Journal of Molecular Evolution
https://doi.org/10.1007/s00239-022-10061-5

ORIGINAL ARTICLE

A Closer Look at Non-random Patterns Within Chemistry Space for a Smaller, Earlier Amino Acid Alphabet

Christopher Mayer-Bacon[1] · Markus Meringer[2] · Riley Havel[3] · José C. Aponte[4,5] · Stephen Freeland[1]

Received: 10 December 2021 / Accepted: 11 May 2022

Abstract

Recent findings, in vitro and in silico, are strengthening the idea of a simpler, earlier stage of genetically encoded proteins which used amino acids produced by prebiotic chemistry. These findings motivate a re-examination of prior work which has identified unusual properties of the set of twenty amino acids found within the full genetic code, while leaving it unclear whether similar patterns also characterize the subset of prebiotically plausible amino acids. We have suggested previously that this ambiguity may result from the low number of amino acids recognized by the definition of prebiotic plausibility used for the analysis. Here, we test this hypothesis using significantly updated data for organic material detected within meteorites, which contain several coded and non-coded amino acids absent from prior studies. In addition to confirming the well-established idea that "late" arriving amino acids expanded the chemistry space encoded by genetic material, we find that a prebiotically plausible subset of coded amino acids generally emulates the patterns found in the full set of 20, namely an exceptionally broad and even distribution of volumes and an exceptionally even distribution of hydrophobicities (quantified as logP) over a narrow range. However, the strength of this pattern varies depending on both the size and composition the library used to create a background (null model) for a random alphabet, and the precise definition of exactly which amino acids were present in a simpler, earlier code. Findings support the idea that a small sample size of amino acids caused previous ambiguous results, and further improvements in meteorite analysis, and/or prebiotic simulations will further clarify the nature and extent of unusual properties. We discuss the case of sulfur-containing amino acids as a specific and clear example and conclude by reviewing the potential impact of better understanding the chemical "logic" of a smaller forerunner to the standard amino acid alphabet.

Keywords Amino acids · Computational chemistry · Evolution · Abiogenesis Chemistry space

Introduction

In a foundational step of molecular evolution, life on Earth established a standard alphabet of twenty amino acids with which to construct genetically encoded proteins. This alphabet appears to have become finalized around the time of LUCA (Fournier and Alm 2015), ~4 billion years ago (Weiss et al. 2018) within a genome of similar complexity to many modern bacteria (Tuller et al. 2010). However, multiple analyses from diverse disciplinary perspectives have converged upon the idea that an earlier stage of life's evolution used a simpler (smaller) amino acid alphabet (Fig. 1). In particular, around half of the twenty encoded amino acids encoded post-LUCA are plausible, almost inevitable, products of prebiotic chemistry. Thus, while LUCA could potentially have evolved to overwrite any signature of this earlier code, it appears not to have done so (Wong and Bronskill

Handling editor: **Andrew Ellington**.

✉ Christopher Mayer-Bacon
cmayerb1@umbc.edu

[1] Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD, USA

[2] Earth Observation Center (EOC), German Aerospace Center (DLR), Oberpfaffenhofen–Wessling, Germany

[3] Department of Physics, University of Central Florida, Orlando, FL 32816, USA

[4] Solar System Exploration Division, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

[5] Department of Physics, Catholic University of America, Washington, DC 20064, USA

Published online: 06 June 2022

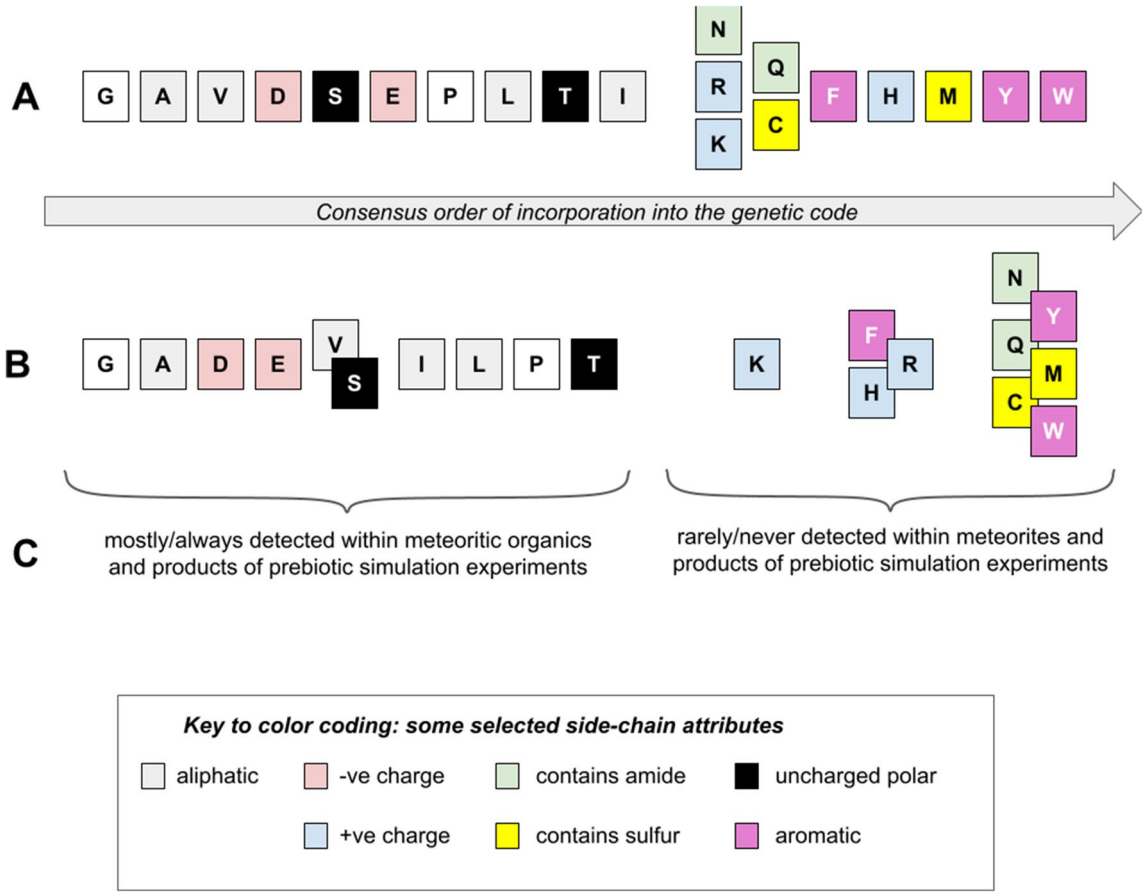


Fig. 1 A comparison of three major syntheses of the scientific literature concerning the antiquity of amino acids within the standard genetic code which motivate ideas for a simpler, earlier genetic code. **A** Trifonov (2000) consensus chronology derived from the conclusions of 40 peer-reviewed analyses of genetic code evolution. **B** Higgs and Pudritz (2009) chronology, considering peer-reviewed literature on synthesis in meteorites, icy grains, atmospheric environments, hydrothermal environments, and other abiotic synthetic routes. **C** Cleaves (2010) review of prebiotic plausibility from 3 perspectives on abiotic synthesis: meteorites, spark tube experiments, and HCN polymerization

1979). In this sense, the amino acid alphabet joins other examples of "molecular fossils" (White 1976; Benner et al. 1989) that suggest a footprint of truly ancient evolutionary history within modern metabolism. Indeed, amino acids represent arguably the single most direct chemical link known between post-LUCA molecular biology and prebiotic chemistry (Fig. 1).

Set against this background, prior literature has identified, with increasing rigor and clarity, some simple, quantitative features that distinguish the post-LUCA set of 20 amino acids from plausible alternatives in terms of fundamental physicochemical properties, namely volume and hydrophobicity (logP). For example, only one in 10^5 amino acid alphabets of size 20 drawn at random from a carefully defined library of ~ 2000 isomers and near-isomers produces a broader range of volumes with such an even distribution (see Fig. 2A and "Methods"). The pattern is approximately one order of magnitude weaker for logP (3×10^{-4} for capped amino acids: see "Methods"), but the chance of a random alphabet of size 20 surpassing the standard alphabet in both criteria simultaneously (i.e., more evenly distributed over a broader range for both logP and volume) is small enough that, of 10 million random alphabets tested, none met these criteria (Fig. 2B).

This degree of non-randomness seems remarkable given the simplicity of the statistical calculation for physicochemical properties that have long been recognized as drivers for protein structure (folding) (e.g., Grantham 1974). If the pattern detected represents a footprint of ancient natural selection, for example, then it provides a useful foundation for developing theory with which synthetic biology might design a xeno amino acid alphabet (Mayer-Bacon et al. 2021). Even without any assumption about selection, just four dimensions (range and evenness in molecular volume and logP) distinguish, from most alternatives, an alphabet with the potential to build protein-based metabolism

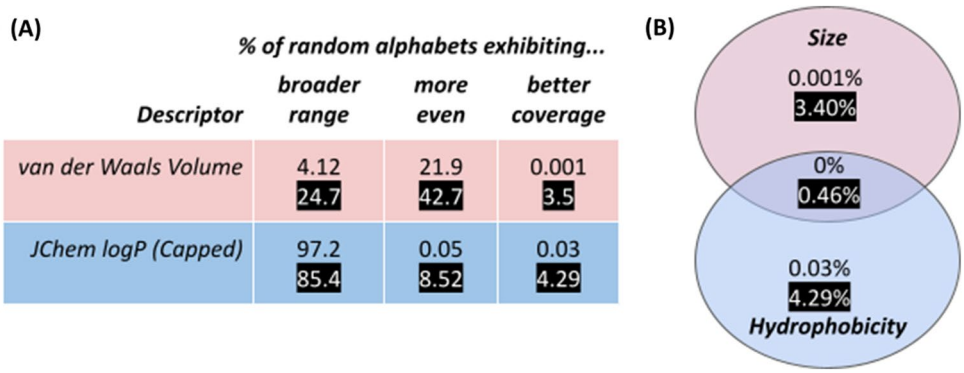
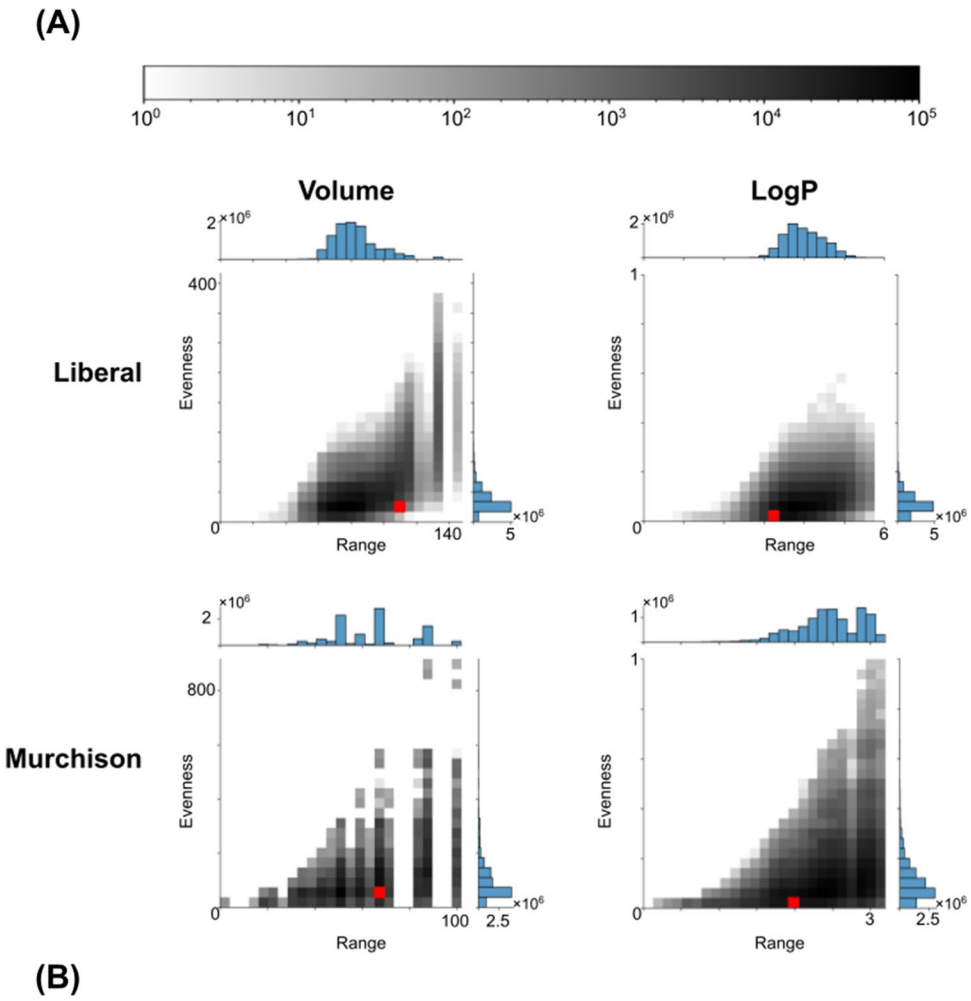


Fig. 2 Comparison of the genetically coded amino acid alphabet against random alphabets. Data adapted from Mayer-Bacon and Freeland (2021) **A** Percentage of random amino acid alphabets exhibiting broader range, more even distribution within that range, or both (relative to a set of coded amino acids), in descriptors for molecular volume and hydrophobicity. **B**: Percentage of random amino acid alphabets exhibiting both better range and more even distribution than a set of coded amino acids for either volume, hydrophobicity, or both. In **A** and **B**, numbers in white text on a black background are for random sets of 8 meteoritic α-amino acids (from a library of 44 α-amino acids) compared to 8 coded amino acids identified in meteorites. Numbers in black text are for random sets of 20 (from a library of 1949) compared to the full genetically coded amino acid alphabet

throughout life's diversification. Given amino acids' facile prebiotic synthesis under a broad range of prebiotic conditions and their near ubiquity in extra-terrestrial samples (Parker et al. 2011; Koga and Naraoka 2017), this would seem a promising direction with which to further develop thinking about independent origin(s) of life and the biosignatures they might imply.

Initial analyses (Philip and Freeland 2011) found qualitatively similar results whether testing the full, standard amino acid alphabet or a subset of these 20 that were plausibly available to an earlier stage of genetic coding. Specifically, they reported a very low number of random amino acid alphabets exhibiting better coverage (a broader range of values and more even distribution within that range) than the coded alphabet for one or more properties of size, charge, or hydrophobicity. However, recent efforts to explore the phenomenon in greater detail have questioned the second (prebiotic) finding: "*Testing eight genetically encoded amino acids which appear routinely in prebiotic simulation experiments and meteorite analyses against the collection of α-amino acids also found there … produces a much less clear picture than previously reported. The chance that a random set of 8 amino acids would achieve better coverage in van der Waals volume … and JChem logP … is orders of magnitude less extreme than the analogous test of the twenty coded amino acids"* (see Fig. 2, white font values). Visual inspection of the underlying distributions of range and evenness for size and hydrophobicity of random alphabets shows that the difference is real (see Fig. 3): whereas range and evenness of the full alphabet lie at the extreme tail of a 2-dimensional distribution, equivalent measures for the putative early alphabet lie firmly within their corresponding distributions.

Notably, the more recent analysis which challenges unusual patterns within the "early" amino acids (Mayer Bacon and Freeland 2021) used improvements to the quality of both amino acid data (e.g., removal of duplicate two-dimensional amino acid structures from consideration) and descriptor calculation (e.g., a consensus estimation of logP). These methodological improvements did not, however, refine the underlying model for prebiotic plausibility for amino acids, which has remained unchanged within this line of analysis for almost 15 years (Lu and Freeland 2008). The model in question was derived from then-current analysis of the Murchison meteorite, which is approximately as old as our planet (~4.5 billion years) and was subject to solar system astrochemistry, including extensive aqueous alteration and organic synthesis, until it fell to Earth in 1969. The Murchison meteorite has long been regarded as providing "*an invaluable sample for the direct analysis of abiotic chemical evolution prior to the onset of life*" (Pizzarello 2007) and has been used widely "*as the standard reference for organic compounds in extraterrestrial material*" (Cooper et al. 2001), retaining this interpretation to the present day (e.g., Aponte et al. 2020). However, Murchison has been re-analyzed repeatedly since 2008, as have other relevant meteorites (Elsila et al. 2016). Improvements to instrumentation and experimental protocols have detected an ever-increasing diversity of organics (e.g., Johnson et al. 2008; Aponte et al. 2020). Given the small total number of α-amino acids detected within Murchison ($N = 44$ plausible prebiotic structures for all previous tests), even small variations in this dataset are likely to change and perhaps clarify the ambiguity over patterns within the coded subset (Philip and Freeland 2011; Mayer-Bacon and Freeland 2021).



	Volume		LogP	
	$\hat{P}_{(range)}$	$\hat{P}_{(evenness)}$	$\hat{P}_{(range)}$	$\hat{P}_{(evenness)}$
Liberal	0.041	0.219	0.961	0.003
Murchison	0.247	0.427	0.859	0.085

Fig. 3 Distribution of random amino acid alphabets according to their range and evenness in van der Waals volume and logP for the "liberal collection" of 1949 plausible alternatives and the 44 amino acids detected within the Murchison meteorite. **A** Joint histograms showing a heatmap of the density of alphabets for a given range and evenness; the red box in each heatmap marks the position of the coded amino acids (all 20 for the "liberal collection", 8 (G,A,P,D,E,V,I,L) for the Murchison collection). Marginal histograms show alphabet distributions in range or evenness. **B** The estimated probability that a random alphabet would show a broader range or a more even distribution than the coded 20. Data for **A** and **B** adapted from Mayer-Bacon and Freeland (2021) (Color figure online)

Beyond the mere existence of a better dataset, motivation to re-analyze unusual statistical patterns of genetically encoded amino acids comes from the exciting frontier of empirical evidence emerging to support an older theory of a simpler (smaller) amino acid alphabet that preceded the post-LUCA set of 20. The initial body of theory was solidified by the three different meta-analyses (Trifonov 2000; Higgs and Pudritz 2009; Cleaves 2010) which converged upon the same subset of 10 amino acids (Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr, and Val) as being both prebiotically plausible and, from diverse angles, the oldest members of the genetic code (Fig. 1), but this raised major, puzzling questions. For example, none of the positively charged, coded amino acids (Lys, Arg, His) occur within the putative early alphabet of 10, nor do any of the aromatic amino acids (Phe, Tyr, Trp). Within post-LUCA biology, cationic amino acids are indispensable to protein-based metabolism, including interaction with nucleic acids (Blanco et al. 2018) and removing (substituting) aromatic amino acids can result in loss of structural stability (Despotović et al. 2020).

Recent experimental work addresses these puzzles by suggesting how an early metabolism could function without "late" coded amino acids (Longo et al. 2020; Giacobelli et al. 2021). For example, an RNA-binding domain

was recently reconstructed using the "early" alphabet of 10 early amino acids by means of an Mg^{2+} cation-mediated interaction between the RNA and negatively charged, early amino acids (Giacobelli et al. 2021). Meanwhile, perfectly adequate protein-core packing is possible for robustly refoldable proteins with a plausible claim on being among those first discovered by molecular evolution: even where aromatic amino acid elimination destabilizes other folds, a halophilic environment can create and stabilize protein structure (Longo et al. 2013, 2015). These findings complement cheminformatics analyses suggesting that "early" amino acids are sufficient for folding and stability while "late" amino acids were recruited by subsequent evolution to improve the catalytic potential of genetically encoded proteins (Shibue et al. 2018; Kimura and Akanuma 2020).

Growing interest in a simpler, earlier amino acid alphabet and better datasets of prebiotic amino acids than anything studied previously, therefore, combine to motivate revisiting whether highly unusual pattern characterizing the full, standard alphabet is something that arrived with the addition of "late" amino acids, or a trait present throughout genetic code evolution. Here, we address this question by combining the most recent improvements to methods and data quality (Mayer Bacon and Freeland 2021) with equally careful improvements to the definition of "prebiotically plausible" and different assumptions about the membership of a simpler, earlier genetic code. Rather than focus on any single model, we ask how the analysis of amino acid coverage changes with the inclusion of additional abiotic amino acids using (i) the convergent analyses and meta-analyses (summarized in Fig. 1) of Trifonov (2000), Cleaves (2010), and Higgs and Pudritz (2009), (ii) a much updated and improved view of meteoritic astrochemistry, and (iii) variations between the two.

Methods

In order to test whether prebiotically plausible subsets of genetically encoded amino acids show similar non-random patterns to those seen in the full set of 20, various definitions of a simpler, earlier genetic code were tested against two structure libraries, each representing a different definition of plausible alternatives. For clarity, we refer to these major components of the analysis as "foregrounds" (subsets of genetically coded amino acids that represent an earlier stage of genetic coding) and "backgrounds" (libraries of alternative amino acids from which random sets are drawn for comparison), respectively. For each combination of foreground and background, we tested for each of two chemical descriptors (logP and van der Waals volume), the frequency with which an alphabet drawn at random from a given background exhibits a broader range of values, populated more evenly within that range, than the foreground. Detailed descriptions of each foreground and background, the test, and the descriptors are provided below.

Backgrounds are Libraries of Plausible Alternative Amino Acid Structures from Which Random Alphabets are Drawn

Two different backgrounds (structure libraries) were used to generate alternative amino acid alphabets: (i) a library of 54 α-amino acids that have been detected within meteorites by analytical chemistry, and (ii) a much larger library of 7155 monosubstituted α-amino acids generated computationally which combines the 54 α-amino acids of the first background with all coded amino acids absent from it and expands this set into one that comprises isomers and near-isomers of their sidechains. We refer to these two libraries as the "conservative background" and "liberal background," respectively, with a detailed description of each library as follows:

(i) The "*conservative background*" represents amino acids detected within carbonaceous chondrites, as reported by the 104 publications cited in Tables 1–4 of Simkus et al. (2019), an authoritative review of organic synthesis in this class of meteorites. Of these 104 publications, 25 were discarded prior to data entry because they did not present a direct, quantitative analysis of organic abundance (Earlier publications, in particular, sometimes estimated organic abundance or reasoned presence/absence rather than presenting direct, analytical results). For each of the remaining 79 publications, abundance data for each chemical compound, error values (if provided), and the associated DOI were copied manually into a spreadsheet according to the classification scheme provided by Simkus et al. (amino acids, amines, monocarboxylic acids, aldehydes, and ketones). Additionally, meteorite analysis papers published after Simkus et al. and before June 2020 were mined for usable data along with all works cited by those in Tables 1–4 but absent from the Tables themselves. All abundance data were recorded as both nmol/g and ppb. Unit conversions from published data were performed using the molecular weight described by either PubChem (https://pubchem.ncbi.nlm.nih.gov/) or ChemSpider (http://www.chemspider.com/). For each organic compound, any published isotopic ratios (commonly used to detect terrestrial contamination) and experimental extraction and detection methods (associated with variability in abundances) were recorded. Simkus et al. Sections 4.1 (Mitigation and Monitoring of Sample Contamination) and 5(Identifying Limitations of Cross-Comparisons between Studies) provide further, detailed information on isotopic ratios, methods of extraction, and

Structure	PubChem CID	Volume (Å^3)	logP	Coded	Special properties	αC substitution	Structure	PubChem CID	Volume (Å^3)	logP	Coded	Special properties	αC substitution	Structure	PubChem CID	Volume (Å^3)	logP	Coded	Special properties	αC substitution
	13314536	156	0.4			2		15817888	156	0.4			2		876	140	-1	✔	S	1
	306131	138	-0			1		791	138	0	✔		1		424	119	-2	✔		1
	351627	156	0.2			1		520755	156	0.4			1		1148	179	0.4	✔	A	1
	6119	104	-1			2		518928	156	0.4			1		1153	168	0	✔	A	1
	229525	138	0			2		235534	156	0.5			2		994	159	0.3	✔	A	1
	272725	156	0.3			1		95206	138	0.1			2		611	136	-2	✔		1
	1182	121	-1	✔		1		6657	104	-1			1		469	153	-1			1
	95515	156	0.3			2		9942136	138	-1			1		779	113	-2			1
	857	138	-0	✔		1		229526	138	0.1			2		617	95.2	-2	✔		1
	220783	156	0.4			1		521918	156	0.4			1		2901	126	-0			2
	237657	153	-1			1		412834	156	0.6			2		750	69.1	-2	✔		0
	5365	153	-1			1		824	121	-0			1		364	97.4	-3			1
	2109	136	-2			2		5105431	156	0.5			2		470	115	-2			1
	95440	153	-1			2		9475	138	0.1			1		389	132	-2			1
	94309	113	-2			2		227939	156	0.5			1		849	126	-1			1
	602	86.4	-1	✔		1		316542	104	-1			0		614	109	-1	✔		1
	5316628	115	-2			1		1088	86.4	-2			0		94744	121	-0			2
	205	113	-2	✔		1		4377	104	-1			1		15042684	156	0.4			1

Fig. 4 The 54 α-amino acids defining the conservative background. For each amino acid, a chemical structure is shown, along with PubChem compound ID and associated van der Waals volume and (capped) logP, calculated as described in Methods text. Genetically coded amino acids are indicated with a checkmark in the "Coded" column. Special properties: S = sulfur containing; A = aromatic. "αC substitution": number of carbon atoms directly attached to the α-carbon of the peptide backbone

detection of organics. The study presented here uses only the amino acid data, available through supplementary information. Readers interested in the wider dataset are encouraged to contact author Riley Havel.

As a result of this methodology, this first background comprises 54 α-amino acids shown in Fig. 4, of which 39 are monosubstituted at the α-carbon, a feature shared by almost all genetically encoded amino acids. The 54 derive from a total of 79 amino acids recorded, of which 25 were rejected for being β- (16), γ- (7), δ- (1) or ε- (1) amino acids. Although both helical and β-sheet-like conformations have been observed in β-amino acid polymers (e.g., Cheng et al. 2001), their exclusion from the analysis presented here reflects the prevailing consensus that β- and γ-amino acids and other prebiotically available compounds (such as hydroxy acids or dicarboxylic acids) would be less prone to form secondary and tertiary structures than α-amino acids (Weber and Miller 1981; Cleaves 2010). Most simply, inclusion of these molecules would add a new layer of assumptions by introducing structures qualitatively different from anything seen within the post-LUCA alphabet: this seems contrary to the goal of achieving greater clarity about whether a prebiotically plausible subset of genetically encoded amino acids retains the unusual statistical features of the full, standard amino acid alphabet. The 54 α-amino acids that remain thus represent a conservative baseline of plausible structural diversity (see, for example, Cleaves (2010) for expanded view of prebiotic availability) that is nevertheless a clear improvement on the Murchison data used in prior studies (Lu and Freeland 2008; Philip and Freeland 2011; Mayer-Bacon and Freeland 2021).

(ii) The "*liberal background*" expands the conservative background described above to include structural homologs of all meteoritic and coded amino acids. This expanded library was generated computationally following the same strategy used to construct the Combined Library of Meringer et al. (2013). Briefly, all coded amino acids and all 54 amino acids of the conservative background were divided into sub-libraries based on their backbone, sidechain heteroatoms and aromaticity to create a total of 10 fuzzy formulae (Fig. 5). By representing atom counts with numerical intervals rather than exact counts, each sublibrary's fuzzy chemical formula implies a set of structures that both cover and "fill in between" the set of structures from which it is derived. Each fuzzy formula was then provided as input to MOLGEN 5 (Gugisch et al. 2015) in order to generate all possible structural isomers implied by the fuzzy formula, except for limits on permissible ring size (< 5 atoms or > 10 atoms), maximum bond order of 2, and implausible substructures defined in three "badlists." These restrictions exclude sterically and energetically unstable amino acids. Three "badlists" containing implausible or unwanted substructures further limited the structures built by MOLGEN. Two of the badlists were distributed with MOLGEN 5.0, defining prohibited cyclic and unsaturated substructures that are universally regarded as structurally implausible within organic chemistry, as well as similarly forbidden bridged aromatic substructures with disallowed ring strain. A third badlist defined restricted substructures specific to α-amino acids, based on principles of chemical reactivity and stability. Badlists are provided in Supplementary Information (files *badist1*, *badlist2*,

Subset	Sidechain	Backbone	Coded AA's included	Structure count (conservative)	Structure count (liberal)
1	H	(i)	Gly	1	1
2	$C_{1\text{-}5}H_{3\text{-}11}$	(i)	Ala, Val, Leu, Ile	16	70
3	$C_{1\text{-}5}H_{3\text{-}11}O_{1\text{-}2}$	(i)	Ser, Thr, Glu, Asp	8	3030
4	$C_{1\text{-}4}H_{4\text{-}10}$	(ii)	Pro	6	63
5	$C_{1\text{-}3}H_{3\text{-}7}S$	(i)	Met, Cys	1	28
6	$C_{6\text{-}7}H_{5\text{-}7}$*	(i)	Phe	1	5
7	$C_{6\text{-}9}H_{6\text{-}12}N$*	(i)	Trp	1	1398
8	$C_{6\text{-}7}H_{5\text{-}7}O$*	(i)	Tyr	1	28
9	$C_{1\text{-}4}H_{3\text{-}14}N_{1\text{-}3}$	(i)	Lys, His, Arg	4	2431
10	$C_{1\text{-}3}H_{3\text{-}10}NO$	(i)	Gln, Asn	0	101
Σ				**39**	**7155**

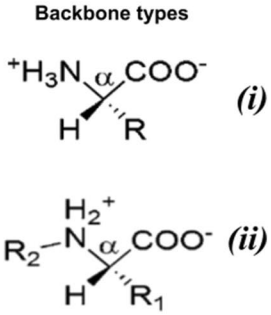


Fig. 5 Fuzzy formulae used to generate the liberal background. Amino acids from the conservative background are organized into groups which share a single fuzzy formula based on sidechain, aromaticity, and backbone composition. All chemically plausible structures implied by these sidechain-backbone combinations are generated by MOLGEN 5. An asterisk (*) structures must have a 6-membered aromatic ring. Note: subset 4 contains 22 peptoidal structures which, from the perspective of protein polymers, carry a "sidechain" branching from the amino group instead of the α-carbon

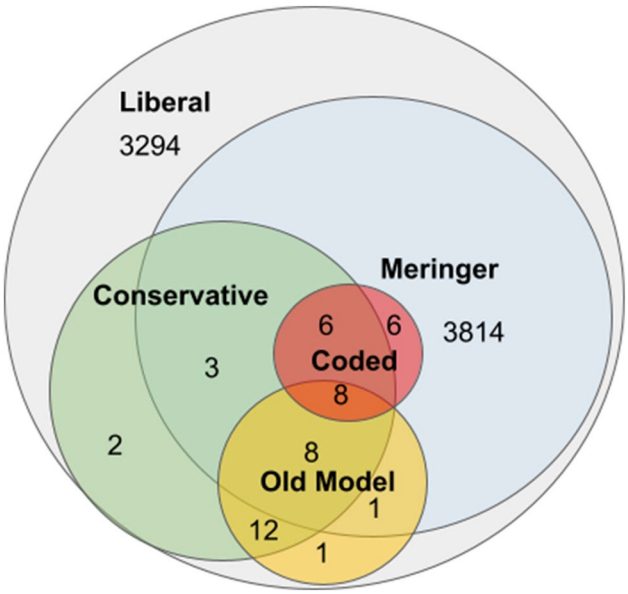


Fig. 6 Venn diagram showing the relationship among the amino acid sets used in this manuscript and prior, related literature (only monosubstituted α-amino acids shown). Meringer: combined library (AACL) described in Meringer et al. (2013) and used with modifications in later publications (Ilardo et al. 2015, 2019; Mayer-Bacon and Freeland 2021); Old Model: α-amino acids detected in the Murchison meteorite, as used in Philip and Freeland (2011) and Mayer-Bacon and Freeland (2021); Conservative: α-amino acids detected in meteorites and reviewed in Simkus et al. (2019) (see "Methods" for more details); Liberal: computational library built using the same principles as the AACL, expanded to include α-amino acids in the Conservative set but not represented in the AACL; Coded: the genetically coded set of 20 amino acids. Two amino acids detected within Murchison by earlier work (3-methylaspartic acid and 2-aminoheptanedioic acid) are not found within the conservative set because they are not associated with peer-reviewed quantitative meteoritic analyses

badlist3, see also *readme.txt)* Further discussion of structure generation and badlists is provided in Meringer et al. (2013). The resulting library comprises 7155 monosubstituted α-amino acids, shown as "Liberal" in Fig. 6 and is included in its entirety within Supplemental Information.

The resulting set contains 22 peptoidal compounds, i.e., those in which the sidechain attaches to the nitrogen of the amino group. This sub-class of α-amino acids is atypical of most coded amino acids, in which the sidechain is attached to the α-carbon. However, we consider amino acids of this type a logical extension of proline (backbone (ii) in Fig. 5), and such amino acids are found in the meteoritic set. Given that the number of non-peptoidal amino acids is more than 6 times greater than peptoidal compounds within the same volume and logP ranges (149 non-peptoidal amino acids with similar volumes and logP values as peptoidal compounds), the inclusion seems unlikely to influence results heavily.

At its simplest this expanded, computational background allows us to test the extent to which the small sample size of meteoritic amino acids leads to an absence of the patterns found within the genetic code. More subtly, this expanded background reflects the fact that additions and improvements to meteoritic detections continue, along with continued improvements to the analytical instrumentation and experimental protocols with which all such samples are analyzed consistently enlarged the scientific community's perception of what amino acids are plausible products of prebiotic chemistry. The liberal background, therefore, represents an estimated upper limit of structural diversity, intended to reveal whether further progress is even capable of changing perception of whether a simpler, earlier genetic code was using an unusual set of amino acids.

Foregrounds: Subsets of Coded Amino Acid Alphabet Reflecting Different Models of Prebiotic Plausibility

Against both of the backgrounds described above, five different "foregrounds" were tested. Each foreground represents a different assumption about the subset of genetically encoded amino acids which could have been present in a forerunner to the standard genetic code:

Foreground #1 comprises 10 amino acids (GAPDEVILST) which form the strongest consensus for components of an early amino acid alphabet (e.g., Higgs and Pudritz 2009).
Foreground #2 is identical to Foreground #1 but adds an eleventh amino acid, methionine (GAPDEVILSTM) which has been detected in meteoritic analyses (e.g., Kotra et al. 1979) and prebiotic simulations (e.g., Parker et al. 2011) but is treated as a "late" amino acid by all three meta-analyses which built early foundations for the idea of a simpler, earlier alphabet. A major reason to single out methionine for specific attention is the point made by Parker et al.'s (2011) analysis that many early prebiotic chemical simulations did not include sulfur as input, in any molecular form, and therefore, could not logically have recorded sulfurous amino acids as output. This limitation exerts a clear and equally logical bias on the relative position of methionine as an "early" or "late" amino acid in meta-analyses that include such literature. The value of this point is heightened by the fact that methionine brings a new atom type (sulfur) into consideration which aligns with a broader history of arguments for the chemistry which produced life (e.g., Ross 2008): an argument we expand upon in the Discussion section.
Foreground #3 is identical to foreground #1 but with the addition of aromatic, coded amino acids (FYW) to produce a total of 13 amino acid structures (GAPDEV-

ILSTFYW). This foreground reflects the detection of benzene-containing amino acids by some meteoritic analyses (Kotra et al. 1979; Chiesl et al. 2009; Pizzarello et al. 2012). The aromatics are worth distinguishing from methionine (Foreground #2), however, because no broader argument within genetic code literature supports aromatics as members of a simpler, earlier code. Indeed, strong consensus in multidisciplinary literature relegates them to the category of latecomers (Trifonov 2000; Higgs and Pudritz 2009; Fournier and Alm 2015).

Foreground #4 combines the additions of Foregrounds 2 and 3, including both Met and the aromatic-coded amino acids, to produce a set of 14 (GAPDEVILSTMFYW). As with Foreground 3, this rejects the consensus view of a multidisciplinary literature reflected in foregrounds 1 and 2 but enables us to establish the effect of separating unusual candidates for a simpler, earlier genetic code on the basis of such literature. Another way to express the value of foreground 4 is that it represents a strict view that meteoritic amino acid contents are the only (or uniquely reliable) guide to which of the 20 genetically encoded amino acids might reasonably have formed a simpler, earlier code.

Foreground #5 comprises the full set of 20 coded amino acids. This foreground allows a further control in the sense of allowing us to test whether random sets of size 20 drawn from the unfiltered conservative or liberal backgrounds can match the unusual features of the full amino acid alphabet's chemistry space.

For all tests involving foregrounds 1–4, all amino acids in the foreground set are also present in the background set, and the background does not contain any coded amino acids that are not present in the given foreground (e.g., no tests used methionine in the foreground but left it absent from the background). For the case of the liberal background, removing or adding specific coded amino acids to a given foreground meant also removing/adding their isomers and near-isomers to the background. For foreground #5, this is not the case as the intention here is merely to ask whether a given background is capable of matching the unusual properties of the full, standard alphabet of 20 amino acids.

Figure 6 shows a Venn diagram of how these various sets of amino acid structures defined for this analysis relate to one another and to the amino acid set(s) used in previous analysis (Mayer Bacon and Freeland 2021).

Testing Foregrounds Against Backgrounds

In order to test foregrounds against backgrounds, random alphabets drawn from each background were evaluated by the exact same procedure as Mayer-Bacon and Freeland (2021), i.e., using the definitions of "range," "evenness," and "coverage" shown in Fig. 7. Thus, the choice of foregrounds and backgrounds represent the only difference between analyses presented here and results published previously. For each specific test, 5 replicates of 1 million alphabets each were drawn at random from the appropriate background, each matching the size of the foreground in question. For example, when testing Foreground 1 (GAPDEVILST) against the conservative background, random sets of 10 amino acids were drawn from a subset comprising the entire conservative background (including all ten amino acids of the foreground), minus the coded amino acids reported from meteorites but not present in the foreground (i.e., Met, Tyr, Trp, and Phe). The 5 replicates of each test were used to generate a mean and standard deviation for the number of random alphabets that exhibit higher range, lower evenness (more even distribution) or both combined (better coverage) in a given descriptor.

Chemical Descriptors for Amino Acid Sidechains

All tests were performed for two chemical descriptors of amino acids, namely logP (hydrophobicity) and van der Waals' volume, following exactly the same procedure as Mayer-Bacon and Freeland (2021). Briefly, van der Waals volume was calculated using the method described by Zhao et al. (2003) as implemented in the Chemistry

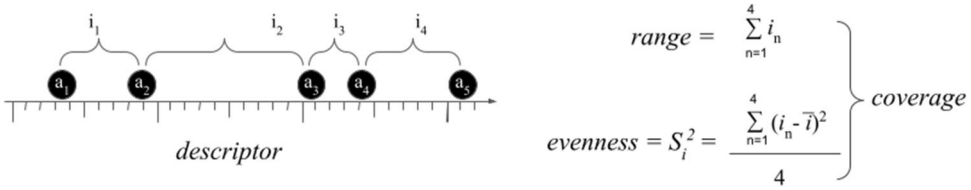


Fig. 7 Visual (left) and mathematical (right) descriptions of range, evenness, and coverage used in this study. Definitions are the same used in prior literature (Philip and Freeland 2011; Ilardo et al. 2015, 2019; Mayer-Bacon and Freeland 2021)**.** For a hypothetical set of 5 amino acids ($a_{1\text{-}5}$) with 4 intervals between them ($i_{1\text{-}4}$) in a given chemical descriptor, "range" is the sum of those intervals while "evenness" is the sample variance of those intervals. These two measurements combined define this set's coverage of the specific descriptor space; in that one alphabet is declared better than another if it exhibits larger range and a more even distribution within that range

Development Kit (CDK; https://cdk.github.io/). Amino acid "capping" was performed using the same methods used previously (Mayer-Bacon and Freeland 2021): the α-amine is acetylated and the α-carboxyl group is converted to a N-methylamide in order to better emulate the chemical properties of an amino acid as it appears within a protein sequence. logP calculation for each of these capped structures was performed using ChemAxon's (https://chemaxon.com/) Instant JChem program (version 19.26.0), a consensus of multiple leading estimation algorithms. The resulting descriptor values for each amino acid used in this analysis are shown in Fig. 4 for the conservative background, and this information along with equivalent, detailed data for the much larger liberal background is provided in Supplementary Information for the liberal background.

Results

Figure 8 shows the results of sampling 1 million alphabets drawn at random from a given library (background) of plausible alternatives and asking what fraction of these show a greater range of values, a more even distribution, or both (coverage = broader range and more even distribution) than a given subset of genetically encoded amino acids (foreground) for two descriptors (JChem logP and van der Waals volume). Overall, for the volume descriptor, the combination of broad range (larger range value than the foreground) and even distribution (lower evenness value than the foreground) exhibited by any foreground appears highly non-random in all tests without exception. For logP, the equivalent analyses are somewhat more heterogeneous: between 10 and 0% of random alphabets exceed the range and evenness of the foreground under scrutiny. That being said, only two tests produce non-significant values (i.e., > 5% of random alphabets "outperform" the foreground by the terms of our investigation) and the single most common result is 0%, to three significant figures. In all tests, 5 replicates of 1 million alphabets each produced less than 2 alphabets difference in the values reported in Fig. 8 (i.e., confidence intervals for all coverage values shown in Fig. 8 are pragmatically zero). Further, context for several of these estimated probabilities may be seen in the underlying distributions of range and evenness for random alphabets in supplementary information (Figure S1). A finer-grained description of results for each test is as follows:

Test E corroborates all previous, published analyses by showing that it is exceedingly difficult, if not impossible, to find a set of 20 prebiotically plausible amino acids, or near-isomers, which emulate the range and evenness of the full standard alphabet in both logP and volume. This finding makes one, minor addition to this robust finding of previous literature by showing that this result is unlikely to change with any foreseeable expansion of perceived prebiotic plausibility based on isomers or near-isomers of those that are known at present.

Test B demonstrates that if one were to focus solely on amino acids detected within meteorites to define which of the 20 genetically encoded amino acids were present in a simpler, earlier genetic code, then these 14 structures exhibit unusually good coverage in both descriptors relative to plausible alternatives: 3% of random alphabets match or exceed coverage in logP for the conservative background, 2% for the liberal background. In both cases, these "better" random alphabets achieve their status mainly by increasing the range of hydrophobicities over that observed within the genetically encoded subset.

Tests A, C, and D reflect different interpretations of a simpler, earlier amino acid alphabet, considering both the multidisciplinary literature from which this idea derives and the vision of prebiotic plausibility informed by meteorites. Test A illustrates the ambiguity which motivated the current analysis by demonstrating that even when the foreground used by prior tests is expanded from a subset of 8 amino acids to 10 by the inclusion of Ser and Thr (reflecting consensus literature; see Fig. 2), the range and evenness of logP for this subset is outperformed by either 8% of random alphabets or 0%, depending on whether one considers a background informed strictly by meteoritic possibilities (conservative background), or one that is far more saturated by computationally generated library of isomers and near-isomers (liberal background). Corresponding results for volume, as noted above, are clear and unequivocal for both backgrounds.

The third and fourth tests (C and D) inform the influence exerted by two anomalies that result from combining the distinct ideas of (i) a simpler, earlier genetic code and (ii) using meteorites to define what amino acids would have been available as products of prebiotic chemistry. Test D includes methionine, a sulfur-containing amino acid, as a prebiotically plausible amino acid in both foreground (genetically encoded amino acids) and background (plausible alternatives) but excludes the aromatic amino acids (Trp, Phe, and Tyr), even though they are reported from some meteorite analyses. Test C includes the aromatics but excludes methionine/sulfur. Of the eight specific coverage results inherent to tests C and D (two descriptors × two backgrounds × two different foregrounds), only one of the two descriptors considered (logP) fails to appear statistically significant by the terms of one foreground (D) but does so against both backgrounds. For the other six tests, 2% or fewer of random alphabets exceed the coverage of the foreground, depending on the exact choice of descriptor, foreground, and background.

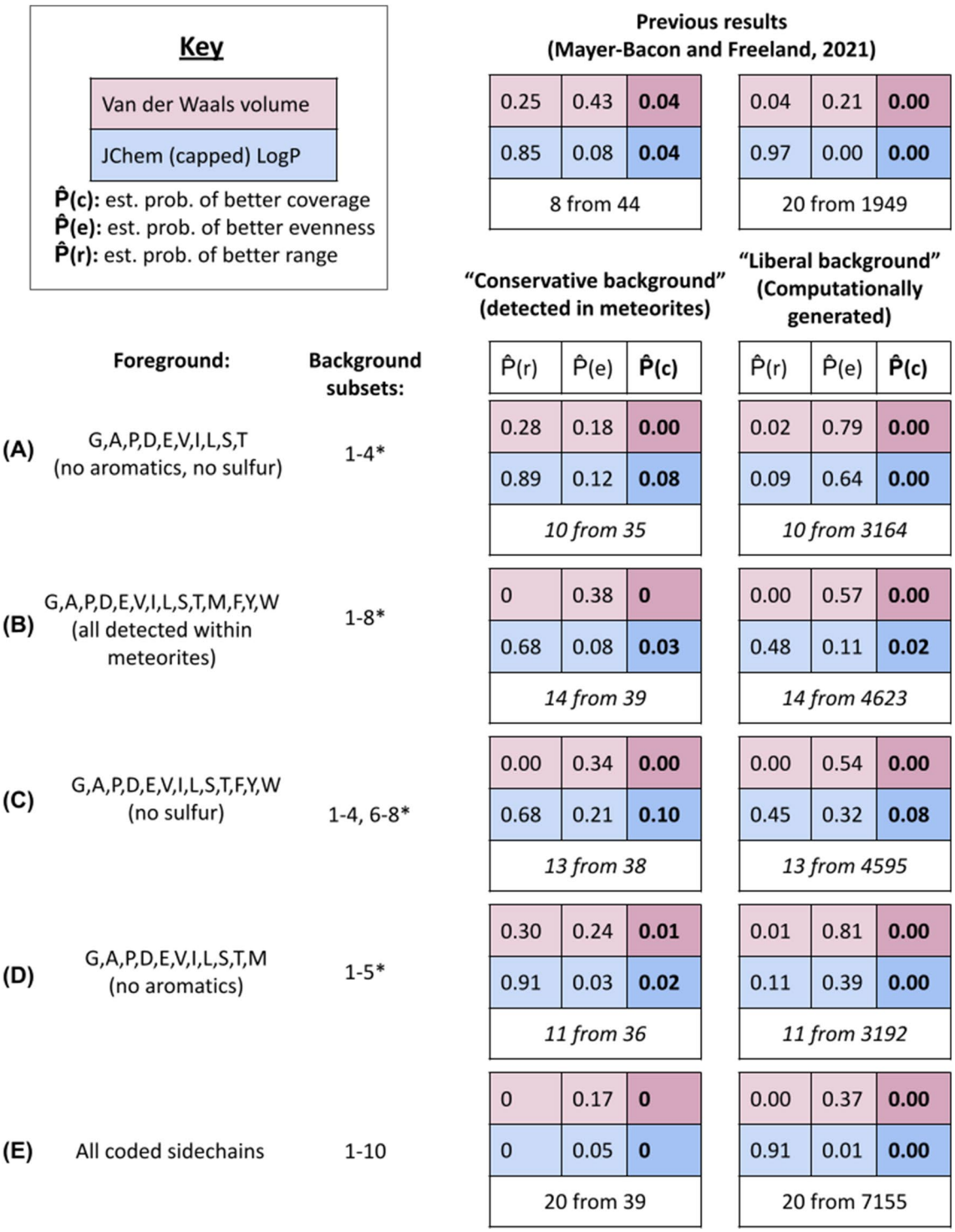


Fig. 8 Estimated probability ($\widehat{P}$) that an alphabet of amino acids drawn at random from a given library (background) exhibits broader range, a more even distribution or both simultaneously (coverage) than a specific subset of coded sidechains (foreground) for the two properties of *logP* and *van der Waals volume*. Values represent the mean of 5 replicates, as described in Methods. Numbers in the "Background subsets" column refer to subsets numbered in Fig. 5. The "*x from y*" nomenclature under each of the 10 tests indicates that *x* amino acids (where *x* is the size of the foreground) are selected at random from a pool of *y* amino acids (where *y* is the size of the background). Most combinations of foreground and background indicate that statistically significant ($\widehat{P}$<0.05) patterns of coverage are present within a prebiotically plausible subset of genetically encoded amino acids for both molecular descriptors. *The conservative set contains 4 amino acids with monoamine sidechains (ornithine, 2,4-diaminobutanoic acid, 2,3-diaminobutanoic acid, 2,3-diaminopropionic acid). These are included in all conservative backgrounds, while only test A uses a liberal background with nitrogenous sidechains

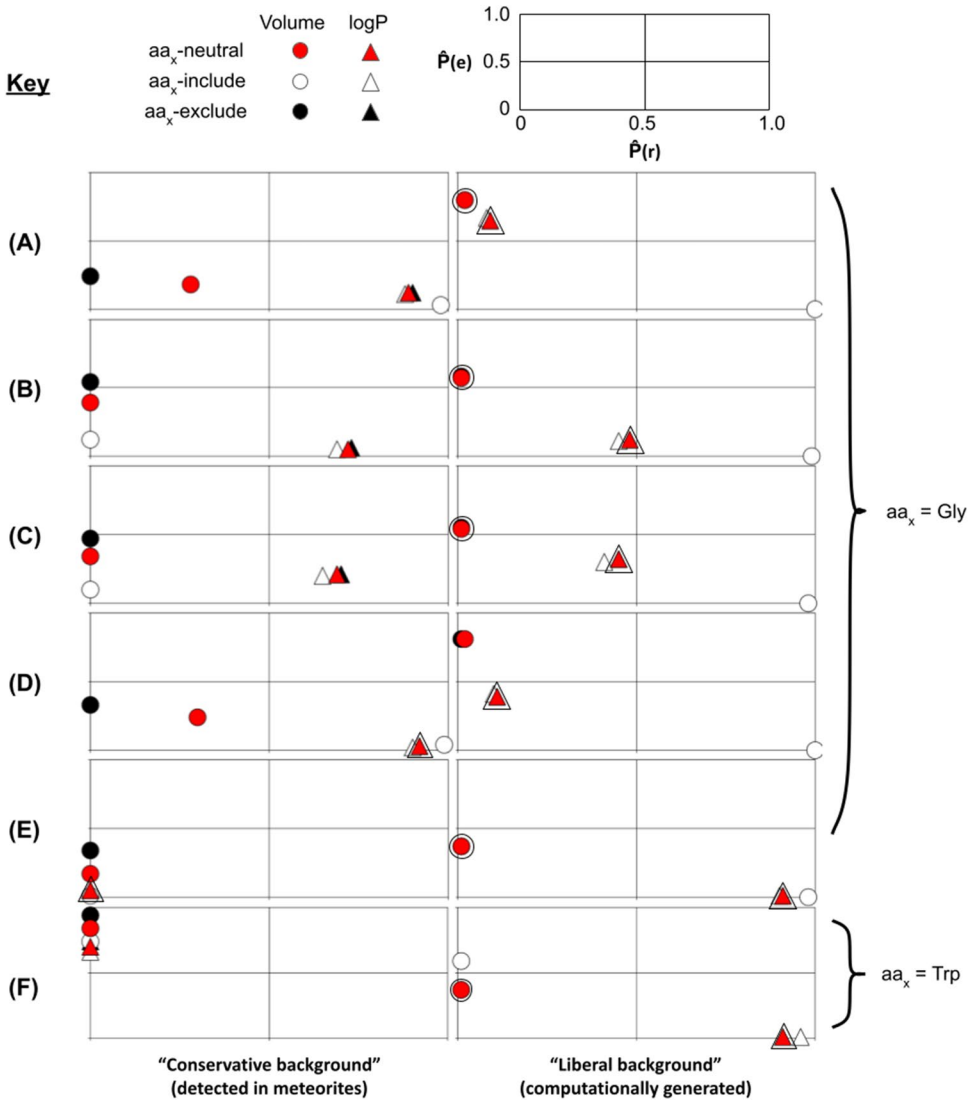


Fig. 9 Summary of changes to the results presented in Fig. 8 when the same tests are adjusted to either force the inclusion of glycine in all random alphabets (white shapes) or force its exclusion (black shapes), shown in comparison with the previously reported results in which glycine is allowed but not required (red shapes). P̂(r) on the horizontal axis plots the estimated probability (observed frequency) that a random alphabet exhibits a broader range in a given descriptor; P̂(e) on the vertical axis plots the estimated probability (observed frequency) that a random alphabet is more evenly distributed within its range for a given descriptor. Row letters A–E correspond to the foregrounds and backgrounds described in Fig. 8; row F uses the same foreground and background as row E, but examines forced inclusion or exclusion of tryptophan instead of glycine. A black border around a red shape indicates a red circle or triangle that is placed behind that shape (Color figure online)

Since the coded amino acids include glycine, which is objectively unique in that no simpler α-amino acid structure exists, a further round of tests investigates the extent to which glycine alone influences the findings reported thus far (Fig. 9). For all tests conducted against the liberal background (right column of Fig. 9), forcing glycine inclusion (white circles) ensures that nearly all random sets exhibit a broader range of volumes: this completely changes the perception of an unusual range (red circles and Fig. 8). However, this higher probability of broader range comes at the cost of a much lower probability of finding a more even volume distribution once glycine is included. Against this same background, excluding glycine had very little effect on volume range and distribution.

For tests using the conservative background, the situation is a little more complex (left column of Fig. 9). In general, glycine inclusion or exclusion has minimal effect on whether a random alphabet has either a broader range or more even distribution of logP values.

Row F of Fig. 9 provides more context for this inclusion/exclusion of glycine by forcing inclusion or exclusion of tryptophan (instead of glycine) from all random alphabets. Similar to glycine exclusion from the conservative tests (rows A-E), Trp exclusion yields more even volume distributions. Neither Trp inclusion nor exclusion appears to affect the range of logP values or how evenly those values are distributed. For alphabets built from the liberal background, Trp inclusion has the opposite effect of glycine inclusion seen in Fig. 9A–E, yielding negligible changes in volume ranges but slightly more even volume distributions. Trp inclusion slightly increases the range of logP values in random alphabets, but this increase is much smaller than the increase seen for the range of volumes under glycine inclusion (rows A-E, liberal background). Similar to the effects of glycine

exclusion, Trp exclusion had little effect on the range and distribution of logP values for random alphabets built from the liberal background.

All data shown in Figs. 8 and 9, as discussed in "Methods", use background libraries filtered to remove all α-disubstituted amino acids since genetically encoded amino acids are all monosubstituted on the α-carbon atom. However, equivalent tests were conducted with the few α-disubstituted amino acids detected within meteorites included (Figure S2). No qualitative differences in coverage occur, and quantitative differences are of small degree. For example, in the analogous test to the conservative background from Fig. 8A (background of 51 amino acids instead of 35), the anomalous result for logP changes from 8% (monosubstituted-only) to 7% (mono- and disubstituted α-amino acids).

Discussion

Results presented here revisit and seek to clarify an ambiguity about the statistical properties of the subset of genetically encoded amino acids that are proposed by a consensus of prior literature to have formed a simpler, earlier stage in genetic code evolution. The ambiguity in question is that the same improvements to methods and data which have strengthened evidence for a strikingly non-random full alphabet of 20 amino acids have simultaneously weakened evidence for a similar pattern in a subset of 8, prebiotically plausible amino acids. Our re-analysis of this latter finding is motivated by the existence of better data about prebiotically plausible amino acids and by developments in experimental protein science which support older, theoretical arguments for this simpler, earlier genetic code.

Broadly speaking, the analyses presented here support the idea of a prebiotically plausible subset of the coded amino acids that does, in fact, emulate the unusual properties of the entire alphabet under a wide range of assumptions about prebiotic plausibility. There are some exceptions for one of the two descriptors studied (logP), but these exceptions mostly involve combinations of foreground and background that are hardest to justify. For example, the test in Fig. 8C which assumed aromatics (but not methionine) were part of an earlier genetic code matches no known claims about the scope of amino acids used by a simpler, earlier code. This does not imply that detection of these "late" amino acids in meteorites is inaccurate: detection of compounds related to Phe, Tyr, and Trp such as phenol (Naraoka et al. 1999) and indole (Remusat et al. 2005) indicates that aromatic structures can form abiotically. Rather, a broad, multidisciplinary literature that has investigated genetic code evolution from multiple perspectives has repeatedly found that, regardless of their availability, aromatic amino acids entered genetic-coding late as biosynthetic modifications of the simpler, earlier alphabet. In this sense, we suggest that the tests shown in Fig. 8B (analysis of all 14 genetically encoded amino acids that have been detected within meteorites) and 8C (aromatics, but no sulfur) are best interpreted less as a serious contender for a simpler, earlier code than a corroboration of the previously reported idea that exceptional size and hydrophobicity are features inherited by the full amino acid alphabet "from its subsets" (Ilardo et al. 2019).

Figure 8A presents the major exception to this overall summary of findings. This particular test represents the single best-defined, consensus view of a simpler, earlier genetic code as one comprising the 10 amino acids GAPDEVILST. Here, we see either the highest (10%) or the lowest (0%) percentage of random alphabets outperforming the foreground, depending on which background one chooses to regard as a better model for plausible alternatives. Thus, the most straightforward test of whether a simpler, earlier genetic code exhibited the same unusual patterns for amino acid distribution as the full, final code depends upon which of these two backgrounds is a better representation of plausible alternatives against which the coded subset should be compared.

In truth, each background presents strengths and weaknesses. While meteorites remain a widely accepted empirical guide to prebiotic chemistry, the diversity of amino acids detected therein has increased over time with both improved instrumentation and the addition of new meteorite samples (Elsila et al. 2016). There is no clear reason to believe that current data have reached an asymptote in this regard. A wider generalization of this point is that while meteorites may provide an invaluable insight into prebiotic chemistry, they are by no means the only guide to what was available to the origin and early evolution of life on Earth. In this sense, the liberal background is more of a theoretical limit to future visions of prebiotic plausibility. It is likely an overestimate in that there is no clear reason why all isomers and near-isomers of current meteoritic detections should have been available to an early genetic code. It seems likely that an accurate background of amino acid possibilities lies, undefined, somewhere between the conservative and liberal models explored here. The results shown in Fig. 8 suggest that future growth to either the foreground of "early" amino acids or the background of possible alternatives is quite likely to strengthen evidence that a simpler, earlier code emulated unusual properties of the full amino acid alphabet; that expansion remains, for now, conjecture. In this sense, the central question motivating this study remains unresolved by the new analyses presented here. However, any frustration is mitigated by clarifying context provided by the network of other tests presented which demonstrate clearly that ambiguity over unusual properties of an early

alphabet stems from the small sample size of "prebiotically plausible" amino acids and from the precise contents of the presumed, early alphabet.

Tests which probe the specific role played by glycine (Fig. 9A–E) illustrate the point. While glycine plays very little role in accounting for the unusual distribution of logP values, this unique amino acid can have significant effects in perceptions of unusual volume distribution. Including glycine in all random alphabets makes it much easier to find an otherwise randomized alphabet with broader range than the coded subset, but only at the cost of making it much more difficult to match the evenness of the coded subset. This is because α-amino acid chemistry space is not populated uniformly. There are exponentially increasing numbers of sidechains possible with linear increases in the number of heavy atoms present within a sidechain. In any library of possible amino acid structures, small amino acids remain invariant in number while larger compounds compose the vast majority of structures, and this density of structures increases rapidly with each additional heavy atom. Therefore, including glycine in all random alphabets forcibly includes an outlier in amino acid chemistry space, making it far easier to find a random alphabet with a broader range in volume but at the cost of making it far harder to find an alphabet which matches the evenness of the coded set. This emphasizes how an answer to the question: does a smaller, earlier amino acid alphabet emulate the unusual properties of the full alphabet? Depends sensitively on the choice of foreground and background.

Similar reasoning explains why the inclusion or exclusion of tryptophan (Fig. 9F) from otherwise random alphabets has much less of an effect on volume and logP distributions. Although tryptophan is the largest and most hydrophobic amino acid in the conservative library, and thus, has a strong effect here, the liberal background is heavily populated by large, aromatic amino acids with side chains that are structural isomers of tryptophan. The forced inclusion or exclusion of tryptophan in otherwise random alphabets, therefore, has little to no effect on results obtained using the liberal background.

Another example, arguably more relevant to thinking about simpler, earlier codes is that adding methionine to the consensus, early alphabet of 10 improves the perception of unusual range and unusual evenness, which then improves further still when the aromatics are added. Conversely, adding aromatics before Met decreases this perception of unusual properties relative to the consensus early alphabet of 10. This difference between the two orders of incorporation offers one clear (if small) way in which to distinguish whether unusual range and evenness were features of an early genetic code and/or consistent features conserved during amino acid alphabet expansion. Methionine of course brings not only an additional side chain but an additional atom type: sulfur.

The identification of Met as a "late" amino acid comes largely from meta-analyses which have synthesized multiple, different and specific models for amino acid alphabet evolution, As noted in "Methods", Parker et al.'s (2011) analysis pointed out that many early prebiotic chemical simulations did not include sulfur as input, in any molecular form, and therefore could not logically have recorded sulfurous amino acids as output. This omission exerts a clear bias towards producing the consensus view that Met was a late amino acid. However, organosulfur compounds are common in meteorites (Shimoyama and Katsumata 2001; Zherebker et al. 2021), and it is, therefore, notable that many organic-solvent extraction procedures remove molecular sulfur prior to analysis (J. C. Aponte, *pers. comm.*). Thus, while exceptional hydrophobicity coverage when methionine is present in the foreground (with or without aromatics) most certainly reflects adaptive properties of the full amino acid alphabet being "inherited from its subsets" (Ilardo et al. 2019), it might well also signify something more. A considerable literature argues for the antiquity of protein sulfur biochemistry, including those who propose a key role for metal sulfide minerals as catalytic centers for the earliest metabolic processes (Wächtershäuser 1992; Martin and Russell 2007). Sulfur certainly seems to offer significant potential to play a structural role in the fundamental chemistry of life (Lavergne et al. 2013; Malyshev et al. 2014; Feldman et al. 2019).

These remaining ambiguities and unknowns suggest that we may usefully end this discussion by reviewing briefly why it matters to understand the chemical logic of an early alphabet. Currently, the deepest challenge for the entire lineage of research to which the present analysis contributes is to interpret clearly the cause of an unusual pattern of amino acid physicochemical properties. We and others have inferred an outcome of natural selection for an optimal set of building blocks with which to construct proteins. Typical reasoning invokes clade selection by arguing that evolving lineages which were best able to approximate a "perfect" combination of size and hydrophobicity at any residue within a given protein sequence were those best able to adapt and diversify within an ever-changing world. The inference of a selective advantage here comes from Anfinsen's Nobel-prize winning demonstration that "*at least for a small globular protein in its standard physiological environment, the native structure is determined only by the protein's amino acid sequence*" (Anfinsen 1973). We certainly defend the plausibility of that interpretation based on current data but accept that other interpretations remain plausible at this point. Amino acids distributed evenly across a broad range of volumes and hydrophobicities could also, for example, represent some as yet unknown version of biochemical

constraint. The nearest evidence for this view of which we are aware comes from a finding that three of the genetically encoded amino acids (Lys, Arg, and His) tend to oligomerize with each other better than with three other, plausible alternatives which did not become part of the genetic code (Frenkel-Pinter et al. 2019). While this is an interesting result, it is difficult to offer more than conjecture what, if anything, it contributes to the findings described here. For example, these three coded amino acids are universally identified as late additions to the genetic code (see Fig. 1): even if selective oligomerization played a role in the incorporation of Lys, His, and Arg into life's amino acid alphabet, this finding does not extrapolate to explain evolutionary forces which shaped earlier, smaller amino acid alphabets. The small number of α-amino acids studied for selective oligomerization (six, plus two α-hydroxy acids) leaves open the question of whether and how this possible constraint extrapolates to the wider variety of amino acids considered here. For example, even the conservative, strictly meteoritic library comprises more than six times as many amino acids. The broader point we draw from Frenkel-Pinter is, therefore that with each demonstration of unusual distribution of physicochemical properties for the coded amino acids comes an increasing motivation to understand better what it signifies. It seems inevitable that clarity about adaptive arguments will come ultimately from experimental work that explores the structure-building potential of different amino acid alphabets.

With this in mind, two overlapping reasons argue for a continued focus on a putative simpler, earlier stage of amino acid alphabet evolution. One is the evidence, steadily developing, that whereas the early amino acid alphabet seems suitable for polymerizing into folding structures, later alphabet additions functioned more as catalysts and antioxidants (Granold et al. 2018; Moosmann 2021). This idea suggests not only that the full alphabet and an earlier forerunner might exhibit different chemical "logic," but also that the early alphabet is the key to understanding how to choose a set of amino acids capable of producing diverse, stable folds.

A more practical second reason for studying the putative early alphabet is that any experimental work to test the structure-building potential of different amino acid alphabets has to contend with the combinatorial mathematics of polymer construction. An oligomer of length n constructed using an alphabet of size s may take one of s^n sequences. Halving the alphabet size from 20 to 10 reduces the search space of possible sequences by orders of magnitude. We, therefore, conclude that while the study presented here leaves core questions unanswered, it presents a valuable foundation for future work. We have collated structure data and associated descriptors for meteoritic amino acids, a computational expansion of this set and baseline analyses that future research may usefully use to clarify what physicochemical properties allowed an early amino acid alphabet to form diverse, stable protein folds.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-022-10061-5.

Acknowledgements R. H.'s work was supported in part by the NASA Astrobiology Institute through funding awarded to the Goddard Center for Astrobiology under proposal 13-13NAI7-0032. We would like to thank UMBC undergraduate Ian Squires for his help in generating the data shown in Fig. 9, along with two anonymous reviewers and the JME editor Andrew Ellington for helpful insights which have improved this manuscript considerably.

Author Contributions CM-B, SF: Conceptualization. MM, RH: Resources. CM-B, SF, MM: Methodology. CM-B, RH, JCA: Data curation. CM-B, MM: Software. CM-B, SF: Investigation. CM-B: Visualization. CM-B, MM, RH, SF: Writing—original draft. CM-B, MM, RH, JCA, SF: Writing—review and editing. SF: Project administration.

References

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223. https://doi.org/10.1126/science.181.4096.223

Aponte JC, Elsila JE, Hein JE et al (2020) Analysis of amino acids, hydroxy acids, and amines in CR chondrites. Meteorit Planet Sci 55:2422–2439. https://doi.org/10.1111/maps.13586

Benner SA, Ellington AD, Tauer A (1989) Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci 86:7054–7058. https://doi.org/10.1073/pnas.86.18.7054

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. https://doi.org/10.1016/j.cub.2018.01.014

Cheng RP, Gellman SH, DeGrado WF (2001) β-peptides: from structure to function. Chem Rev 101:3219–3232. https://doi.org/10.1021/cr000045i

Chiesl TN, Chu WK, Stockton AM et al (2009) Enhanced amine and amino acid analysis using pacific blue and the mars organic analyzer microchip capillary electrophoresis system. Anal Chem 81:2537–2544. https://doi.org/10.1021/ac8023334

Cleaves HJ (2010) The origin of the biologically coded amino acids. J Theor Biol 263:490–498. https://doi.org/10.1016/j.jtbi.2009.12.014

Cooper G, Kimmich N, Belisle W et al (2001) Carbonaceous meteorites as a source of sugar-related organic compounds for the early earth. Nature 414:879–883. https://doi.org/10.1038/414879a

Despotović D, Longo LM, Aharon E et al (2020) Polyamines mediate folding of primordial hyperacidic helical proteins. Biochemistry 59:4456–4462. https://doi.org/10.1021/acs.biochem.0c00800

Elsila JE, Aponte JC, Blackmond DG et al (2016) Meteoritic amino acids: diversity in compositions reflects parent body histories. ACS Cent Sci 2:370–379. https://doi.org/10.1021/acscentsci.6b00074

Feldman AW, Dien VT, Karadeema RJ et al (2019) Optimization of replication, transcription, and translation in a semi-synthetic organism. J Am Chem Soc 141:10644–10653. https://doi.org/10.1021/jacs.9b02075

Fournier GP, Alm EJ (2015) Ancestral reconstruction of a Pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. J Mol Evol 80:171–185. https://doi.org/10.1007/s00239-015-9672-1

Frenkel-Pinter M, Haynes JW, Martins C et al (2019) Selective incorporation of proteinaceous over nonproteinaceous cationic amino acids in model prebiotic oligomerization reactions. Proc Natl Acad Sci 116:16338–16346. https://doi.org/10.1073/pnas.1904849116

Giacobelli VG, Fujishima K, Lepsik M, et al (2021) In vitro evolution reveals primordial RNA-protein interaction mediated by metal cations. https://doi.org/10.1101/2021.08.01.454623

Granold M, Hajieva P, Toşa MI et al (2018) Modern diversification of the amino acid repertoire driven by oxygen. Proc Natl Acad Sci 115:41–46. https://doi.org/10.1073/pnas.1717100115

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Gugisch R, Kerber A, Kohnert A et al (2015) Chapter 6—MOLGEN 50, a molecular structure generator. In: Basak SC, Restrepo G, Villaveces JL (eds) Advances in mathematical chemistry and applications. Bentham Science Publishers, United Arab Emirates, pp 113–138

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. https://doi.org/10.1089/ast.2008.0280

Ilardo M, Meringer M, Freeland S et al (2015) Extraordinarily adaptive properties of the genetically encoded amino acids. Sci Rep 5:9414. https://doi.org/10.1038/srep09414

Ilardo M, Bose R, Meringer M et al (2019) Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. Sci Rep 9:12468. https://doi.org/10.1038/s41598-019-47574-x

Johnson AP, Cleaves HJ, Dworkin JP et al (2008) The miller volcanic spark discharge experiment. Science 322:404–404. https://doi.org/10.1126/science.1161527

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. https://doi.org/10.1007/s00239-020-09938-0

Koga T, Naraoka H (2017) A new family of extraterrestrial amino acids in the murchison meteorite. Sci Rep 7:636. https://doi.org/10.1038/s41598-017-00693-9

Kotra RK, Shimoyama A, Ponnamperuma C, Hare PE (1979) Amino acids in a carbonaceous chondrite from Antarctica. J Mol Evol 13:179–183. https://doi.org/10.1007/BF01739477

Lavergne T, Degardin M, Malyshev DA et al (2013) Expanding the scope of replicable unnatural DNA: stepwise optimization of a predominantly hydrophobic base pair. J Am Chem Soc 135:5408–5419. https://doi.org/10.1021/ja312148q

Longo LM, Lee J, Blaber M (2013) Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc Natl Acad Sci 110:2135–2139. https://doi.org/10.1073/pnas.1219530110

Longo LM, Tenorio CA, Kumru OS et al (2015) A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding: aromatic amino acids and mesophile adaptation. Protein Sci 24:27–37. https://doi.org/10.1002/pro.2580

Longo LM, Despotović D, Weil-Ktorza O et al (2020) Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. Proc Natl Acad Sci 117:15731. https://doi.org/10.1073/pnas.2001989117

Lu Y, Freeland SJ (2008) A quantitative investigation of the chemical space surrounding amino acid alphabet formation. J Theor Biol 250:349–361. https://doi.org/10.1016/j.jtbi.2007.10.007

Malyshev DA, Dhami K, Lavergne T et al (2014) A semi-synthetic organism with an expanded genetic alphabet. Nature 509:385–388. https://doi.org/10.1038/nature13314

Martin W, Russell MJ (2007) On the origin of biochemistry at an alkaline hydrothermal vent. Philos Trans R Soc B Biol Sci 362:1887–1926. https://doi.org/10.1098/rstb.2006.1881

Mayer-Bacon C, Freeland SJ (2021) A broader context for understanding amino acid alphabet optimality. J Theor Biol 520:110661. https://doi.org/10.1016/j.jtbi.2021.110661

Mayer-Bacon C, Agboha N, Muscalli M, Freeland S (2021) Evolution as a guide to designing xeno amino acid alphabets. Int J Mol Sci. https://doi.org/10.3390/ijms22062787

Meringer M, Cleaves HJ, Freeland SJ (2013) Beyond terrestrial biology: charting the chemical universe of α-amino acid structures. J Chem Inf Model 53:2851–2862. https://doi.org/10.1021/ci400209n

Moosmann B (2021) Redox biochemistry of the genetic code. Trends Biochem Sci 46:83–86. https://doi.org/10.1016/j.tibs.2020.10.008

Naraoka H, Shimoyama A, Harada K (1999) Molecular distribution of monocarboxylic acids in Asuka carbonaceous chondrites from Antarctica. Orig Life Evol Biosph 29:187–201. https://doi.org/10.1023/A:1006547127028

Parker ET, Cleaves HJ, Dworkin JP et al (2011) Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. Proc Natl Acad Sci 108:5526–5531. https://doi.org/10.1073/pnas.1019191108

Philip GK, Freeland SJ (2011) Did evolution select a nonrandom "alphabet" of amino acids? Astrobiology 11:235–240. https://doi.org/10.1089/ast.2010.0567

Pizzarello S (2007) The chemistry that preceded life's origin: a study guide from meteorites. Chem Biodivers 4:680–693. https://doi.org/10.1002/cbdv.200790058

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 109:11949–11954. https://doi.org/10.1073/pnas.1204865109

Remusat L, Derenne S, Robert F, Knicker H (2005) New pyrolytic and spectroscopic data on orgueil and murchison insoluble organic matter: a different origin than soluble? Geochim Cosmochim Acta 69:3919–3932. https://doi.org/10.1016/j.gca.2005.02.032

Ross D (2008) A quantitative evaluation of the iron-sulfur world and its relevance to life's origins. Astrobiology 8:267–272. https://doi.org/10.1089/ast.2007.0199

Shibue R, Sasamoto T, Shimada M et al (2018) Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. Sci Rep 8:1227. https://doi.org/10.1038/s41598-018-19561-1

Shimoyama A, Katsumata H (2001) Polynuclear aromatic thiophenes in the murchison carbonaceous chondrite. Chem Lett 30:202–203. https://doi.org/10.1246/cl.2001.202

Simkus DN, Aponte JC, Elsila JE et al (2019) Methodologies for analyzing soluble organic compounds in extraterrestrial samples: amino acids, amines, monocarboxylic acids, aldehydes, and ketones. Life 9:47. https://doi.org/10.3390/life9020047

Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261:139–151. https://doi.org/10.1016/S0378-1119(00)00476-5

Tuller T, Birin H, Gophna U et al (2010) Reconstructing ancestral gene content by coevolution. Genome Res 20:122–132. https://doi.org/10.1101/gr.096115.109

Wächtershäuser G (1992) Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog Biophys Mol Biol 58:85–201. https://doi.org/10.1016/0079-6107(92)90022-X

Weber AL, Miller SL (1981) Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol 17:273–284. https://doi.org/10.1007/BF01795749

Weiss MC, Preiner M, Xavier JC et al (2018) The last universal common ancestor between ancient earth chemistry and the onset of genetics. PLOS Genet 14:e1007518. https://doi.org/10.1371/journal.pgen.1007518

White HB (1976) Coenzymes as fossils of an earlier metabolic state. J Mol Evol 7:101–104. https://doi.org/10.1007/BF01732468

Wong JT-F, Bronskill PM (1979) Inadequacy of prebiotic synthesis as origin of proteinous amino acids. J Mol Evol 13:115–125. https://doi.org/10.1007/BF01732867

Zhao YH, Abraham MH, Zissimos AM (2003) Fast calculation of Van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. J Org Chem 68:7368–7373. https://doi.org/10.1021/jo034808o

Zherebker A, Kostyukevich Y, Volkov DS et al (2021) Speciation of organosulfur compounds in carbonaceous chondrites. Sci Rep 11:7410. https://doi.org/10.1038/s41598-021-86576-6