49. Fialho DM, Karunakaran SC, Greeson KW, Martínez I, Schuster GB, Krishnamurthy R, Hud NV. 2021 Depsipeptide nucleic acids: prebiotic formation, oligomerization, and self-assembly of a new proto-nucleic acid candidate. *J. Am. Chem. Soc.* **143**, 13 525–13 537. (doi:10.1021/jacs.1c02287)
50. Wittung P, Nielsen PE, Buchardt O, Egholm M, Nordén B. 1994 DNA-like double helix formed by peptide nucleic acid. *Nature* **368**, 561–563. (doi:10.1038/368561a0)
51. Mulkidjanian AY, Bychkov AY, Dibrova DV, Galperin MY, Koonin EV. 2012 Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl Acad. Sci. USA* **109**, E821–E830. (doi:10.1073/pnas.1117774109)
52. Kalervo Airas R. 1996 Differences in the magnesium dependences of the class I and class II aminoacyl-tRNA synthetases from Escherichia coli. *Eur. J. Biochem.* **240**, 223–231. (doi:10.1111/j.1432-1033.1996.0223h.x)
53. Bray MS, Lenz TK, Haynes JW, Bowman JC, Petrov AS, Reddi AR, Hud NV, Williams LD, Glass JB. 2018 Multiple prebiotic metals mediate translation. *Proc. Natl Acad. Sci. USA* **115**, 12 164–12 169. (doi:10.1073/pnas.1803636115)
54. Sosnick TR. 2008 Kinetic barriers and the role of topology in protein and RNA folding. *Protein Sci.* **17**, 1308–1318. (doi:10.1110/ps.036319.108)
55. Thirumalai D, Woodson SA. 1996 Kinetics of folding of proteins and RNA. *Acc. Chem. Res.* **29**, 433–439. (doi:10.1021/ar9500933)
56. Higgs PG, Pudritz RE. 2009 A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490. (doi:10.1089/ast.2008.0280)
57. Trifonov EN. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151. (doi:10.1016/S0378-1119(00)00476-5)
58. Granold M, Hajieva P, Toşa MI, Irimie FD, Moosmann B. 2018 Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl Acad. Sci. USA* **115**, 41–46. (doi:10.1073/pnas.1717100115)
59. Zaia DAM, Zaia CTBV, De Santana H. 2008 Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.* **38**, 469–488. (doi:10.1007/s11084-008-9150-5)
60. Brooks DJ, Fresco JR, Lesk AM, Singh M. 2002 Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655. (doi:10.1093/oxfordjournals.molbev.a003988)
61. Fournier GP, Andam CP, Alm EJ, Gogarten JP. 2011 Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. *Orig. Life Evol. Biosph.* **41**, 621–632. (doi:10.1007/s11084-011-9261-2)
62. Bradley LH, Thumfort PP, Hecht MH. 2006 De novo proteins from binary-patterned combinatorial libraries. *Methods Mol. Biol.* **340**, 53–69. (doi:10.1385/1-59745-116-9:53)
63. Etchebest C, Benros C, Bornot A, Camproux AC, De Brevern AG. 2007 A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur. Biophys. J.* **36**, 1059–1069. (doi:10.1007/s00249-007-0188-5)
64. Fellouse FA, Wiesmann C, Sidhu SS. 2004 Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc. Natl Acad. Sci. USA* **101**, 12 467–12 472. (doi:10.1073/pnas.0401786101)
65. Reader J, Joyce GF. 2002 A ribozyme composed of only two different nucleotides. *Nature* **420**, 841–844. (doi:10.1038/nature01185)
66. Greenwald J, Friedmann MP, Riek R. 2016 Amyloid aggregates arise from amino acid condensations under prebiotic conditions. *Angew. Chem. Int. Ed.* **55**, 11 609–11 613. (doi:10.1002/anie.201605321)
67. Georgelin T, Jaber M, Bazzi H, Lambert JF. 2013 Formation of activated biomolecules by condensation on mineral surfaces: a comparison of peptide bond formation and phosphate condensation. *Orig. Life Evol. Biosph.* **43**, 429–443. (doi:10.1007/s11084-013-9345-2)
68. Wolos A, Roszak R, Zadlo-Dobrowolska A, Beker W, Mikulak-Klucznik B, Spólnik G, Dygas M, Szymkuć S, Grzybowski BA. 2020 Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, aaw1955. (doi:10.1126/science.aaw1955)
69. Kim SC, O'Flaherty DK, Giurgiu C, Zhou L, Szostak JW. 2021 The Emergence of RNA from the heterogeneous products of prebiotic nucleotide synthesis. *J. Am. Chem. Soc.* **143**, 3267–3279. (doi:10.1021/jacs.0c12955)
70. Weber AL, Pizzarello S. 2006 The peptide-catalyzed stereospecific synthesis of tetroses: a possible model for prebiotic molecular evolution. *Proc. Natl Acad. Sci. USA* **103**, 12 713–12 717. (doi:10.1073/pnas.0602320103)
71. Pearce BKD, Tupper AS, Pudritz RE, Higgs PG. 2018 Constraining the time interval for the origin of life on Earth. *Astrobiology* **18**, 343–364. (doi:10.1089/ast.2017.1674)
72. Amit M, Cheng G, Hamley IW, Ashkenasy N. 2012 Conductance of amyloid β based peptide filaments: structure-function relations. *Soft Matter* **8**, 8690–8696. (doi:10.1039/c2sm26017d)
73. Mehta AK *et al.* 2008 Facial symmetry in protein self-assembly. *J. Am. Chem. Soc.* **130**, 9829–9835. (doi:10.1021/ja801511n)
74. Maury CPJ. 2009 Self-propagating β-sheet polypeptide structures as prebiotic informational molecular entities: the amyloid world. *Orig. Life Evol. Biosph.* **39**, 141–150. (doi:10.1007/s11084-009-9165-6)
75. Carny O, Gazit E. 2005 A model for the role of short self-assembled peptides in the very early stages of the origin of life. *FASEB J.* **19**, 1051–1055. (doi:10.1096/fj.04-3256hyp)
76. Hordijk W. 2017 Autocatalytic confusion clarified. *J. Theor. Biol.* **435**, 22–28. (doi:10.1016/j.jtbi.2017.09.003)
77. Lee DH, Granja JR, Martinez JA, Severin K, Ghadiri MR. 1996 A self-replicating peptide. *Nature* **382**, 525–528. (doi:10.1038/382525a0)
78. Rout SK, Friedmann MP, Riek R, Greenwald J. 2018 A prebiotic template-directed peptide synthesis based on amyloids. *Nat. Commun.* **9**, 234. (doi:10.1038/s41467-017-02742-3)
79. Takahashi Y, Mihara H. 2004 Construction of a chemically and conformationally self-replicating system of amyloid-like fibrils. *Bioorg. Med. Chem.* **12**, 693–699. (doi:10.1016/j.bmc.2003.11.022)
80. Lee DH, Severin K, Yokobayashi Y, Ghadiri MR. 1997 Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Nature* **390**, 591–594. (doi:10.1038/37569)
81. Rufo CM, Moroz YS, Moroz OV, Stöhr J, Smith TA, Hu X, Degrado WF, Korendovych IV. 2014 Short peptides self-assemble to produce catalytic amyloids. *Nat. Chem.* **6**, 303–309. (doi:10.1038/nchem.1894)
82. Zhang C *et al.* 2014 Self-assembled peptide nanofibers designed as biological enzymes for catalyzing ester hydrolysis. *ACS Nano* **8**, 11 715–11 723. (doi:10.1021/nn5051344)
83. Omosun TO *et al.* 2017 Catalytic diversity in self-propagating peptide assemblies. *Nat. Chem.* **9**, 805–809. (doi:10.1038/nchem.2738)
84. Bonfio C, Godino E, Corsini M, Fabrizi de Biani F, Guella G, Mansy SS. 2018 Prebiotic iron–sulfur peptide catalysts generate a pH gradient across model membranes of late protocells. *Nat. Catal.* **1**, 616–623. (doi:10.1038/s41929-018-0116-3)
85. Söding J, Lupas AN. 2003 More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837–846. (doi:10.1002/bies.10321)
86. Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. 2021 Bridging themes: short protein segments found in different architectures. *Mol. Biol. Evol.* **38**, 2191–2208. (doi:10.1093/molbev/msab017)
87. Chaudhuri I, Söding J, Lupas AN. 2008 Evolution of the β-propeller fold. *Proteins Struct. Funct. Genet.* **71**, 795–803. (doi:10.1002/prot.21764)
88. Alva V, Söding J, Lupas AN. 2015 A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410. (doi:10.7554/eLife.09410)

89. Longo L, Jabłońska J, Vyas P, Kanade M, Kolodny R, Ben-Tal N, Tawfik DS. 2020 On the emergence of P-Loop NTPase and Rossmann enzymes from a beta-alpha-beta ancestral fragment. *eLife* **9**, e64415. (doi:10.7554/eLife.64415)
90. Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD. 2020 Root of the tree: the significance, evolution, and origins of the ribosome. *Chem. Rev.* **120**, 4848–4878. (doi:10.1021/acs.chemrev.9b00742)
91. Lupas AN, Alva V. 2017 Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol.* **198**, 74–81. (doi:10.1016/j.jsb.2017.04.007)
92. Fox GE. 2010 Origin and evolution of the ribosome. *Cold Spring Harb. Perspect Biol.* **2**, a003483. (doi:10.1101/cshperspect.a003483)
93. Belousoff MJ *et al.* 2010 Ancient machinery embedded in the contemporary ribosome. *Biochem. Soc. Trans.* **38**, 422–427. (doi:10.1042/BST0380422)
94. Petrov AS *et al.* 2015 History of the ribosome and the origin of translation. *Proc. Natl Acad. Sci. USA* **112**, 15 396–15 401. (doi:10.1073/pnas.1509761112)
95. Kovacs NA, Petrov AS, Lanier KA, Williams LD. 2017 Frozen in time: the history of proteins. *Mol. Biol. Evol.* **34**, 1252–1260. (doi:10.1093/molbev/msx086)
96. Li P, Holliger P, Tagami S. 2021 Hydrophobic-cationic peptides enhance RNA polymerase ribozyme activity by accretion. *bioRxiv*. (doi:10.1101/2021.02.22.432394)
97. Dale T. 2006 Protein and nucleic acid together: a mechanism for the emergence of biological selection. *J. Theor. Biol.* **240**, 337–342. (doi:10.1016/j.jtbi.2005.09.027)
98. Lahav N. 1993 The RNA-world and co-evolution hypotheses and the origin of life: implications, research strategies and perspectives. *Orig. Life Evol. Biosph.* **23**, 329–344. (doi:10.1007/BF01582084)
99. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS. 2016 De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell* **164**, 476–486. (doi:10.1016/j.cell.2015.12.024)
100. Grishin NV. 2001 KH domain: one motif, two folds. *Nucleic Acids Res.* **29**, 638–643. (doi:10.1093/nar/29.3.638)
101. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS. 2020 Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl Acad. Sci. USA* **117**, 15 731–15 739. (doi:10.1073/pnas.2001989117)
102. Rebeaud ME, Mallik S, Goloubinoff P, Tawfik DS. 2021 On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. *Proc. Natl Acad. Sci. USA* **118**, e2020885188. (doi:10.1073/pnas.2020885118)
103. De Los Rios P, Ben-Zvi A, Slutsky O, Azem A, Goloubinoff P. 2006 Hsp70 chaperones accelerate protein translocation and the unfolding of stable protein aggregates by entropic pulling. *Proc. Natl Acad. Sci. USA* **103**, 6166–6171. (doi:10.1073/pnas.0510496103)
104. Tretyachenko V, Vymětal J, Neuwirthová T, Vondrášek J, Fujishima K, Hlouchová K. 2021 Structured proteins are abundant in unevolved sequence space. *bioRxiv*. (doi:10.1101/2021.08.29.458031)
105. Akanuma S, Kigawa T, Yokoyama S. 2002 Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc. Natl Acad. Sci. USA* **99**, 13 549–13 553. (doi:10.1073/pnas.222243999)
106. Cleaves HJ. 2010 The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498. (doi:10.1016/j.jtbi.2009.12.014)
107. Ilardo M, Meringer M, Freeland S, Rasulev B, Cleaves HJ. 2015 Extraordinarily adaptive properties of the genetically encoded amino acids. *Sci. Rep.* **5**, 9414. (doi:10.1038/srep09414)
108. Ilardo M *et al.* 2019 Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. *Sci. Rep.* **9**, 12468. (doi:10.1038/s41598-019-47574-x)
109. Philip GK, Freeland SJ. 2011 Did evolution select a nonrandom 'alphabet' of amino acids? *Astrobiology* **11**, 235–240. (doi:10.1089/ast.2010.0567)
110. Weber AL, Miller SL. 1981 Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273–284. (doi:10.1007/BF01795749)
111. Liu CC, Schultz PG. 2010 Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79**, 413–444. (doi:10.1146/annurev.biochem.052308.105824)
112. Passioura T, Suga H. 2014 Reprogramming the genetic code in vitro. *Trends Biochem. Sci.* **39**, 400–408. (doi:10.1016/j.tibs.2014.07.005)
113. Mayer-Bacon C, Agboha N, Muscalli M, Freeland S. 2021 Evolution as a guide to designing xeno amino acid alphabets. *Int. J. Mol. Sci.* **22**, 1–13. (doi:10.3390/ijms22062787)
114. Burton AS, Stern JC, Elsila JE, Glavin DP, Dworkin JP. 2012 Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev.* **41**, 5459–5472. (doi:10.1039/c2cs35109a)
115. Glavin DP, Aubrey AD, Callahan MP, Dworkin JP, Elsila JE, Parker ET, Bada JL, Jenniskens P, Shaddad MH. 2010 Extraterrestrial amino acids in the Almahata Sitta meteorite. *Meteorit. Planet Sci.* **45**, 1695–1709. (doi:10.1111/j.1945-5100.2010.01094.x)
116. Pizzarello S, Schrader DL, Monroe AA, Lauretta DS. 2012 Large enantiomeric excesses in primitive meteorites and the diverse effects of water in cosmochemical evolution. *Proc. Natl Acad. Sci. USA* **109**, 11 949–11 954. (doi:10.1073/pnas.1204865109)
117. Johnson AP, Cleaves HJ, Dworkin JP, Glavin DP, Lazcano A, Bada JL. 2008 The Miller volcanic spark discharge experiment. *Science* **322**, 404. (doi:10.1126/science.1161527)
118. Alvarez-Carreño C, Becerra A, Lazcano A. 2013 Norvaline and norleucine may have been more abundant protein components during early stages of cell evolution. *Orig. Life Evol. Biosph.* **43**, 363–375. (doi:10.1007/s11084-013-9344-3)
119. Mascarenhas AP, An S, Rosen AE, Martinis SA, Musier-Forsyth K. 2009 Fidelity mechanisms of the aminoacyl-tRNA synthetases. *Protein Eng.* 155–203. (doi:10.1007/978-3-540-70941-1_6)
120. Bredehöft J, Thiemann W, Jessberger E, Carob G, Meierhenrich U. 2004 Identification of diamino acids in the Murichison meteorite. *Proc. Natl Acad. Sci. USA* **25**, 9182–9186. (doi:10.1073/pnas.0403043101)
121. Blanco C, Bayas M, Yan F, Chen IA. 2018 Analysis of evolutionarily independent protein-RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Curr. Biol.* **28**, 526–537.e5. (doi:10.1016/j.cub.2018.01.014)
122. Jukes TH. 1973 Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophys. Res. Commun.* **53**, 709–714. (doi:10.1016/0006-291X(73)90151-4)
123. Vázquez-Salazar A, Becerra A, Lazcano A. 2018 Evolutionary convergence in the biosyntheses of the imidazole moieties of histidine and purines. *PLoS ONE* **13**, e0196349. (doi:10.1371/journal.pone.0196349)
124. Raggi L, Bada JL, Lazcano A. 2016 On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Phys. Chem. Chem. Phys.* **18**, 20 028–20 032. (doi:10.1039/C6CP00793G)
125. Cheng RP, Gellman SH, DeGrado WF. 2001 β-peptides: from structure to function. *Chem. Rev.* **101**, 3219–3232. (doi:10.1021/cr000045i)
126. Chandru K, Guttenberg N, Giri C, Hongo Y, Butch C, Mamajanov I, Cleaves HJ. 2018 Simple prebiotic synthesis of high diversity dynamic combinatorial polyester libraries. *Commun. Chem.* **1**, 30. (doi:10.1038/s42004-018-0031-1)
127. Tian YF, Hudalla GA, Han H, Collier JH. 2013 Controllably degradable β-sheet nanofibers and gels from self-assembling depsipeptides. *Biomater. Sci.* **1**, 1037–1045. (doi:10.1039/c3bm60161g)
128. Fahnestock S, Rich A. 1971 Ribosome-catalyzed polyester formation. *Science* **173**, 340–343. (doi:10.1126/science.173.3994.340)
129. Ohta A, Murakami H, Hiroaki S. 2008 Polymerization of α-hydroxy acids by ribosomes. *ChemBioChem* **9**, 2773–2778. (doi:10.1002/cbic.200800439)
130. Trifonov EN. 2004 The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22**, 1–11. (doi:10.1080/07391102.2004.10506975)
131. Moosmann B, Schindeldecker M, Hajieva P. 2020 Cysteine, glutathione and a new genetic code: biochemical adaptations of the primordial cells that spread into open water and survived biospheric oxygenation. *Biol. Chem.* **401**, 213–231. (doi:10.1515/hsz-2019-0232)
132. Tze-Fei Wong J. 2005 Coevolution theory of the genetic code at age thirty. *Bioessays* **27**, 416–425. (doi:10.1002/bies.20208)

133. Fujishima K, Wang KM, Palmer JA, Abe N, Nakahigashi K, Endy D, Rothschild LJ. 2018 Reconstruction of cysteine biosynthesis using engineered cysteine-free enzymes. *Sci. Rep.* **8**, 1776. (doi:10.1038/s41598-018-19920-y)
134. Eck RV, Dayhoff MO. 1966 Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152**, 363–366. (doi:10.1126/science.152.3720.363)
135. Holliday GL, Fischer JD, Mitchell JBO, Thornton JM. 2011 Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.* **278**, 3835–3845. (doi:10.1111/j.1742-4658.2011.08190.x)
136. Shen C, Mills T, Oró J. 1990 Prebiotic synthesis of histidyl-histidine. *J. Mol. Evol.* **31**, 175–179. (doi:10.1007/BF02109493)
137. White SH. 1994 The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. *J. Mol. Evol.* **38**, 383–394. (doi:10.1007/BF00163155)
138. Woese C. 1998 The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859. (doi:10.1073/pnas.95.12.6854)
139. Tong CL, Lee KH, Seelig B. 2021 De novo proteins from random sequences through in vitro evolution. *Curr. Opin. Struct. Biol.* **68**, 129–134. (doi:10.1016/j.sbi.2020.12.014)
140. Tretyachenko V *et al.* 2017 Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **7**, 15449. (doi:10.1038/s41598-017-15635-8)
141. De Lucrezia D, Franchi M, Chiarabelli C, Gallori E, Luisi PL. 2006 Investigation of de novo totally random biosequences. Part IV. Folding properties of de novo, totally random RNAs. *Chem. Biodivers.* **3**, 869–877. (doi:10.1002/cbdv.200690090)
142. Davidson AR, Sauer RT. 1994 Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA* **91**, 2146–2150. (doi:10.1073/pnas.91.6.2146)
143. LaBean TH, Butt TR, Kauffman SA, Schultes EA. 2011 Protein folding absent selection. *Genes (Basel)* **2**, 608–626. (doi:10.3390/genes2030608)
144. Keefe AD, Szostak JW. 2001 Functional proteins from a random-sequence library. *Nature* **410**, 715–718. (doi:10.1038/35070613)
145. Chao FA *et al.* 2013 Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. *Nat. Chem. Biol.* **9**, 81–83. (doi:10.1038/nchembio.1138)
146. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH. 2011 De novo designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. *PLoS ONE* **6**, e15364. (doi:10.1371/journal.pone.0015364)
147. Tanaka J, Doi N, Takashima H, Yanagawa H. 2010 Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* **19**, 786–795. (doi:10.1002/pro.358)
148. Newton MS, Arcus VL, Gerth ML, Patrick WM. 2018 Enzyme evolution: innovation is easy, optimization is complicated. *Curr. Opin Struct. Biol.* **48**, 110–116. (doi:10.1016/j.sbi.2017.11.007)
149. Solis AD. 2019 Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. *BMC Evol. Biol.* **19**, 1–19. (doi:10.1186/s12862-019-1464-6)
150. Kozlowski LP. 2017 Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116. (doi:10.1093/nar/gkw978)
151. Anbar A. 2008 Elements and evolution. *Science* **332**, 1481–1483. (doi:10.1126/science.1163100)
152. Hazen RM, Sverjensky DA. 2010 Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harb. Perspect. Biol.* **2**, a002162. (doi:10.1101/cshperspect.a002162)
153. Nitschke W, McGlynn SE, Milner-White EJ, Russell MJ. 2013 On the antiquity of metalloenzymes and their substrates in bioenergetics. *Biochim. Biophys. Acta* **1827**, 871–881. (doi:10.1016/j.bbabio.2013.02.008)
154. Reinhard CT, Planavsky NJ, Robbins LJ, Partin CA, Gill BC, Lalonde SV, Bekker A, Konhauser KO, Lyons TW. 2013 Proterozoic ocean redox and biogeochemical stasis. *Proc. Natl Acad. Sci. USA* **110**, 5357–5362. (doi:10.1073/pnas.1208622110)
155. Garcia AK, McShea H, Kolaczkowski B, Kaçar B. 2020 Reconstructing the evolutionary history of nitrogenases: evidence for ancestral molybdenum-cofactor utilization. *Geobiology* **18**, 394–411. (doi:10.1111/gbi.12381)
156. Kacar B, Garcia AK, Anbar AD. 2021 Evolutionary history of bioessential elements can guide the search for life in the universe. *ChemBioChem* **22**, 114–119. (doi:10.1002/cbic.202000500)
157. Jones C, Nomosatryo S, Crowe SA, Bjerrum CJ, Canfield DE. 2015 Iron oxides, divalent cations, silica, and the early earth phosphorus crisis. *Geology* **43**, 135–138. (doi:10.1130/G36044.1)
158. Andreini C, Cavallaro G, Lorenzini S, Rosato A. 2013 MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **41**, D459–D464. (doi:10.1093/nar/gks1063)
159. van der Gulik P, Massar S, Gilis D, Buhrman H, Rooman M. 2009 The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* **261**, 531–539. (doi:10.1016/j.jtbi.2009.09.004)
160. Rozov A, Khusainov I, El Omari K, Duman R, Mykhaylyk V, Yusupov M, Westhof E, Wagner A, Yusupova G. 2019 Importance of potassium ions for ribosome structure and function revealed by long-wavelength X-ray diffraction. *Nat. Commun.* **10**, 2519. (doi:10.1038/s41467-019-10409-4)
161. Watson JD, Milner-White EJ. 2002 A novel main-chain anion-binding site in proteins: the nest. A particular combination of φ,ψ values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. *J. Mol. Biol.* **315**, 171–182. (doi:10.1006/jmbi.2001.5227)
162. Bianchi A, Giorgi C, Ruzza P, Toniolo C, Milner-White EJ. 2012 A synthetic hexapeptide designed to resemble a proteinaceous p-loop nest is shown to bind inorganic phosphate. *Proteins Struct. Funct. Bioinform.* **80**, 1418–1424. (doi:10.1002/prot.24038)
163. James Milner-White E. 2019 Protein three-dimensional structures at the origin of life. *Interface Focus* **9**, 20190057. (doi:10.1098/rsfs.2019.0057)
164. Tajima M, Urabe I, Yutani K, Okada H. 1976 Role of calcium ions in the thermostability of thermolysin and bacillus subtilis var. amylosaccharicticus neutral protease. *Eur. J. Biochem.* **64**, 243–247. (doi:10.1111/j.1432-1033.1976.tb10293.x)
165. Jaiswal JK. 2001 Calcium: how and why? *J. Biosci.* **26**, 357–363. (doi:10.1007/BF02703745)
166. Kazmierczak J, Kempe S, Kremer B. 2013 Calcium in the Early evolution of living systems: a biohistorical approach. *Curr. Org. Chem.* **17**, 1738–1750. (doi:10.2174/13852728113179990081)
167. Bonfio C *et al.* 2018 UV light-driven prebiotic synthesis of iron-sulfur clusters. *Nat. Chem.* **9**, 1229–1234. (doi:10.1038/nchem.2817)
168. Muchowska KB, Varma SJ, Moran J. 2020 Nonenzymatic metabolic reactions and life's origins. *Chem. Rev.* **120**, 7708–7744. (doi:10.1021/acs.chemrev.0c00191)
169. Preiner M *et al.* 2020 A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism. *Nat. Ecol. Evol.* **4**, 534–542. (doi:10.1038/s41559-020-1125-6)
170. Muchowska KB, Varma SJ, Moran J. 2019 Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107. (doi:10.1038/s41586-019-1151-1)
171. Kitadai N, Nakamura R, Yamamoto M, Takai K, Yoshida N, Oono Y. 2019 Metals likely promoted protometabolism in early ocean alkaline hydrothermal systems. *Sci. Adv.* **5**, eaav7848. (doi:10.1126/sciadv.aav7848)
172. Kitadai N, Kameya M, Fujishima K. 2017 Origin of the reductive tricarboxylic acid (RTCA) cycle-type CO_2 fixation: a perspective. *Life* **7**, 39. (doi:10.3390/life7040039)
173. Ji HF, Kong DX, Shen L, Chen LL, Ma BG, Zhang HY. 2007 Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **8**, R176. (doi:10.1186/gb-2007-8-8-r176)
174. Denessiouk KA, Rantanen VV, Johnson MS. 2001 Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins Struct. Funct. Genet.* **44**, 282–291. (doi:10.1002/prot.1093)
175. Wächtershäuser G. 1990 The case for the chemoautotrophic origin of life in an iron-sulfur world. *Orig. Life Evol. Biosph.* **20**, 173–176. (doi:10.1007/BF01808279)
176. White III HB. 1976 Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104. (doi:10.1007/BF01732468)

177. Gilbert W. 1986 Origin of life: the RNA World. *Nature* **319**, 618. (doi:10.1038/319618a0)
178. Kim HS, Mittenthal JE, Caetano-Anollés G. 2006 MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinf.* **7**, 351. (doi:10.1186/1471-2105-7-351)
179. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G. 2007 Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585. (doi:10.1101/gr.6454307)
180. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. 2009 The origin, evolution and structure of the protein world. *Biochem. J.* **417**, 621–637. (doi:10.1042/BJ20082063)
181. Goldman AD, Samudrala R, Baross JA. 2010 The evolution and functional repertoire of translation proteins following the origin of life. *Biol. Direct* **5**, 15. (doi:10.1186/1745-6150-5-15)
182. To P, Whitehead B, Tarbox HE, Fried SD. 2021 Non-refoldability is pervasive across the *E. coli* proteome. *J. Am. Chem. Soc.* **143**, 11 435–11 448. (doi:10.1021/jacs.1c03270)
183. Kerner MJ *et al.* 2005 Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell* **122**, 209–220. (doi:10.1016/j.cell.2005.05.028)
184. Romero Romero ML *et al.* 2018 Simple yet functional phosphate-loop proteins. *Proc. Natl Acad. Sci. USA* **115**, E11943–E11950. (doi:10.1073/pnas.1812400115)
185. Petrov AS *et al.* 2014 Evolution of the ribosome at atomic resolution. *Proc. Natl Acad. Sci. USA* **111**, 10 251–10 256. (doi:10.1073/pnas.1407205111)
186. Thirumalai D, Lorimer GH, Hyeon C. 2020 Iterative annealing mechanism explains the functions of the GroEL and RNA chaperones. *Protein Sci.* **29**, 360–377. (doi:10.1002/pro.3795)
187. Thirumalai D, Lorimer GH. 2001 Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 245–269. (doi:10.1146/annurev.biophys.30.1.245)
188. To P, Lee SO, Xia Y, Devlin T, Fleming KG, Fried SD. 2021 Systematic interrogation of protein refolding under cellular-like conditions. *bioRxiv*. (doi:10.1101/2021.11.20.469408)
189. Evans MS, Sander IM, Clark PL. 2008 Cotranslational folding promotes β-helix formation and avoids aggregation in vivo. *J. Mol. Biol.* **383**, 683–692. (doi:10.1016/j.jmb.2008.07.035)
190. Thanaraj TA, Argos P. 1996 Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594–1612. (doi:10.1002/pro.5560050814)
191. Liu K, Maciuba K, Kaiser CM. 2019 The ribosome cooperates with a chaperone to guide multi-domain protein folding. *Mol. Cell* **74**, 310–319.e7. (doi:10.1016/j.molcel.2019.01.043)
192. Ma H, Proctor DJ, Kierzek E, Kierzek R, Bevilacqua PC, Gruebele M. 2006 Exploring the energy landscape of a small RNA hairpin. *J. Am. Chem. Soc.* **128**, 1523–1530. (doi:10.1021/ja0553856)
193. Ditzler MA, Rueda D, Mo J, Håkansson K, Walter NG. 2008 A rugged free energy landscape separates multiple functional RNA folds throughout denaturation. *Nucleic Acids Res.* **36**, 7088–7099. (doi:10.1093/nar/gkn871)
194. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995 Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins Struct. Funct. Bioinform.* **21**, 167–195. (doi:10.1002/prot.340210302)
195. Dill KA, Chan HS. 1997 From levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19. (doi:10.1038/nsb0197-10)
196. Proctor JR, Meyer IM. 2013 CoFold: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.* **41**, e102. (doi:10.1093/nar/gkt174)
197. Hua B, Panja S, Woodson S, Ha T. 2018 Mimicking co-transcriptional RNA folding using a superhelicase. *Biophys. J.* **114**, 433a–434a. (doi:10.1016/j.bpj.2017.11.2401)
198. Woodson SA. 2010 Taming free energy landscapes with RNA chaperones. *RNA Biol.* **7**, 677–686. (doi:10.4161/rna.7.6.13615)
199. Delaye L, Becerra A, Lazcano A. 2004 The nature of the last common ancestor. In *The genetic code and the origin of life*, pp. 34–47. Boston, MA: Springer. (doi:10.1007/0-387-26887-1_3)
200. Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020 Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138–147. (doi:10.1038/s41559-019-1040-x)
201. Nasir A, Mughal F, Caetano-Anollés G. 2021 The tree of life describes a tripartite cellular world. *Bioessays* **43**, 2000343. (doi:10.1002/bies.202000343)