CHARLES UNIVERSITY IN PRAGUE

Faculty of Science

Department of Developmental and Cell Biology



Mikhail Makarov

Biophysical and structural constraints underlying amino acid alphabet evolution

Biofyzikální a strukturní faktory v evoluci aminokyselinové abecedy

Doctoral thesis

Supervisor: Mgr. Klára Hlouchová, Ph.D.

Prague 2024

Declaration

I hereby declare that I have written this thesis independently, and all resources as well as co-authors were properly indicated. I also declare that I did not submit this thesis or a part of it to obtain any other academic degree.

Prague, 2024 Mikhail Makarov

Abstract

Despite the universal scientific acclaim of the RNA-centred view on the origin and early evolution of cellular life, the numerous lines of experimental evidence indicate the high feasibility of prebiotic synthesis and condensation of amino acids into simple yet functional ancestral peptides that aided the establishment of proto-cellular life and predated the emergence of modern structurally and functionally diverse proteins. Both exogenous and endogenous scenarios of amino acid formation indicate the prebiotic availability of more than 90 different amino acids, however, only ten of them (referred to as early canonical amino acids) persisted throughout life evolution and were recruited to the genetic code for cellular protein synthesis. The remaining half of the genetically encoded amino acids (referred to as late canonical amino acids) were presumably introduced after the establishment of their corresponding biosynthetic pathways. This work addresses the lack of evolutionary continuity between the prebiotic availability of amino acids and the amino acid composition of the extant proteome. The particular aims of this work were to determine the selection factors that led to (i) selective inclusion of early canonical amino acids over their non-proteinogenic alternatives; and (ii) recruitment of late canonical amino acids to the genetic code during early protein evolution. The comparative biophysical and structural analysis of 25-mer combinatorial peptide libraries performed in the first part of the dissertation indicates that the early canonical amino acids were selected based on their propensity to induce both the high solubility and high secondary structure formation in the earliest oligopeptides. The functional analysis of ancestral ribosomal peptides enriched in late canonical amino acids conducted in the second part of the dissertation indicates that late canonical amino acids could be recruited to the genetic code to support the early RNA-protein co-evolution. Overall, the findings of our studies support the idea of the highly non-random and evolutionary adaptive expansion of the amino acid alphabet.

Key words: origin of life, amino acid alphabet evolution, non-canonical amino acids, early ribosome evolution.

Abstrakt

Ačkoliv poslední dvě dekády výzkumu vzniku života byly výrazně pod vlivem teorií tzv. RNA světa, četné experimenty nasvědčují prebiotické syntéze a kondenzaci aminokyselin do jednoduchých peptidů. Prebiotická dostupnost a odolnost takových polymerů je navíc výrazně vyšší než u nukleových kyselin a peptidy tak (nezávisle či v součinnosti s dalšími probiotickými hráči) velmi pravděpodobně předcházely vzniku moderních strukturálně a funkčně různorodých proteinů. Exogenní i endogenní scénáře syntézy aminokyselin ukazují na prebiotickou dostupnost více než 90 různých aminokyselin, avšak pouze deset z nich (označovaných jako rané kanonické aminokyseliny) přetrvávaly po celou dobu evoluce života a byly zahrnuty do genetického kódu pro buněčnou proteosyntézu. Zbývající polovina geneticky kódovaných aminokyselin (označovaných jako pozdní kanonické aminokyseliny) byla pravděpodobně zahrnuta po vytvoření jejich biosyntetických drah. Tato práce se zabývá nedostatkem evoluční kontinuity mezi prebiotickou dostupností aminokyselin a aminokyselinovým složením moderního proteomu. Konkrétní cílem této práce bylo určit selekční faktory, které mohly vést k (i) selektivnímu zařazení raných kanonických aminokyselin před jejich neproteinogenními alternativami; a (ii) zařazení pozdních kanonických aminokyselin do genetického kódu během časné evoluce proteinu. Biofyzikální a strukturní analýza 25-merních randomizovaných peptidových knihoven provedená v první části disertační práce ukazuje, že rané kanonické aminokyseliny vykazovaly adaptivní vlastnosti na základě jejich schopnosti indukovat jak vysokou rozpustnost, tak i vysokou tvorbu sekundární struktury v ancestrálních oligopeptidech. Funkční analýza ancestrálních ribozomálních peptidů obohacených o pozdní kanonické aminokyseliny provedená v druhé části disertační práce dale ukazuje, že inkorporace pozdních kanonických aminokyselin mohla být klíčová pro stabilitu a kompartmentalizaci RNA-proteinové koevoluce, jak bylo pozorováno na příkladu rané evoluce ribozomu. Celkově výsledky našich studií podporují myšlenku vysoce nenáhodné a evolučně adaptivní expanze aminokyselinové abecedy.

Klíčová slova: vznik života, evoluce aminokyselinové abecedy, nekanonické aminokyseliny, raná evoluce ribozomů.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Klára Hlouchová, for the patient guidance, encouragement and support throughout these years. I would also like to thank our long-time collaborators, Prof. Stephen Fried and Dr. Kosuke Fujishima, for providing valuable advice during the implementation of our projects. But most of all, I would like to thank my late mother, Stepanova Eleonora. Thank you so much, mom, it all would be nothing without you. I wish you could see this.

Table of contents

List of abbreviations 8
1. Introduction 10
1.1. Abiotic formation of amino acids 10
1.1.1. Endogenous sources of amino acids 10
1.1.2. Exogenous sources of amino acids 14
1.2. Prebiotic condensation of amino acids into peptides 18
1.2.1. Thermal approaches to prebiotic peptide synthesis 18
1.2.2. Prebiotic peptide synthesis with organic condensation agents 20
1.2.3. Prebiotic peptide synthesis with inorganic condensation agents 21
1.2.4. Abiotic co-condensation of amino acids with other biological precursors 23
1.3. The role of peptides in the origin of life 25
1.3.1. Catalytic properties of short peptides 25
1.3.2. Self-assembly of short peptides 27
1.3.3. Compartmentalisation and molecular recognition by short peptides 30
1.4. Formation of proto-ribosome and early translation machinery 31
1.5. Evolution of the genetic code 38
1.5.1. The frozen accident theory of the genetic code evolution 38
1.5.2. The adaptation theory of the genetic code evolution 39
1.5.3. The stereochemical theory of the genetic code evolution 39
1.5.4. The co-evolution theory of the genetic code evolution 40
2. Aims of the thesis 46
3. Materials and methods 47
3.1. Biophysical and structural analysis of early, alternative early, and late combinatorial peptide libraries 47
3.1.1. Design of combinatorial peptide libraries 47
3.1.2. Chemical synthesis and quality control of combinatorial peptide libraries 47
3.1.3. Solubility analysis of combinatorial peptide libraries 48
3.1.4. Aggregation propensity analysis of combinatorial peptide libraries 49
3.1.5. Secondary structure analysis of combinatorial peptide libraries 49
3.2. Biophysical and structural analysis of differently charged combinatorial peptide libraries 50
3.2.1. Design of combinatorial peptide libraries 50
3.2.2. Chemical synthesis and quality control of combinatorial peptide libraries 51

3.2.3. Solubility analysis of combinatorial peptide libraries 51
3.2.4. Aggregation propensity analysis of combinatorial peptide libraries 51
3.2.5. Secondary structure analysis of combinatorial peptide libraries 51
3.2.6. Isolation of monomer, aggregate, and total fractions from 10E peptide library 51
3.2.7. Compositional, biophysical, and structural analysis of monomer, aggregate, and total fractions from 10E peptide library 52
3.3. Characterization of ancestral PTC rRNA-rPeptide interactions 52
3.3.1. Preparation of ancestral PTC rRNA constructs 52
3.3.2. Fluorescent labelling of ancestral PTC rRNA constructs 55
3.3.3. Design and chemical synthesis of rPeptides 55
3.3.4. Ancestral PTC rRNA-rPeptide interaction analysis by microscale thermophoresis 55
4. Results 57
4.1. Exploration of biophysical and structural properties of early, alternative early, and modern amino acid alphabets 57
4.1.1. Biophysical analysis of combinatorial peptide libraries 60
4.1.2. Structural analysis of combinatorial peptide libraries 65
4.2. Exploration of biophysical and structural properties of differently charged amino acid alphabets 70
4.2.1. Biophysical analysis of combinatorial peptide libraries 72
4.2.2. Structural analysis of combinatorial peptide libraries 78
4.2.3. Compositional, biophysical, and structural analysis of monomer, aggregate, and total fractions from 10E peptide library 82
4.3. Characterization of RNA-binding propensity of ancestral ribosomal peptides enriched in late proteinogenic amino acids 86
5. Discussion 92
5.1. Foldability defined the selection of early canonical amino acid alphabet 92
5.2. Random negatively charged peptides undergo spontaneous self-assembly 98
5.3. Late proteinogenic amino acids were crucial for the early ribosome evolution 104
6. Conclusions 109
7. References 110
8. Publications and submitted manuscripts 138

List of abbreviations

10E	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu
10U	subset of Gly, Ala, Aba, Nva, Nle, Pro, Ser, Thr, Asp, Glu
11D	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu, Dab
11R	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu, Arg
11Y	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu, Tyr
19F	full canonical amino acid set except Cys
8E	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr
8E+Aad,Api	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Aad, Api
8E+Asp,Glu	10E subset
8E+Dap,Dab	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Dap, Dab
8E+Dap,Dab,Aad,Api	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Dap, Dab, Aad, Api
8E+Dap,Dab,Asp,Glu	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Dap, Dab, Asp, Glu
8E+Lys,Arg	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Lys, Arg
8E+Lys,Arg,Aad,Api	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Lys, Arg, Aad, Api
8E+Lys,Arg,Asp,Glu	subset of Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Lys, Arg, Asp, Glu
A215	absorbance at 215 nm
AA	amino acid
Aba	2-aminobutyric acid
Aad	2-aminoadipic acid
ABP	acetate-borate-phosphate
Api	2-aminopimelic acid
Boc	tert-butyloxycarbonyl
BSA	bovine serum albumin
CD	circular dichroism
Dab	2,4-diaminobutyric acid
Dap	2,3-diaminopropionic acid
DHB	2,5-dihydroxybenzoic acid
DMF	N,N-dimethylformamide
DMSO	dimethyl sulfoxide
dNTP	deoxynucleotide triphosphate
dsDNA	double-stranded DNA
Fmoc	9-fluorenylmethoxycarbonyl
FTSC	fluorescein-5-thiosemicarbazide
g^6A	N6-glycinylcarbamoyladenosine
HPLC	high performance liquid chromatography
K_d	dissociation constant
LUCA	last universal common ancestor
MALDI-TOF	matrix-assisted laser desorption/ionization-time of flight
mdeg	millidegree
MMW	mean molecular weight
mnm^5U	5-methylaminomethyl uridine
MST	microscale thermophoresis

MWCO	molecular weight cut-off
Nle	norleucine
NMR	nuclear magnetic resonance
Nva	norvaline
OPA	o-phthalaldehyde
Orn	ornithine
PAGE	polyacrylamide gel electrophoresis
PDB	Protein Data Bank
pK_a	acid dissociation constant
pNPA	p-nitrophenyl acetate
PTC	peptidyl transferase centre
rpm	revolutions per minute
SDS	sodium dodecyl sulphate
SEC	size-exclusion chromatography
ssDNA	single-stranded DNA
t^6A	N6-threonylcarbamoyladenosine
TAE	Tris-acetate-EDTA
TFE	2,2,2-trifluoroethanol
UV-Vis	ultraviolet-visible
v/v	volume per volume

1. Introduction

The literature review is divided into sections that cover main stages of early protein evolution starting from abiotic synthesis of amino acids from inorganic precursors and ending with the establishment of the genetic code by the time of the last universal common ancestor (LUCA). The further biological evolution of proteins that accompanied divergence of the three domains of life from the LUCA is not covered in this review. It is important to note that this literature review serves to highlight the plausibility of formation and significance of early peptides and proteins in the origin and early evolution of life but does not imply peptides/proteins to be the first and only biomolecules responsible for the emergence of life. The questions surrounding the origin of life invoke many chicken-and-egg dilemmas of "which came first?". The absence of definitive answers to the origin of life questions led to numerous contradicting and sometimes even mutually exclusive hypotheses that serve as bases for divisive "Me-first" scenarios with "RNA world" scenario being the most popularised in scientific literature. The ever-growing evidence on abiotic formation of all building blocks of life (amino acids, nucleobases, sugars, fatty acids) as well as the more recent origin-of-life studies focusing on the functional cooperation between them indicate that emergence of life is the result of symbiotic chemical evolution of numerous major biomolecules. This review aims to provide a structured summary of the findings that elucidate different aspects of early protein evolution as well as highlight possible symbiotic co-evolution of proteins, lipids and nucleic acids.

1.1. Abiotic formation of amino acids

Formation of organic matter from simple inorganic precursors is considered the earliest stage of abiogenesis. The concept of abiogenesis takes its roots from the milestone Miller-Urey experiment, in which a mixture of simple gases (H_2, NH_3, CH_4, H_2O) subjected to mild heating and electric discharges yielded a complex mixture of organic molecules including five amino acids: Gly, Ala, Asp, β-alanine, and 2-aminobutyric acid (Miller 1953). The relative easiness of synthesis of organic compounds from inorganic precursors under exposure to various sources of energy led to numerous studies on abiotic formation of amino acids that support their either endogenous or exogenous origin.

1.1.1. Endogenous sources of amino acids

Endogenous (terrestrial) origin implies formation of amino acids from (i) atmospheric gases present in prebiotic Earth; or (ii) from dissolved small organic molecules present in hydrothermal vents.

It is generally agreed that Earth was formed 4.56 billion years ago (Dalrymple 2001). Abiotic formation of amino acids from atmospheric gases could have presumably occurred after the emergence of liquid water on prebiotic Earth ~4.4 billion years ago (Wilde *et al.* 2001). The classic experiments on formation of amino acids from atmospheric gases assumed that the primitive Earth atmosphere was highly reducing and consisted of simple reduced gases such as H_2, NH_3, CH_4, C_2H_6, H_2O, and H_2S, most of which were introduced into the primordial atmosphere by volcanic eruptions (Kasting 1993; Tian *et al.* 2005). Formation of amino acids from these gases could be promoted by different forms of energy such as lightning (electric spark), volcanic eruptions (high temperatures), sunlight and cosmic rays (ultraviolet (UV), γ- and X-ray radiation) (Miller 1953; Miller and Urey 1959; Sagan and Khare 1971; Lawless and Boynton 1973; Takahashi *et al.* 1999). Amino acids are produced from atmospheric gases as racemic mixtures of diverse C2−C6 isomers with the yields that decrease exponentially with the increasing chemical complexity, i.e. number of carbon atoms. The precise amino acid repertoire of the mixtures obtained in the numerous experiments varies depending on the gas mixture composition and the source of energy. To date, 10 proteinogenic amino acids (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu) and 21 non-proteinogenic amino acids have been obtained in quantifiable amounts from reduced atmospheric gas mixtures (CH_4, NH_3/N_2, and H_2O) (Miller 1953; Miller 1955; Ring *et al.* 1972; Ruiz-Bermejo 2007; Parker *et al.* 2014) (Fig. 1.1). The inclusion of H_2S in reduced gas mixtures (CH_4, NH_3, N_2, H_2O) led to additional identification of S-methylcysteine and Met as well as several organosulphur compounds including cysteamine, which could be formed by Cys degradation (Parker *et al.* 2011).

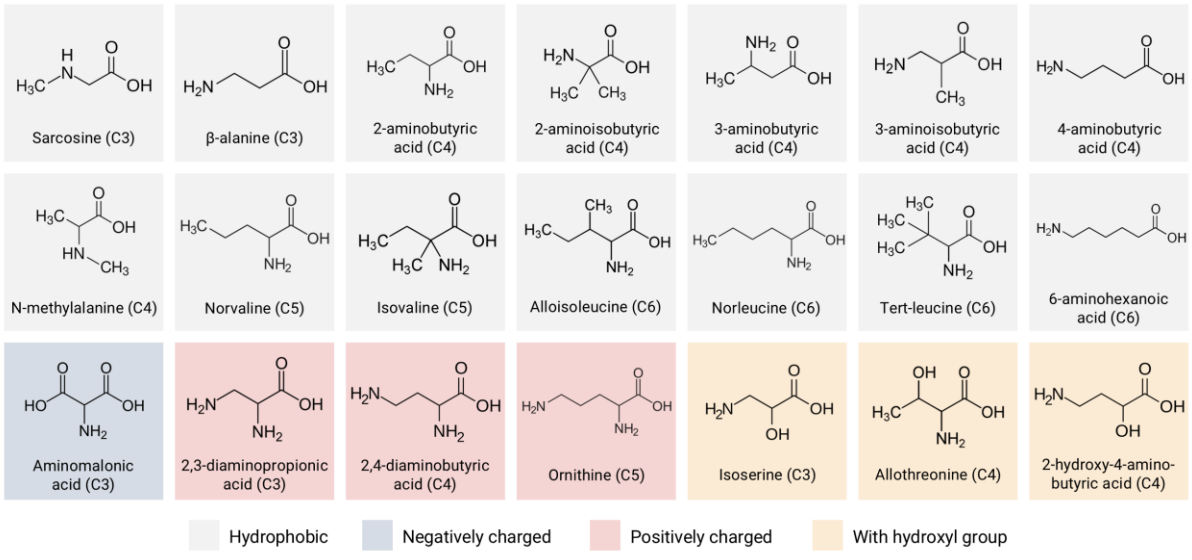


Fig. 1.1. Chemical structures of non-proteinogenic amino acids obtained from the reduced gas mixtures. Amino acids are divided into four groups based on the side chain character.

The assumption of highly reducing conditions of the prebiotic Earth atmosphere is divisive due to the lack of definitive geologic evidence (Miyakawa *et al.* 2002; Chyba 2005). Several geochemical studies indicate that the primitive terrestrial atmosphere was most likely neutral or weakly reducing and was composed of N_2, CO, CO_2, H_2O with small amounts of H_2, CH_4, NH_3, and H_2S (Kasting 1993). The gravitational forces of the primordial Earth were presumably significantly weaker, which could lead to the escape of H_2 from the early atmosphere, while the strong UV irradiation of the Earth by the primordial Sun could lead to photodecomposition of CH_4, H_2S and NH_3, which, however, still could have been present in the atmosphere in smaller concentrations due to volcanic outgassing and hydrothermal vent production (Levine *et al.* 1982; Kasting *et al.* 1989; Kasting 1993). The initial attempts to simulate abiotic synthesis of amino acids from neutral Earth's atmosphere (the mixture of N_2, CO/CO_2, and H_2O) led to insufficient yields of amino acids due to significant formation of nitrogen oxides (NO and NO_2), which leads to accumulation of nitrous (HNO_2) and nitric (HNO_3) acids that cause oxidative and hydrolytic degradation of organic compounds (Schlesinger and Miller 1983; Hirose *et al.* 1990; Cleaves *et al.* 2008). The addition of reducing agents (such as $FeSO_4$) and calcium carbonate ($CaCO_3$), which would neutralise nitrous and nitric acids, significantly improves the yields of amino acids making them comparable to those obtained from simulated reducing Earth atmosphere (Cleaves *et al.* 2008). In general, simulations on the neutral Earth atmosphere led to less diverse amino acid mixtures that are represented by 6 proteinogenic (Gly, Ala, Val, Ser, Asp, Glu) and 5 non-proteinogenic (β-alanine, 2-aminobutyric acid, 2-aminoisobutyric acid, 4-aminobutyric acid, norvaline) amino acids (Schlesinger and Miller 1983; Hirose *et al.* 1990; Miyakawa *et al.* 2002; Cleaves *et al.* 2008).

Abiotic formation of amino acids in these experiments starts with gas-phase reactions between radical and ion species produced from atmospheric gases by electric discharges (or other forms of energy), which leads to formation of hydrogen cyanide (HCN) and simple organic compounds such as formaldehyde (HCHO), higher aldehydes/ketones, amines and nitriles. Once dissolved in water, HCN and simple organic compounds can undergo diverse condensation and oligomerization reactions producing a big variety of biological precursors including amino acids, sugars and even nitrogenous bases that presumably constituted the "primordial soup" (Schwartz 1983; Cleaves *et al.* 2008). α-amino acids can be formed by several mechanisms including (i) Bucherer-Bergs synthesis, which involves formation of hydantoins from HCN, $(NH_4)_2CO_3$ and respective aldehydes/ketones followed by its hydrolysis; (ii) hydrolysis of HCN oligomerization products; and (iii) Strecker reaction, which

involves condensation reaction between HCN, NH_3, and respective aldehydes (Miller 1957; Harada 1967; Taillades *et al.* 1998). β-amino acids in turn can be formed by Michael addition of ammonia to α,β-unsaturated nitriles resulting in β-aminonitriles, which are subsequently hydrolysed into β-amino acids (Miller 1957). All mechanisms leading to formation of α- and β-amino acids require the presence of water (Fig. 1.2). The chemical mechanisms underlying formation of ɣ- and δ-amino acids are poorly understood.

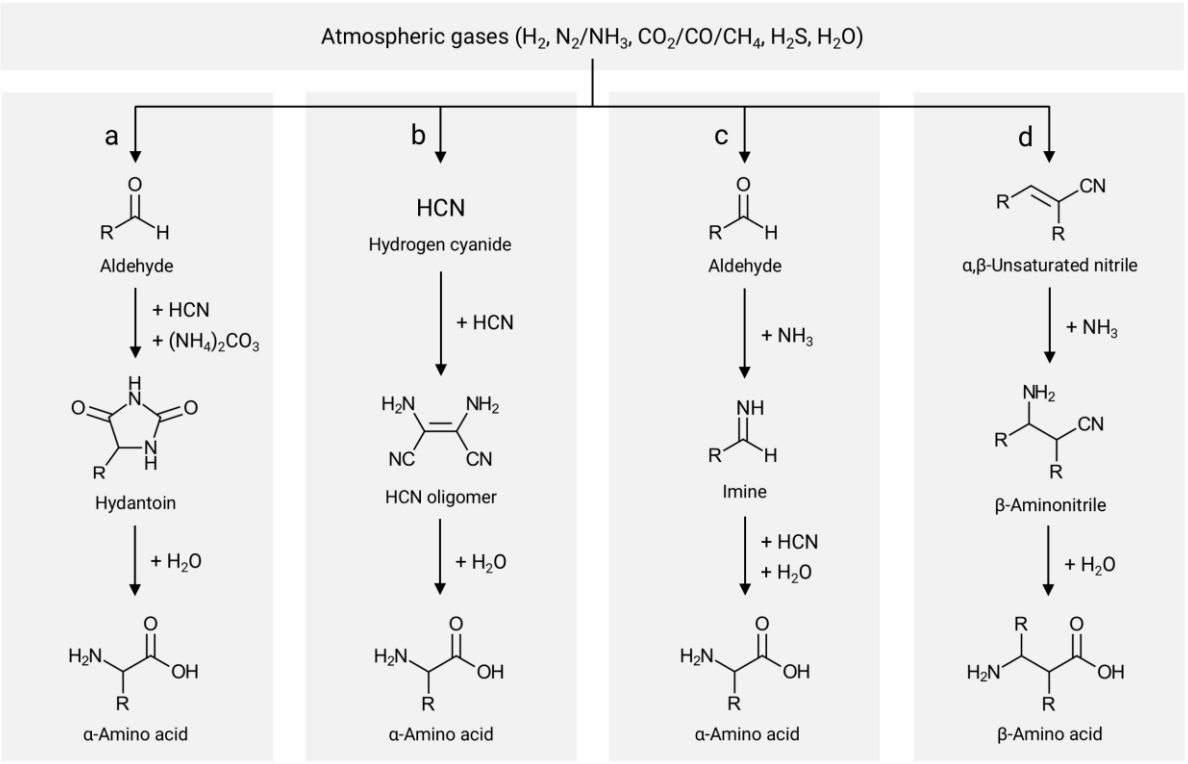


Fig. 1.2. Chemical mechanisms underlying formation of α- and β-amino acids. (a) The Bucherer-Bergs synthesis of amino acids from hydantoins; (b) Formation of α-amino acids by HCN oligomerization; (c) The Strecker synthesis of α-amino acids by the addition of NH_3 and HCN to aldehydes and ketones; and (d) Formation of β-amino acids by Michaelis addition of ammonia to α,β-unsaturated nitriles (Miller 1957; Harada 1967; Taillades *et al.* 1998).

An alternative (and mutually not exclusive) scenario proposes that the first building blocks of life including amino acids were formed at deep sea hydrothermal vents. Deep sea hydrothermal vents produce hot (up to 400 °C) alkaline fluids enriched in molecular hydrogen that forms as a by-product of serpentinization – oxidation of ferrous (Fe^{2+}-rich) ferromagnesian minerals such as olivine and pyroxene into ferric (Fe^{3+}-rich) serpentinite coupled with reduction of water to hydrogen (Mével 2003). Such reducing conditions could promote formation of simple organic molecules from inorganic carbon species (McCollom and Seewald 2007). Hydrothermal fluids are enriched in various organic compounds that are mostly formed by either metabolic processes of hydrothermal biomass (biogenic origin) or its thermal decomposition (thermogenic origin) (McCollom and Seewald 2007). Several abiotically

synthesised small organic molecules such as simple hydrocarbons (CH_4, C_2H_6, C_3H_8, C_4H_{10}) and formate were reported to be formed in the hydrothermal vents by Fischer-Tropsch reactions between CO and H_2 and reduction of CO_2, respectively (Konn *et al.* 2015). Their abiotic formation is thought to be driven not only by high temperatures provided by hydrothermal vents but also supported by hydrothermal minerals via adsorption (in case of zeolites and layered silicate minerals such as montmorillonite) or transition metal catalysis (in cases of Fe/Ni-enriched minerals such as olivine and pyroxene) (Holm and Neubeck 2009). Computational studies and laboratory simulations indicate that hydrothermal conditions favour reduction of CO_2 to CO/HCHO/CH_4 and N_2/nitrite/nitrate to NH_3/HCN that could serve as the precursors for amino acid synthesis (Holm and Neubeck 2009; Konn *et al.* 2015). Several reports demonstrate formation of 12 proteinogenic (Gly, Ala, Val, Leu, Ile, Pro, Asp, Glu, Ser, Thr, Lys, Phe) and 5 non-proteinogenic (β-alanine, sarcosine, 2-aminobutyric acid, 2-aminoisobutyric acid, 4-aminobutyric acid) amino acids from the gas mixtures of CH_4 and N_2 or highly concentrated solutions of HCHO, NH_3, and HCN or their derivatives (NH_4HCO_3, NaCN/KCN) under plausible hydrothermal conditions (Fox and Windsor 1970; Hennet *et al.* 1992; Yanagawa and Kobayashi 1992; Marshall 1994; Aubrey *et al.* 2009). The additional inclusion of reduced sulphur species such as H_2S and NaHS to the reaction mixtures did not lead to the emergence of sulphur-containing amino acids in these experiments (Fox and Windsor 1970; Aubrey *et al.* 2009). The analysis of water collected from the Lost City hydrothermal field shows its enrichment in amino acids compared to local sea water, however, these amino acids are most likely of biotic origin (Lang *et al.* 2013). The hydrothermal scenario of amino acid synthesis has been highly disputed by many origin-of-life researchers due to the unknown equilibrium between formation and thermal decomposition of amino acids at elevated temperatures (Ito *et al.* 2006; Aubrey *et al.* 2009). Despite the questionable plausibility of hydrothermal amino acid formation, hydrothermal vents could still supply the primordial Earth atmosphere with reduced atmospheric gases (CH_4, H_2, NH_3) that are necessary for electric spark-induced amino acid synthesis.

1.1.2. Exogenous sources of amino acids

Exogenous (extraterrestrial) origin implies formation of amino acids in space and their delivery to prebiotic Earth by meteorites. The evidence of extraterrestrial formation of amino acids comes from the amino acid analyses of carbonaceous chondrites, a specific group of stony (non-metallic) meteorites that contain water and organic matter of extraterrestrial origin (Kerridge 1985). To date, the most chemically complex and extensively studied carbonaceous chondrite, Murchinson meteorite, which fell in Australia in 1969, has been shown to contain

96 amino acids, 12 of which are genetically encoded (Gly, Ala, Val, Leu, Ile, Ser, Pro, Thr, Asp, Glu, Phe, Tyr) (Koga and Naraoka 2017; Glavin *et al.* 2018). Amino acids found in the Murchinson meteorite range from two to ten carbons in length and are predominantly represented by α-amino isomers (Koga and Naraoka 2017; Glavin *et al.* 2018). Meteoritic amino acids are present as quasi-equimolar racemic mixtures of relatively high abundance (parts per million) and are highly enriched in deuterium, ^{13}C and ^{15}N isotopes indicating their extraterrestrial origin (Epstein *et al.* 1987; Engel and Macko 1997). Similarly to the abiotic syntheses from atmospheric gas mixtures, the relative abundances of meteoritic amino acids drastically decrease with the increasing chemical complexity. Even though most of the meteoritic amino acids are scarce or not present in the Earth biosphere, it is important to note that meteoritic amino acids may not fully reflect formation of amino acids on planetesimals (asteroids and comets) due to possible destruction/alterations caused by initial collision with Earth, exposure to atmospheric water and oxygen and terrestrial contamination by organic compounds from soil (Burton *et al.* 2012; Lee *et al.* 2021). The absolute and relative contents of meteoritic amino acids reported by different studies of the Murchinson meteorite vary drastically due to the significant heterogeneity of meteorite samples, differences in protocols for amino acid extraction (water extraction vs acidic hydrolysis), different analytical sensitivities of instruments and possible contamination by terrestrial amino acids (Zaia *et al.* 2008; Burton *et al.* 2012).

The clearer evidence on extraterrestrial origin of amino acids comes from the sample-return missions such as Stardust mission (that brought samples from the comet 81P/Wild-2 in 2006); JAXA Hayabusa mission (that brought samples from the Itokawa asteroid in 2010) and the recently completed JAXA Hayabusa2 mission (that brought samples from the Ryugu asteroid in 2020). The cometary dust grain samples from the comet 81P/Wild-2 were generally poor in organic material but were shown to contain Gly whose extraterrestrial origin was confirmed by carbon isotope analysis (Elsila *et al.* 2009). Several non-proteinaceous amino acids (β-alanine, 3-aminobutyric acid, 3-aminoisobutyric acid) found in the Itokawa asteroid were confirmed to be of extraterrestrial origin (Parker *et al.* 2022). The Ryugu asteroid fragments have been shown to contain 11 proteinogenic (Gly, Ala, Val, Leu, Ile, Ser, Thr, Asp, Glu, Phe, Tyr) and 17 non-proteinogenic (N,N-dimethylglycine, β-alanine, 2-aminobutyric acid, 2-aminoisobutyric acid, 3-aminobutyric acid, 3-aminoisobutyric acid, 4-aminobutyric acid, 3-amino-2,2-dimethylpropanoic acid, norvaline, isovaline, 3-aminopentanoic acid, 5-aminopentanoic acid, 2-aminoadipic acid, 2-aminopimelic acid, pipecolic acid, cycloleucine,

6-aminohexanoic acid) amino acids of extraterrestrial origin (Potiszil *et al.* 2023; Parker *et al.* 2023).

Simple amino acids are hypothesised to have formed by gas-phase reactions between radicals and ion-radicals present in the interstellar medium, even prior to the emergence of the solar system (Blagojevic *et al.* 2003). Several decades of astrophysical studies have led to identification of 256 inorganic and organic molecules in the interstellar medium, most of which exist in the ionised or radical form due to strong ionising cosmic radiation (Guelin and Cernicharo 2022). To date, no amino acid has been detected in interstellar space. Trp has been recently claimed to be identified in the IC 348 star cluster of the Perseus molecular cloud, however, this discovery has been later disproven (Iglesias-Groth 2023; Hudson 2023). Even though the simplest amino acid, Gly, has yet to be identified in the interstellar medium, four plausible precursors of Gly (methylamine, formamide, glycolonitrile, and aminoacetonitrile) have been previously identified, which hints at the possibility of the discovery of interstellar Gly (Guelin and Cernicharo 2022; Tadeus et al. 2023).

The more likely scenario assumes formation of extraterrestrial amino acids by free radical reactions that occurred in the solid state, i.e. on the surface of interstellar ice, rather than the gas phase. The specific regions of interstellar medium called dense molecular clouds are heavily populated by carbon and silicate dust particles covered with amorphous ice mantles that adsorb and entrap atoms and simple molecular species protecting them against ionising cosmic radiation (Tsuge and Watanabe 2023). Adsorption of atoms and simple molecular species on the surface of interstellar ice ensures their accumulation, surface diffusion and collision resulting in formation of neutral inorganic (H_2O, CO, CO_2, NH_3) and simple organic (formaldehyde, methanol) molecules abundant in interstellar ice mantles (Tsuge and Watanabe 2023). Numerous laboratory simulations demonstrate that certain amino acids such as Gly, Ala, Val, Ser, Asp, Pro, β-alanine, sarcosine, 2-aminobutyric acid, 2,3-diaminopropionic acid, 2,3-diaminobutyric acid and 3,3'-diaminoisobutyric acid can be produced by UV irradiation of interstellar ice analogues via free radical solid-state reactions (Muñoz Caro *et al.* 2002; Oba *et al.* 2016; Modica *et al.* 2018). Icy dust grains containing amino acids could have aggregated into progenitors of comets and asteroids of the early solar system, which in turn delivered extraterrestrial amino acids to the prebiotic Earth (Modica *et al.* 2018; Potiszil *et al.* 2023). Despite the high feasibility of this scenario, UV irradiation of interstellar ice analogues does not provide sufficient chemical complexities and quantities of amino acids that can be found in meteorites (Modica *et al.* 2018; Potiszil *et al.* 2023). Hence, extraterrestrial amino acids had to be additionally formed on asteroids and comets due to the aqueous alteration, a change in

the mineralogical (and organic) composition of asteroids and comets caused by the exposure of primary anhydrous silicate minerals (olivine, pyroxene) to liquid water (Potiszil *et al.* 2023). The presence of secondary minerals such as phyllosilicate minerals and magnetite that can be produced through aqueous alteration serves as the evidence towards the former existence of liquid water on asteroids and comets (Brearley 2006). Amino acids could be produced during aqueous alteration by the same liquid-water reactions (HCN oligomerization, Strecker synthesis, Michael addition) that could give rise to terrestrial amino acids (Nakamura *et al.* 2022; Potiszil *et al.* 2023).

The most striking feature of amino acids found in carbonaceous chondrites is the enantiomeric excess of L-isomers, which is not observed during electric discharge experiments under simulated primordial Earth conditions. The L-enantiomeric excess (up to 15 % in the Murchison meteorite) is most notable for α-dialkyl amino acids such as isovaline, α-methylvaline, α-methylnorvaline, α-methylnorleucine which are rare or not found in terrestrial organisms and do not undergo spontaneous racemization in aqueous solutions due to the absence of hydrogen atom at the asymmetric centre (Bonner *et al.* 1979; Pizzarello and Cronin 2000). The higher abundance of L-isomers is also observed for proteinogenic amino acids found in meteorites but is disputable due to possible terrestrial contamination (Engel and Nagy 1982; Koga and Naraoka 2017). The enantiomeric excess of extraterrestrial L-amino acids could explain the universal homochirality of modern biologic systems, which represents one of the major unresolved problems in the origin of life research. The L-enantiomeric excess cannot be a consequence of abiotic synthesis but could be plausibly caused by the circular polarisation of UV light possibly originating from neutron stars, which could induce preferential absorption of the UV light by D-amino acids resulting in their photodegradation (Takano *et al.* 2007). However, the most recent enantiomeric analysis of amino acids from the Ryugu asteroid samples did not reveal statistically significant biases towards L-isomers, which contrasts with the observations from the carbonaceous chondrites (Parker *et al.* 2023).

Lastly, it is important to note that both terrestrial and extraterrestrial scenarios of amino acid formation are not exclusive to amino acids and lead to simultaneous formation of basically all building blocks of life including α-hydroxy and α-oxo acids, fatty acids, sugars and nucleobases (Kitadai and Maruyama 2018). Non-biological purine (unsubstituted purine, 2,6-diaminopurine, 6,8-diaminopurine, isoguanine) and pyrimidine (isocytosine) nucleobases of clear extraterrestrial origin have been identified in the carbonaceous chondrites, albeit in much lower quantities (parts per billion) than amino acids (parts per million) most likely due to their higher chemical complexity (Callahan *et al.* 2011; Oba *et al.* 2022). Extraterrestrial uracil and

its alkylated homologs have been recently found in the Ryugu asteroid samples in similar quantities (Oba *et al.* 2023). All biologically relevant nucleobases (cytosine, guanosine, uracil, thymine, adenine, xanthine, hypoxanthine) have been reported in interstellar ice simulations and electric discharge experiments (Ferus *et al.* 2017; Oba *et al.* 2019). The simultaneous formation of nucleobases, sugars, and amino acids in all these experiments is not a coincidence since their abiotic synthesis can occur via oligomerization and condensation of the common precursor, HCN (Ruiz-Bermejo *et al.* 2013). The similarity between nucleobase and amino acid abiotic synthetic pathways is striking and has led to the establishment of the cyanosulfidic protometabolism model, a hypothetical network of HCN oligomerization/reduction (mediated by H_2S) reactions that produce all major precursors of lipids, proteins and RNAs (Patel *et al.* 2015). The plausible conditions of the cyanosulfidic protometabolism reactions have not been fully elucidated, however, this model demonstrates putative simultaneous abiotic emergence of all building blocks of life.

1.2. Prebiotic condensation of amino acids into peptides

In contrast to the relative easiness of amino acid formation from simple inorganic precursors, their further condensation into linear peptides poses a significant thermodynamic and kinetic challenge due to the high free Gibbs and activation energies of peptide bond formation (Martin 1998). The simple heating of aqueous solutions of α-amino acids does not produce linear peptides in sufficient amounts since α-amino acids undergo cyclization into diketopiperazines or thermal decomposition by decarboxylation/deamination reactions (Nagayama *et al.* 1990). The formation of diketopiperazines is especially challenging since their high thermodynamic stability inhibits formation of oligomers longer than dipeptides (Parker *et al.* 2016). However, the trace amounts of the simplest dipeptide, glycylglycine, have been detected in electric discharge experiments indicating plausibility of the abiotic formation of peptides (Hirose *et al.* 1990; Miyakawa *et al.* 2002). A wide variety of physico-chemical factors and their combinations have been proposed to promote peptide formation on primordial Earth under plausible prebiotic conditions and will be described below.

1.2.1. Thermal approaches to prebiotic peptide synthesis

Condensation of amino acids to peptides is commonly known to be thermodynamically unfavourable in water due to formation of water as one of the reaction products. Hence, the simplest approach to shifting the equilibrium towards formation of peptides is to perform the reaction under water-scarce or dehydrating conditions. The earliest studies reported formation of small amounts of short peptides by simple dry melting of amino acids (Harada and Fox

1958; Vegotsky *et al.* 1958). In such conditions, formation of peptides requires the presence of acidic (Asp, Glu) or basic (Lys) amino acids that can undergo side chain-dependent intramolecular cyclization into reactive intermediates (lactams or succinimides) in a molten phase (Harada and Fox 1958; Vegotsky *et al.* 1958; Harada 1959). The side chain-independent formation of peptides can be achieved by thermal evaporation of amino acid solutions in the presence of deliquescent (water-absorbing) salts. Dipeptides and tripeptides can be formed in substantial yields by heating amino acids (Gly, Ala, Val, Leu, Pro, Asp, Glu, Lys, His) in highly concentrated NaCl solutions (Rode and Schwendinger 1990; Suwannachot and Rode 1999). At NaCl concentrations above 3M, the hydration shell surrounding sodium ions is not saturated with water molecules, which enables sodium ions to act as a strong dehydrating agent and overcome the high thermodynamic barrier of peptide bond formation (Suwannachot and Rode 1999). The kinetic limitations of peptide bond formation are reduced by addition of small amounts of transition metal salts such as $CuCl_2$, which serves as a catalyst (Rode and Schwendinger 1990).

The alternative approach relies on the repetitive temperature-induced dehydration of the reaction by wet-dry cycling. Wet-dry cycling involves iteration of heating and cooling steps that mimic dehydration and rehydration of the prebiotic Earth surface during day/night cycles (Lahav *et al.* 1978). Dehydration steps drive the condensation of amino acids into peptides whereas hydration steps enable efficient diffusion of amino acids for the following dehydration steps. Addition of small amounts of monovalent ion salts during wet-dry cycling simulations leads to their transient accumulation to very high concentrations during dehydration steps, which further promotes peptide formation by deliquescent (salt-induced) mechanism (Campbell *et al.* 2019). Addition of KCl/NaCl or K_2HPO_4/Na_2HPO_4 enables formation of much longer (up to 12-mer) oligomers from Gly with potassium ions being more efficient than sodium ions due to their higher deliquescence (Campbell *et al.* 2019). Formation of longer peptides (up to 6-mer for Gly) during wet-dry simulations can also be achieved by the addition of layered silicate minerals such as kaolinite and montmorillonite, which aid peptide elongation by adsorbing and protecting initially forming di- and tripeptides against hydrolysis as well as providing transition metal ions for catalysis (Lahav *et al.* 1978; Fuchida *et al.* 2014; Dalai *et al.* 2017). Similarly, wet-dry cycling drives condensation of imidazole-activated ribonucleotides (nucleoside phosphoimidazolides) into oligonucleotides up to 50 nucleotides long on the surface of silicate minerals (Riggi *et al.* 2023). Interestingly, the addition of aliphatic (Gly, Ala, Val, Leu) and acidic (Asp, Glu) amino acids was reported to greatly enhance montmorillonite-catalysed polymerization of adenosine 5'-phospho-2-methylimidazolide,

indicating possible cooperation between prebiotic amino acids and nucleotides towards higher-order oligomers (Namani *et al.* 2021).

1.2.2. Prebiotic peptide synthesis with organic condensation agents

Apart from dehydration-driven approaches, a wide range of organic and inorganic condensation agents of high prebiotic plausibility have been reported to induce peptide synthesis in aqueous solutions through formation of highly reactive amino acid derivatives. Organic condensation agents are represented by cyanamide and its derivatives (dicyanamide, dicyandiamide), diaminomaleonitrile and urea (Fig. 1.3). Cyanamide is formed abiotically by UV irradiation of diluted aqueous solutions of HCN or electron beam irradiation of CH_4/NH_3 gas mixtures (Schimpl *et al.* 1968). Upon mild heating, cyanamide undergoes dimerization into dicyandiamide, which in turn can undergo thermal decomposition producing dicyanamide (Schimpl *et al.* 1968). Cyanamide and its derivatives were the first condensation agents proposed for abiotic condensation of amino acids due to their tautomerization into carbodiimides, one of the first discovered coupling agents that are still extensively used in peptide synthesis (Sheehan and Hess 1955; Steinman *et al.* 1964). The cyanamide-mediated condensation of amino acids is thought to occur via their tautomerization into carbodiimides that react with amino acids producing highly reactive O-acylisoureas, which undergo condensation into peptides (Sheehan and Hess 1955). The cyanamide-mediated condensation of amino acids was proven to be highly insufficient producing dipeptides from Gly and Ala with very low conversion rates (up to 1 %) at highly acidic conditions and thus has a rather historical significance (Hulshof and Ponnamperuma 1976). On the other hand, diaminomaleonitrile, a tetramer of HCN, was shown to induce Gly dimerization under prebiotically plausible neutral or slightly alkaline conditions with sufficient yields (up to 4.8 %) and is putatively responsible for formation of the trace amounts of glycylglycine in the electric discharge experiments (Chang *et al.* 1969; Hulshof and Ponnamperuma 1976). Alternatively, amino acids can undergo condensation in the presence of urea, one of the most widely recognised abiotic condensation agents. Urea is abiotically produced in electric discharge experiments or by UV irradiation of aqueous solutions of ammonium cyanide (similarly to cyanamide) and has been shown to play a key role in abiotic synthesis of pyrimidines (cytosine, uracil) as well as abiotic phosphorylation reactions leading to formation of nucleotides (Lohrmann and Orgel 1971; Robertson and Miller 1995; Menor Salván *et al.* 2020). Reaction of urea with amino acids under moderate temperatures and alkaline pH leads to formation of N-carbamoyl amino acids, which can undergo intramolecular cyclization into highly reactive N-carboxyanhydrides that condense into dipeptides (Danger *et al.* 2006; Gan

et al. 2023). Formation of longer peptides (up to 7-mers) can be achieved by heating proteinogenic (Gly, Ala, Val) and non-proteinogenic (2-aminobutyric acid, β-alanine, norvaline, norleucine) amino acids in molten urea, however, the prebiotic plausibility of such an extreme water-scarce environment is questionable (Mita *et al.* 2005).

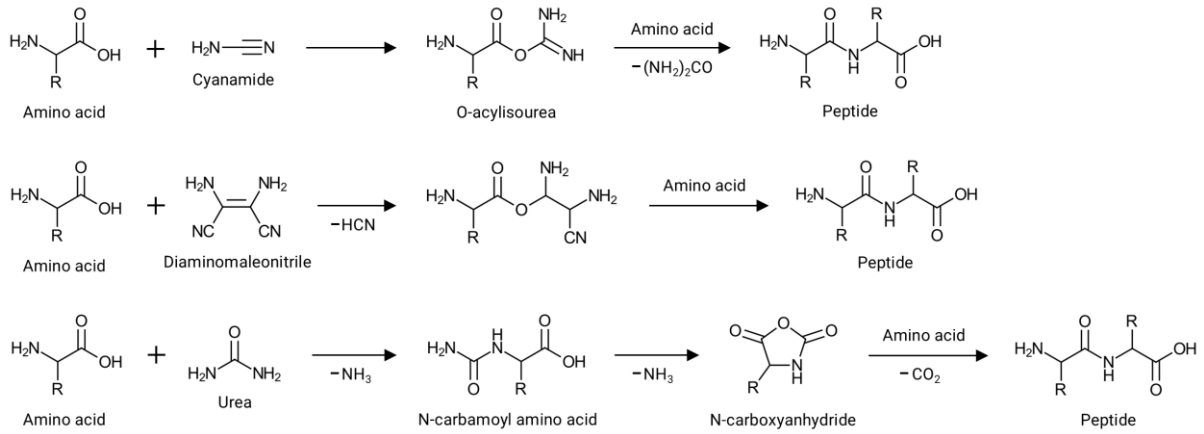


Fig. 1.3. The proposed reaction mechanisms of abiotic peptide syntheses mediated by organic condensation agents: cyanamide (Steinman *et al.* 1964), diaminomaleonitrile (Chang *et al.* 1969), and urea (Gan *et al.* 2023).

1.2.3. Prebiotic peptide synthesis with inorganic condensation agents

Inorganic condensation agents include various sulphur compounds (COS, CS_2, SO_2), polyphosphates and diamidophosphate (Fig. 1.4). Carbonyl sulphide (COS), carbon disulphide (CS_2), and sulphur dioxide (SO_2) are simple inorganic compounds produced in large amounts by volcanic emissions and weathering of sulphide minerals (Chen and Yang 2007; Leman *et al.* 2004; Leman *et al.* 2015). Passing COS or CS_2 through aqueous solutions of amino acids (Leu, Ser, Ala, Tyr, Phe) results in formation of aminoacyl thiocarbamates and dithiocarbamates that undergo intramolecular cyclization into respective N-carboxyanhydrides and 2-thiono-5-oxazolidones under mildly alkaline conditions (Leman *et al.* 2004; Leman *et al.* 2015). The resulting N-carboxyanhydrides and 2-thiono-5-oxazolidones are highly reactive and undergo condensation, forming di-, tri- and tetrapeptides in sufficient yields reaching up to 20 %. The peptide formation is facilitated by the addition of stoichiometric amounts of transition metal salts ($PbCl_2$, $CdCl_2$, $FeCl_2$) or oxidising agents ($K_3[Fe(CN)_6]$) that facilitate removal of the thiolate (RS-) group and improve the overall yield up to 80% (Leman *et al.* 2004; Leman *et al.* 2015). Potassium ferrocyanide ($K_3[Fe(CN)_6]$), a potent oxidising agent of high prebiotic plausibility, was also shown to induce oxidation of sulphur dioxide (SO_2) to sulphur trioxide (SO_3) in reducing conditions of the primitive Earth atmosphere, i.e. the absence of atmospheric oxygen (Chen and Yang 2007). The resulting sulphur trioxide can

mediate amino acid (Ala, Ser, Phe) condensation through formation of reactive carboxylic sulphuric anhydrides that produce dipeptides in relatively lower yields (~5%).

The more attractive alternative to sulphur compounds is inorganic polyphosphates. Inorganic polyphosphates harbour energy-rich phosphoanhydride bonds that can be used for abiotic activation of amino acids as well as phosphorylation of other biological precursors such as nucleosides, glycerol and fatty acids. Short linear (pyrophosphate, tripolyphosphate, tetrapolyphosphate) and cyclic (trimetaphosphate) polyphosphates can be abiotically produced by the action of volcanic magma on water-insoluble terrestrial phosphate minerals (apatite, $Ca_5(PO_4)_3(OH)$) (endogenous origin) or by the aqueous corrosion of meteorite phosphide mineral, schreibersite ($(Fe,Ni)_3P$) (exogenous origin) (Yamagata *et al.* 1991; Pasek *et al.* 2008). Mild heating of short polyphosphates with simple amino acids (Gly and Ala) leads to formation of linear phosphoramidates that are hydrolysed into highly reactive cyclic phosphoramidates, which subsequently undergo condensation into respective dipeptides (Rabinowitz *et al.* 1969; Chung *et al.* 1971). Simple dipeptides such as glycylglycine can undergo further oligomerization into longer (up to 6-mer) oligopeptides (Yamagata and Inomata 1997; Sibilska *et al.* 2017). Polyphosphate-induced formation of peptides takes place over a wide range of temperatures and pH (including ambient temperature and neutral pH) and is further promoted by the addition of magnesium ions that neutralise electrostatic repulsion between negatively charged phosphate residues (Yamagata and Inomata 1997; Sibilska *et al.* 2018). The efficiency of phosphate-driven condensation reactions can be further improved if inorganic polyphosphates are first converted into highly reactive diamidophosphate by reaction with aqueous ammonia (Gibard *et al.* 2018). Mixing of amino acids (Gly, Asp, Glu) with diamidophosphate and imidazole results in formation of acylphosphoramidates that undergo subsequent oligomerization into short (up to 8-mer) peptides with high overall yields (up to 23%). Diamidophosphate has gained increasing attention as a potent condensation agent due to its ability to not only phosphorylate biological precursors but also drive their polymerization or self-assembly into higher-order structures. Unlike inorganic polyphosphates, phosphorylation of nucleosides with diamidophosphate is regioselective and leads to formation of 2',3'-cyclic nucleotide monophosphates that readily polymerize into short (up to 12-mer) oligonucleotides with high conversion rates (up to 66 %). Mild heating of glycerol and nonanoic (C9) acid in the presence of diamidophosphate leads to formation of cyclic phospholipids that can self-assemble into bilayer vesicles, mimicking cell membranes, as well as other multilamellar structures. Although the reactions require high concentrations of diamidophosphate and minimal presence of water, diamidophosphate-driven condensation

demonstrates the plausibility of simultaneous abiotic formation of all three major classes of biomolecules (lipids, proteins, nucleic acids) under the same reaction conditions.

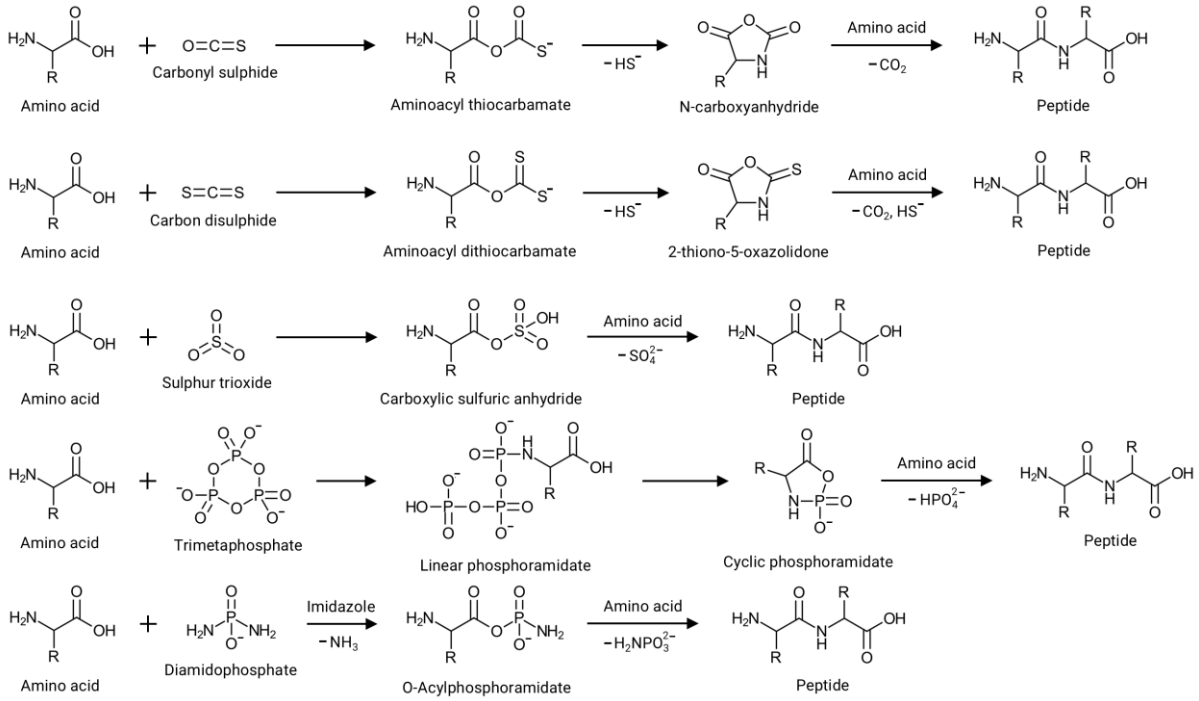


Fig. 1.4. The proposed reaction mechanisms of abiotic peptide syntheses mediated by inorganic condensation agents: carbonyl sulphide (Leman *et al.* 2004), carbon disulphide (Leman *et al.* 2015), sulphur trioxide (Chen and Yang 2007), trimetaphosphate (Chung *et al.* 1971), and diamidophosphate (Gibard *et al.* 2018).

1.2.4. Abiotic co-condensation of amino acids with other biological precursors

Peptides could also be synthesised by symbiotic cooperation of amino acids with other biological precursors (Fig. 1.5). One highly acclaimed hypothesis proposes co-condensation of amino acids with hydroxy acids, which are also formed in electric discharge experiments and carbonaceous chondrites (Forsythe *et al.* 2015). Dry-wet cycling of α-amino acids with two simplest α-hydroxy acids (glycolic and lactic acids) under acidic conditions leads to formation of depsipeptides, in which building blocks are connected by alternating amide (peptide) and ester bonds (Forsythe *et al.* 2015; Yu *et al.* 2017; Doran *et al.* 2019). The reaction mechanism implies the dehydration-induced condensation of two α-hydroxy acids into an ester, which is subsequently attacked by an amino acid resulting in the exchange of an ester bond to an amide (peptide) bond (Forsythe *et al.* 2015). Formation of depsipeptides follows the reaction mechanism similar to translation, which does not occur by direct condensation of free amino acids but by the ester-amide bond exchange between aminoacyl-tRNAs. The ester-amide bond exchange (or ester aminolysis) is thermodynamically more favourable than peptide bond formation between two amino acids and results in formation of longer (up to 14-mer) oligomers

in substantial yields (up to 50 %) (Forsythe *et al.* 2015). Ester bonds are more liable towards hydrolysis, which causes their eventual breakdown during hydration steps and the gradual enrichment of the newly forming depsipeptides with peptide bonds during the subsequent wet-dry cycles. Similarly, amino acids can undergo co-condensation with α-mercapto acids such as thioglycolic acid, which leads to formation of thiodepsipeptides (Frenkel-Pinter *et al.* 2022). Analogously to depsipeptides, the reaction mechanism involves formation of a thioester bond between the thiol group of α-mercapto acid and the carboxyl group of an amino acid, which is subsequently attacked by another amino acid leading to formation of a peptide bond. Formation of thiodepsipeptides is achieved at a wider pH range (from acidic to neutral), milder temperatures and lower reactant concentrations due to the higher reactivity (nucleophilicity) of α-mercapto acids compared to α-hydroxy acids.

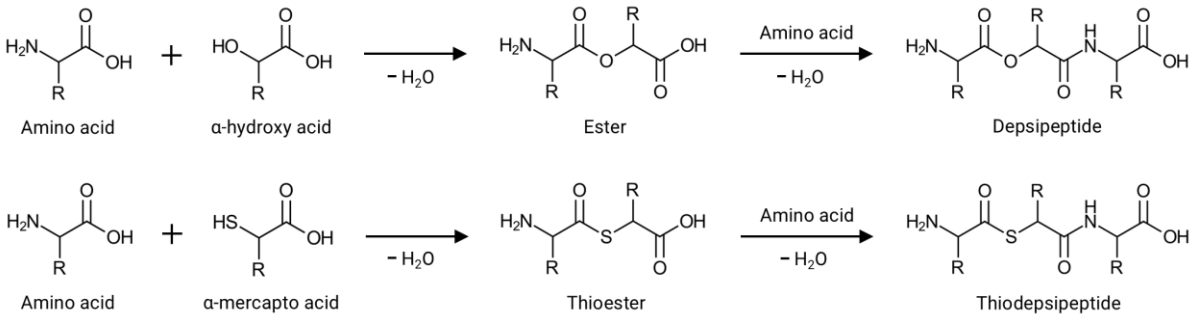


Fig. 1.5. Co-condensation of amino acids with α-hydroxy acids and α-mercapto acids, leading to formation of depsipeptides (Forsythe *et al.* 2015) and thiodepsipeptides (Frenkel-Pinter *et al.* 2022). Iterative wet-dry cycling leads to the gradual enrichment of depsipeptides and thiopeptides with peptide bonds over ester/thioester bonds due to their lower stability.

Alternatively, Ying *et al.* (2018) showed that trimetaphosphate can mediate non-specific condensation of neutral amino acids (Val, Leu, Phe, Tyr, Trp) with adenosine leading to formation of the isomeric (2'-, 3'-, 5'-) mixture of aminoacyl adenylates. The mixture of isomeric aminoacyl adenylates was also shown to be produced by the reaction of adenosine monophosphates with N-carboxyanhydrides formed by urea- or COS-mediated condensation of amino acids (Biron *et al.* 2005). The resulting aminoacyl adenylates can undergo further condensation yielding small amounts (less than 2 %) of dipeptides (Ying *et al.* 2018). The condensation of aminoacyl adenylates into oligopeptides is greatly enhanced by layered silicate minerals (Paecht-Horowitz 1974). Trimetaphosphate-induced condensation is not exclusive to adenosine and can be performed with any of the four canonical nucleosides (uridine, cytidine, guanosine) leading to formation of the respective nucleotide amidates (Ying *et al.* 2018). This scenario suggests the abiotic origin of aminoacylation since 5'-aminoacyl adenylate is

synthesised as a key intermediate of amino acid activation by aminoacyl-tRNA synthetases in modern living organisms.

All the aforementioned scenarios of amino acid condensation imply the terrestrial formation of short peptides on primordial Earth from endogenous or exogenous amino acids in their free form. Oligomerization (up to 6-mer) of Gly was also achieved by thermal cycling of its highly concentrated solutions under high pressure that resembles the hydrothermal circulation of seawater (Imai *et al.* 1999). The trace amounts (parts per trillion) of glycylglycine of extraterrestrial origin were also detected in the carbonaceous chondrites (Shimoyama and Ogasawara 2002). The extraterrestrial formation of glycylglycine has been recently suggested to occur not by condensation of Gly but via polymerization of aminoketene (NH_2CHCO) that is formed by the reaction between atomic carbon, CO and NH_3 on the surface of interstellar icy dust grains (Krasnokutski *et al.* 2022).

1.3. The role of peptides in the origin of life

The ever-growing evidence supporting the high feasibility of the abiotic formation and condensation of amino acids indicates that short peptides existed at the earliest stages of life evolution prior to the introduction of proteins, which could not be formed by simple condensation reactions. Similarly to proteins, short peptides are capable of performing various biological functions including (but not limited to) molecular recognition (binding), compartmentalization, and catalysis, which would be sufficient for formation of a proto-cellular system. The prebiotically relevant functionalities of short peptides will be described in this section.

1.3.1. Catalytic properties of short peptides

The ability of short peptides to catalyse a wide variety of chemical reactions either in free form or in complex with metal ions and organic cofactors is widely recognised. Historically, the two originally proposed scenarios of origin of life assigned the role of abiotic catalysts to either minerals and metal ions (metabolism-first scenario) or ribozymes and nucleotide-derived cofactors (RNA-first scenario) (Gilbert 1986; Wächtershäuser 1988). Considering this fact, the plausible involvement of peptides in proto-metabolism has not been extensively highlighted in literature, even though it is generally believed that first functional proteins emerged by fusion, duplication and recombination of primordial peptides (Söding and Lupas 2003; van der Gulik *et al.* 2009; Kolodny *et al.* 2021). Some of these primordial peptides can still be identified as short protein segments that are repeated within the same protein chain

(such as ancient double β-hairpin) or found across different protein families (such as ancestral KH peptide) (Remmert *et al.* 2010; Pereira and Lupas 2018).

In most extreme cases, the evolutionary origin of an enzyme family can be traced back to a few highly conserved catalytic residues as illustrated by the case of Ser-His, the shortest catalytic peptide, which has been extensively studied in the context of origin of life as a plausible evolutionary ancestor of modern hydrolases. Ser and His (along with Asp) residues form a highly conserved catalytic triad that can be found in the active sites of serine proteases, esterases, and lipases (Li *et al.* 2000). Li *et al.* (2000) demonstrated that Ser-His-Asp tripeptide as well as its shorter analogue, Ser-His dipeptide, display proteolytic, nucleolytic and esterase activities over a wide range of pH and temperatures. The same range of catalytic activities, albeit inferior, was also observed for a panel of related peptides (Cys-His, Ser-His-Gly, Ser-His-His) but not for the structural isomer of Ser-His, His-Ser. The further studies revealed that Ser-His is able to promote reverse reactions and catalyse dimerization of imidazole-activated GMP into GppG and pGpG oligonucleotides (Wieczorek *et al.* 2013). The reaction is carried out in eutectic water/ice mixtures, in which most of the water is trapped in the form of ice crystals thus increasing the local concentration of guanosine-5'-imidazole monophosphate and shifting the thermodynamic equilibrium towards phosphodiester (and phosphoanhydride) bond formation. Ser-His was also shown to catalyse condensation of N-acetylated amino acid esters into respective dipeptides (Gorlero *et al.* 2009). The reaction is facilitated by N-acetylation and O-esterification of hydrophobic amino acids (Phe, Leu) that increase the overall hydrophobicity of the resulting dipeptides (Ac-Phe-Leu-NH_2 and Ac-Phe-Phe-NH_2), leading to their precipitation from solution and thus shifting the thermodynamic equilibrium towards peptide bond formation. The same reaction performed in the presence of oleic acid vesicles (used as model prebiotic protocells) leads to symbiotic cooperation between Ser-His and oleic acid vesicles – the resulting hydrophobic product (Ac-Phe-Leu-NH_2) dissolves in vesicle membranes, which increases the catalytic efficiency of Ser-His and simultaneously promotes the vesicle growth by decreasing the fluidity of oleic acid membranes (Adamala and Szostak 2013). Interestingly, Ser-His was shown to be exclusively synthesised over similar dipeptides (Ser-Ala, Ser-Ser, Ser-Pro, Ser-Asp) from the simultaneous mixture of *N*-diisopropyl phosphoryl serine with all five respective amino acids (Ala, Ser, Pro, Asp, and His) under acidic or neutral conditions (Shu *et al.* 2018). The sole formation of Ser-His is thought to be promoted by hydrogen bonding between the hydroxyl group of Ser and the imidazolyl ring of His.

The catalytic versatility of His-containing dipeptides has been further illustrated by Maguire *et al.* (2024) who developed a plausible prebiotic phosphate transfer system consisting

of imidazole phosphate and histidyl peptides. Imidazole phosphate can be abiotically produced from the mixture of imidazole, cyanate, and orthophosphate under very mild conditions (22 °C, pH 7.3) and has been proposed as a prebiotic phosphate donor analogue of ATP (Maguire *et al.* 2021). Wet-dry cycling of the mixture containing imidazole phosphate, glycerol, and His-containing peptides (His-Asp, His-Lys, His-Gly-Gly) under the same ambient conditions leads to formation of three phosphorylated glycerol species, namely glycerol-1-phosphate, glycerol-2-phosphate, and cyclo-glycerol-1,2-phosphate, with the high (up to 50 %) overall yield. The reaction mechanism involves the transfer of phosphate group from imidazole phosphate to the imidazolyl ring of His residue, which subsequently transfers phosphate group to glycerol forming glycerol phosphates.

Considering the general non-specificity of abiotic condensation reactions, prebiotic formation of specific His-containing catalytically active peptides is stochastically a less likely process. Weber and Pizzarello (2006) demonstrated that various di- and tripeptides composed of Gly, Ala, Val, Ile, Leu, and Glu can act as asymmetric catalysts in intermolecular aldol condensation of glycolaldehyde leading to the preferential formation of D-erythrose over L-erythrose with enantiomeric excesses reaching up to 82 %. Possibly, the strongest support towards abiotically formed proto-catalytic peptides comes from a recent study showing that wet-dry cycling of a mixture of Leu, Glu, Lys and glycolic acid leads to the crude mixture of short depsipeptides that exhibit esterase-like activity as evidenced by their ability to hydrolyse p-nitrophenyl acetate (pNPA) (Doran *et al.* 2019). The bulk structural characterization of the depsipeptide mixture by far-UV circular dichroism (CD) spectroscopy revealed the presence of a substantial population of β-sheet structure.

1.3.2. Self-assembly of short peptides

Short peptides display inferior catalytic activity and weaker molecular recognition than proteins due to their higher conformational flexibility and lower surface area. Both of these drawbacks can be effectively overcome by the self-assembly of short peptides into diverse rigid higher-order architectures such as helix bundles, amyloid fibrils, and other motifs by non-covalent interactions between amino acid residues (Zozulia *et al.* 2018). Self-assembly of short peptides has gained recent attention in the origin of life research as a possible prebiotic path towards self-replicating catalytic molecular entities. Analogously to self-ligating ribozymes developed in the support of the “RNA world” hypothesis, self-replication of a peptide (in its simplest form) can be demonstrated by the ability to catalyse peptide bond formation between its two shorter fragments. Self-replication of short peptides was first demonstrated by Lee *et al.* (1996), who derived an autocatalytic α-helical peptide from the leucine-zipper domain of

the yeast transcription factor GCN4. This peptide could template its own synthesis from two thiol-activated fragments by formation of a coiled coil, which brings two fragments in close proximity to each other and accelerates peptide bond formation between them up to 500-fold. Following this discovery, other α-helical peptides utilising coiled-coil dimerization for their self-replication from thiol-activated fragments have been described in literature (Yao *et al.* 1997; Saghatelian *et al.* 2001; Altay *et al.* 2018). Interestingly, Saghatelian *et al.* (2001) demonstrated that α-helical peptide replicator is able to selectively catalyse formation of homochiral products from a racemic mixture of peptide fragments, providing a plausible mechanism for the origin of homochirality.

The same fragment-ligation strategy has been employed for the development of self-replicating systems based on amyloidogenic peptides (Takahashi and Mihara 2004; Rubinov *et al.* 2009). Amyloid aggregation itself can be viewed as a particular case of self-replication, in which misfolded or partially folded monomeric peptide/protein molecules first associate into a "nucleus" (primary nucleation) that grows and elongates into a fibril by recruiting and conferring the same secondary structure to other monomeric molecules and at the same time promotes formation of other "nuclei" on the fibril surface (secondary nucleation) (Riek and Eisenberg 2016). The resulting amyloid fibrils are highly ordered peptide/protein aggregates characterised by the cross-β-sheet conformation that consists of two tightly interacting β-sheet layers stabilised by various noncovalent interactions (Nelson *et al.* 2005) (Fig. 1.6). Each β-sheet layer consists of thousands of identical β-strand copies that are densely packed via extensive intermolecular hydrogen bonding and lie perpendicular to the fibril axis. The cross-β-sheet conformation of amyloid fibrils confers them extraordinary stability and resistance towards hash environmental conditions such as high temperatures and hydrostatic pressure, extreme pH conditions and high salt concentrations (Meersman and Dobson 2006; Friedmann *et al.* 2015). The lack of sequence and structural homology in amyloidogenic proteins as well as the recent studies characterising amyloidogenic propensities of *de novo* designed sequences imply that amyloid aggregation is a phenomenon universal to nearly all peptides and proteins, albeit in many cases amyloid fibrils are formed under strong perturbations such as heating, sonication, shaking, or extreme pH (Chiti *et al.* 1999; Fowler *et al.* 2007; Meisl *et al.* 2022).

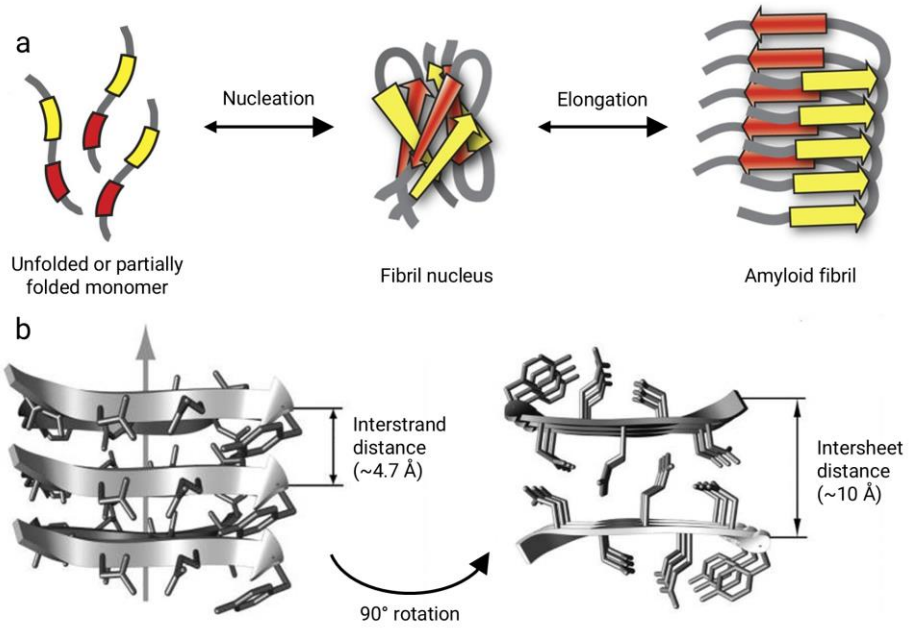


Fig. 1.6. (a) The simplified "nucleation and growth" model of amyloid fibril formation; (b) The side and top views of the cross-β core structure of the amyloidogenic peptide fragment from the yeast protein Sup35. Adopted and modified from Tiller and Tessier (2013) and Fändrich *et al.* (2009)

Initially discovered and extensively studied in the context of human neurodegenerative diseases, amyloidogenic proteins were shown to carry out diverse biological functions across all kingdoms of life (Fowler *et al.* 2007). Some of the examples include bacterial CsgA protein (the main structural component of curli, a proteinaceous fibrillar matrix that enables surface adhesion and colony formation in *Escherichia coli*) or mammalian Pmel17 protein (the structural monomer of Mα amyloid that catalyses polymerization of indole-5,6-quinone into melanin in mammals) (Chapman *et al.* 2002; Fowler *et al.* 2006). The amyloidogenic propensity of naturally occurring polypeptides can be usually scaled down to a few core amino acid residues as demonstrated by the case of Phe19-Phe20, the shortest amyloidogenic fragment of amyloid-β peptide responsible for Alzheimer disease (Reches and Gazit 2003). Numerous studies have shown that even very short non-active peptides can assemble into amyloid fibrils harbouring different functional properties due to their high conformational rigidity and structural periodicity (Riek and Eisenberg 2016). Amyloid fibrils assembled from non-catalytic peptides, ranging from 7 to 13 residues in length, can catalyse a wide range of reactions such as tetrose/pentose formation, ester and phosphoester hydrolysis, CO_2 hydration, ATP hydrolysis, and phenol oxidation (Rufo *et al.* 2014; Makhlynets *et al.* 2016; Tena-Solsona *et al.* 2016; Al-Garawi *et al.* 2017; Monasterio *et al.* 2017). In most cases, amyloid fibrils require coordination with transition metal ions (Mn^{2+}, Zn^{2+}, Cu^{2+}) for their catalytic activity and exhibit Michaelis-Menten saturation kinetics, similarly to enzymes. The catalytic amyloid

fibril-metal complexes can be assembled even from a single amino acid, Phe, which forms amyloid-like nanofibrillar structures from millimolar solutions spontaneously or after co-precipitation with zinc ions (Adler-Abramovich *et al.* 2012; Makam *et al.* 2019). The resulting Phe-Zn(II) crystals display esterase-like and carbonic anhydrase-like activities, as evidenced by pNPA hydrolysis and CO_2 hydration experiments (Makam *et al.* 2019). The numerous observations on both self-replicative and catalytic properties of amyloid fibrils led to the formulation of the "amyloid world" hypothesis, according to which structurally stable self-replicating and catalytic amyloidogenic peptides served as precursors of life that preceded or co-existed with self-replicating and catalytic RNA molecules (Maury 2018).

1.3.3. Compartmentalisation and molecular recognition by short peptides

Besides self-replication and metabolism, the emergence of cellular life would not be possible without formation of compartmentalised structures. In fact, depending on the conditions, short fragments of amyloid-β peptide (scaling down to Phe19-Phe20) were shown to assemble into amyloid fibrils, insoluble hollow nanotubes or the mixture of both (Reches and Gazit 2003; Dong *et al.* 2006; Mehta *et al.* 2008). At neutral pH, the amyloid-β (16–22) fragment, KLVFFAE, forms a mixture of amyloid fibrils and nanotubes that have a uniform diameter of 52 nm and the wall thickness of 4.2 nm (Mehta *et al.* 2008). The nanotube walls are built of two stacked antiparallel β-sheet layers with Lys residues both placed at the solvent exposed interface and buried in the bilayer interface (Mehta *et al.* 2008). Coordination of two alanine mutants of amyloid-β (13–21) fragment, HAQKLVFFA (H14A) and HHQALVFFA (K16A), with zinc ions results in formation of larger nanotubes with an average diameter of 200–300 nm that are composed of β-sheet layers stacked via intermolecular His-Zn(II)-His bridges (Dong *et al.* 2006). Apart from amyloidogenic peptides, two different groups of amphiphilic peptides are known to assemble into phospholipid membrane-like structures. Peptides with alternating hydrophobic and hydrophilic residues such as poly(Lys-Val), poly(Leu-Lys), and poly(Leu-Glu) form soluble membrane-like β-sheet bilayers, in which hydrophobic residues are sequestered inside the bilayer and hydrophilic residues are exposed to solution, similarly to phospholipids (Brack and Orgel 1975; Brack 1977). Short (7- to 8-mer) surfactant-like peptides composed of one or two negatively or positively charged residues (Asp, Glu, Lys, Arg, His) and a stretch of hydrophobic residues (Gly, Ala, Val, Leu, Ile) can form a variety of monolayer (nanorods, nanofibers, and micelles of 5–10 nm in size) and bilayer (nanotubes and nanovesicles of 30–50 nm in size) structures in solution (Santoso *et al.* 2002; Vauthey *et al.* 2002; von Maltzahn *et al.* 2003; Qiu *et al.* 2009). The resulting nanostructures were shown to stabilise certain membrane proteins such as photosynthetic complex I, G

protein-coupled receptor bovine rhodopsin, glycerol-3-phosphate dehydrogenase (Das *et al.* 2004; Yeh *et al.* 2005; Zhao *et al.* 2006).

All the aforementioned discoveries confirm that short peptides can achieve remarkable functionality without requiring high chemical complexity. Despite the fact that short peptides can fulfil all the essential requirements for the emergence of life (self-replication, metabolism and compartmentalization), formation of a proto-cellular molecular system based solely on peptides is plausible but unlikely. Rather than being stand-alone biological precursors, short peptides and their assemblies were more likely engaged in complex chemical networks with other prebiotically available molecules such as fatty acids and oligonucleotides. In the prebiotic absence of phospholipids, protocell membranes could be composed of prebiotically available fatty acids that form bilayer vesicles susceptible to extreme pH, high ionic strength and divalent metal ions (Mg^{2+}) (Budin *et al.* 2012). Leu-containing dipeptides (as well as single hydrophobic amino acids) were shown to bind to decanoic acid vesicles, accelerating their growth and improving their stability against divalent metal ions and high salt concentrations (Cornell *et al.* 2019; Cohen *et al.* 2024). In turn, oleic acid vesicles were shown to effectively encapsulate and enhance formation of amyloidogenic peptides from the stoichiometric mixtures of activated amino acids due to their selective permeability (Kwiatkowski *et al.* 2021). The countless studies on ribozymes led to the universal acclaim of the involvement of catalytic RNA molecules (ribozymes) in proto-cellular replication and metabolism, both of which could be enhanced or modulated by interaction with short cationic peptides (Tagami *et al.* 2017; Drobot *et al.* 2018; Le Vay *et al.* 2021; Li *et al.* 2022). The symbiotic cooperation between peptides and RNA was also essential for the further protein evolution ultimately leading to the emergence of the ribosome that will be discussed in the next section.

1.4. Formation of proto-ribosome and early translation machinery

Peptide bonds can be formed in living organisms by ribosomes or non-ribosomal peptide synthetases. Even though non-ribosomal peptide synthetases have a broader activity, which enables the synthesis of linear, branched, and circular peptides from both proteinogenic and non-proteinogenic amino acids, their prebiotic relevance is unlikely due to the non-universal phylogenetic distribution and a high modular complexity required for peptide formation (Finking and Marahiel 2004). On the other hand, ribosomes are large ribonucleoprotein complexes that are responsible for synthesis of all cellular proteins and are universally conserved across all domains of life (Mears *et al.* 2002; Fox 2010; Bernier *et al.* 2018). The ribosome is composed of at least three ribosomal RNAs (rRNAs) and more than 50

ribosomal proteins that together constitute the large and small ribosomal subunits. The large ribosomal subunit (LSU) contains the peptidyl transferase centre (PTC) that catalyses peptide bond synthesis, and the small ribosomal subunit (SSU) contains the decoding centre (DC) that is responsible for recognition of codons in mRNA. The two main functional regions of the ribosome (PTC and DC) are composed solely of highly conserved rRNA and are considered to be among the oldest structures in biology (Nissen *et al.* 2000; Ogle *et al.* 2001; Davidovich *et al.* 2009; Fox 2010). Even though the details of the origin and early evolution of the translation machinery remain vastly unclear, it is generally believed that the LSU and SSU evolved independently from each other followed by the late acquisition of the subunit interface by the time of the LUCA (Smith *et al.* 2008; Fox 2010; Petrov *et al.* 2015).

The PTC is arguably the oldest component of the modern ribosome that emerged from ancestral RNA molecules even prior to the LUCA (Smith *et al.* 2008; Davidovich *et al.* 2009; Bokov and Steinberg 2009; Fox 2010; Petrov *et al.* 2015), albeit an alternative view supporting the earlier origin of the DC exists (Harish and Caetano-Anollés 2012). In prokaryotes, the PTC is situated in the domain V of 23S rRNA and consists of 180 highly conserved nucleotides that represent approximately 6 % of bacterial 23S rRNA (Agmon *et al.* 2005; Prosdocimi *et al.* 2020) (Fig. 1.7). The universally conserved PTC structure is characterised by distinct two-fold rotational quasi-symmetry, which led to a very popular hypothesis on its origin, according to which the PTC emerged by dimerization, association or fusion of two smaller ancestral RNA molecules (Agmon *et al.* 2005; Belousoff *et al.* 2010; Tamura 2011). These ancestral RNA molecules are widely speculated to be amino acid-bearing proto-tRNAs, whose association or fusion could bring activated amino acids together for the non-coded formation of short random peptide sequences (Tamura 2011; Torres de Farias *et al.* 2014; Demongeot and Seligmann 2022). This hypothesis has been supported by several computational studies inferring structural or sequence similarity between rRNAs and tRNAs. Tamura (2011) demonstrated a topological similarity between the secondary structure of the PTC region of 23S rRNA from *Deinococcus radiodurans* and the canonical "cloverleaf" secondary structure of tRNAs. Following this observation, Torres de Farias *et al.* (2013) performed ancestral tRNA sequence reconstruction by applying maximum likelihood analysis to an extensive dataset of 9758 tRNA sequences from 361 organisms distributed in all three kingdoms of life. The inferred ancestral tRNA sequences were aligned individually against the PTC region of 23S rRNA from *Thermus thermophilus*, which led to identification of the concatemer of ancestral tRNAs (tRNAPTCancestor) that shared 50.5 % sequence and 92.7 % three-dimensional structural identity with the PTC from *Thermus Thermophilus* (Torres de Farias *et al.* 2014; Torres de

Farias *et al.* 2017). The sequence alignment of 5S, 16S, and 23S rRNAs with the core transcription and translation components of *Escherichia coli*, performed by Root-Bernstein and Root-Bernstein (2015), revealed that the sequences homologous (>50 % identity) to all tRNAs can be found along or are complementary to the sequences of 16S and 23S rRNAs, implying that rRNAs most likely emerged by concatenation and accretion of ancestral tRNAs. Lastly, Demongeot and Seligmann (2022) pointed out the resemblance between the secondary structures of tRNA sub-elements and the most ancient regions of 16S and 23S rRNAs supporting the proto-tRNA concatenation/accretion hypothesis. The latter two observations lead to a broader speculation, according to which rRNAs may represent "molecular fossils" of an ancestral self-replicating RNA genome that encoded all components of transcription and translation machinery, including tRNAs (Root-Bernstein and Root-Bernstein 2015; Demongeot and Seligmann 2022).

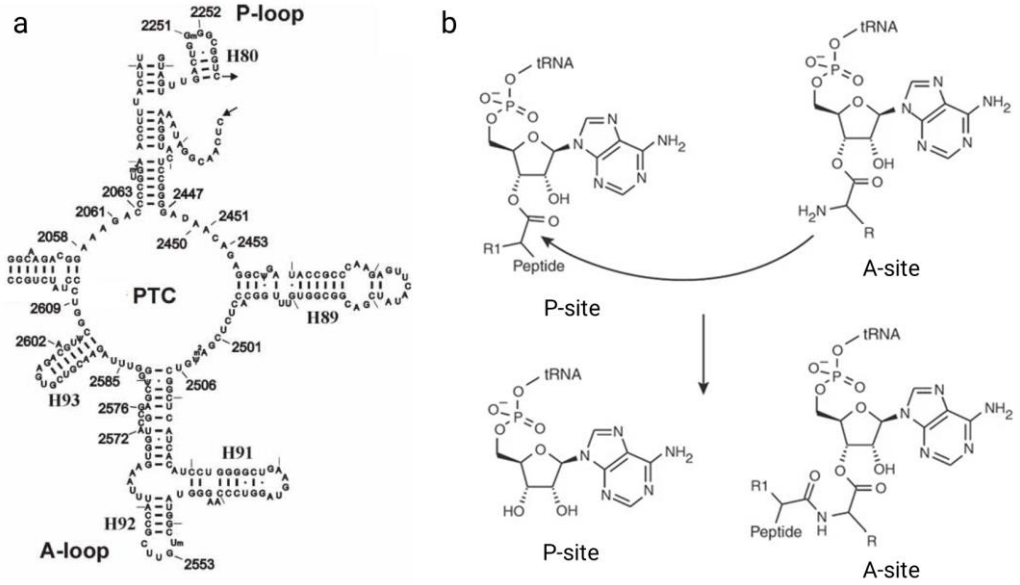


Fig. 1.7. (a) The secondary structure of the peptidyl transferase loop of domain V of 23S rRNA from *Escherichia coli*; (b) The peptidyl transfer reaction between aminoacyl-tRNA and peptidyl-tRNA catalysed by the PTC. Adopted and modified from Polacek and Mankin (2005) and Bieling *et al.* (2006).

There are several hypotheses on the possible structure of ancestral tRNAs that predated both modern tRNAs and the PTC. tRNAs fold into an L-shaped tertiary structure that consists of two distinct domains: acceptor domain is formed by stacking of acceptor stem and TΨC arm into a minihelix and contains the amino acid attachment site (3'-CCA end), whereas the anticodon domain is formed by stacking of D and anticodon arms and contains anticodon (Schimmel and Ribas de Pouplana 1995). In the absence of the coding system and codon-anticodon recognition, ancestral tRNAs were most likely smaller and consisted solely of the acceptor-TΨC minihelix that roughly constitutes a half of tRNA (Schimmel and Ribas de

Pouplana 1995; Sun and Caetano-Anollés 2008; Tamura 2011). The acceptor-TΨC minihelices have been shown to be sufficient substrates for aminoacyl-tRNA synthetases and even preferred over the full-length tRNAs by urzymes, the plausible evolutionary ancestors of aminoacyl-tRNA synthetases (Francklyn and Schimmel 1989; Tang *et al.* 2024). The resulting aminoacyl-minihelices can be effectively used for peptide bond synthesis catalysed by the 50S ribosomal subunit (Sardesai *et al.* 1999). Furthermore, Tamura and Schimmel (2001) demonstrated the ribosome-free peptidyl transfer reaction between the 35-mer minihelix, derived from $tRNA^{Ala}$ from *Escherichia coli* and charged with N-acetyl-L-alanine (mimicking peptidyl-tRNA), and the complementary short puromycin-bearing oligonucleotide, 5'-UGGU-3' (mimicking aminoacyl-tRNA). The reaction is facilitated by the base-pairing between the 3'-CCA end of the minihelix and the puromycin-containing oligonucleotide and requires imidazole as a catalyst. The alternative hypothesis suggests that tRNAs emerged by the direct duplication of a short RNA hairpin (Di Giulio 1992; Dick and Schamel 1995). The duplication hypothesis is supported by the inherent quasi-symmetry of the canonical "cloverleaf" secondary structure and the apparent sequence similarity/homology between 5' and 3' halves of tRNA separated by anticodon (Di Giulio 1992; Widmann *et al.* 2005; Zuo *et al.* 2013). Another alternative model suggests formation of tRNA by ligation of three 31-nt minihelices that consisted of the universal acceptor stem and 16-nt microhelices containing D-, T- or acceptor loops (Burton 2020). All the aforementioned hypotheses can be supported by the discovery of split tRNA genes that separately encode two or three fragments of several archaeal tRNAs (Randau *et al.* 2005; Fujishima *et al.* 2009). Lastly, in their earliest form, ancestral tRNAs could also be represented by single nucleotides (Lacey *et al.* 1975; Jash *et al.* 2021). All these theories are not entirely mutually exclusive and could represent different stages of tRNA evolution.

Ancestral tRNAs could be coupled with amino acids even in the absence of protein-based aminoacyl-tRNA synthetases. The direct non-catalysed reaction between a free amino acid and tRNA resulting in formation of aminoacyl-tRNA is possible but requires extremely high pressure of 600 MPa (Krzyzaniak *et al.* 1994). Biotic aminoacylation is a two-step process that involves (i) formation of aminoacyl-adenylate by reaction between amino acid and ATP; and (ii) transfer of amino acid residue from aminoacyl-adenylate to 2' or 3' hydroxyl group of 3' terminal adenosine of tRNA. Both of the steps were shown to be carried out abiotically or catalysed by ribozymes that have been extensively developed from the pools of random RNA sequences by directed evolution and presumably predated the emergence of aminoacyl-tRNA synthetases (Wong 2024). Aminoacyl-adenylates can be produced by trimetaphosphate-

mediated condensation of adenosine and neutral amino acids (Val, Leu, Phe, Tyr, Trp) or the reaction of AMP with N-carboxyanhydrides of amino acids (Biron *et al.* 2005; Ying *et al.* 2018). The resulting aminoacyl-adenylates could be later employed as substrates for tRNA activation by self- or trans-aminoacylation ribozymes that do not have to achieve high chemical complexity for sufficient catalysis (Illangasekare and Yarus 1999; Lee and Suga 2001; Murakami *et al.* 2003; Turk *et al.* 2010). The shortest known ribozyme with trans-aminoacylation activity, 5'-GUGGC-3', consists solely of 5 nucleotides and can catalyse reaction between phenylalanyl-adenylate (Phe-AMP) and 5'-GCCU-3' (Turk *et al.* 2010). The 5'-GUGGC-3' ribozyme selectively transfers Phe from Phe-AMP to 2'-hydroxyl group of 3'-terminal uridine leading to formation of GCCU-Phe as well as GCCU-Phe_2 and GCCU-Phe_3 products.

Alternatively, entire aminoacylation could be carried out abiotically by intra- or intermolecular transfer of amino acid residue from 5'-terminal phosphate to the hydroxyl group of 3' terminal nucleotide of ancestral tRNAs (Tamura and Schimmel 2004; Wu *et al.* 2021; Roberts *et al.* 2022; Su *et al.* 2023) (Fig. 1.8). Amino acids can be initially coupled to 5'-terminal phosphate via their amino group in the form of phosphoramidates that can be abiotically produced from imidazole-activated ribonucleotides and free amino acids (Weber and Orgel 1978; Roberts *et al.* 2022). Alternatively, amino acids can be attached to 5'-terminal phosphate via their carboxylic group in the form of aminoacyl phosphate mixed anhydrides that can be formed from free ribonucleotides and COS/urea-activated amino acid derivatives, N-carboxyanhydrides (Leman *et al.* 2006; Wu *et al.* 2021). In both cases, the amino acid transfer from 5'-terminal phosphate to the hydroxyl group of 3' terminal nucleotide has been demonstrated to be more efficient for L-amino acids than D-amino acids, indicating another plausible mechanism for the origin of homochirality (Tamura and Schimmel 2004; Wu *et al.* 2021). The resulting aminoacyl-RNAs can spontaneously react with aminoacyl phosphate-bearing oligonucleotides in a template-directed manner leading to formation of peptidyl-RNAs (Tamura and Schimmel 2003).

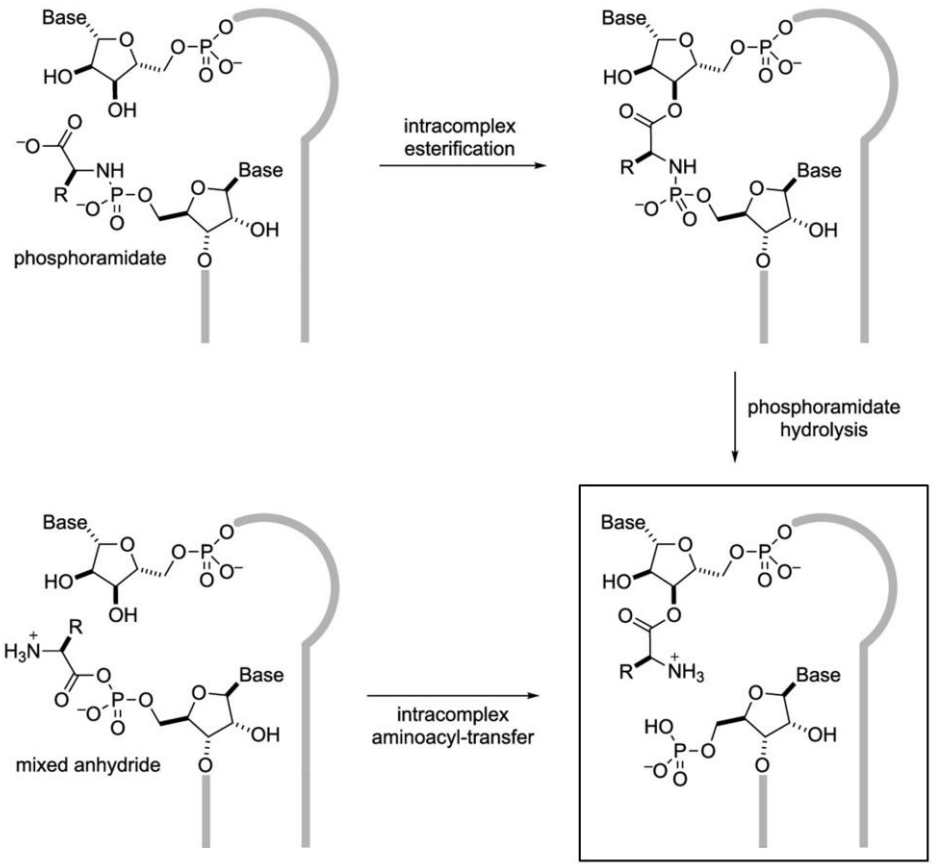


Fig. 1.8. A possible mechanism of abiotic tRNA aminoacylation through the transfer of amino acid residue from 5'-terminal phosphate to the hydroxyl group of 3'-terminal nucleotide. Adopted and modified from Su *et al.* (2023).

A rather interesting mechanism of ancestral tRNA activation and subsequent peptide synthesis has been recently proposed to occur via nucleobase modifications instead of aminoacylation. tRNAs contain a wide variety of chemically modified nucleobases that were proposed to be "molecular vestiges" of diverse abiotic reactions between nucleobases and other prebiotically available molecules, including amino acids (Schneider *et al.* 2018). In particular, N6-substituted adenosines bearing simple amino acids such as Gly (g^6A) and Thr (t^6A), are commonly found in tRNAs across all three kingdoms of life and were shown to be abiotically produced from the mixture of adenosine, methyl isocyanate and corresponding amino acids (Schneider *et al.* 2018). In tRNAs, g^6A and t^6A reside next to the anticodon loop at position 37 and in close proximity to amino-bearing non-canonical nucleoside, 5-methylaminomethyl uridine (mnm^5U), at position 34 (Müller *et al.* 2022) (Fig. 1.9). Müller *et al.* (2022) showed that wet-dry cycling of two short complementary RNA strands bearing 5'-terminal g^6A (donor strand) and 3'-terminal mnm^5U (acceptor strand) leads to the efficient Gly transfer from g^6A to the amino group of mnm^5U and subsequent formation of glycylglycine in the following reaction cycle. The same reaction was successfully performed for a panel of N6-substituted

adenosines containing other amino acids (Ala, Val, Pro, Ser, Thr, Asp, Phe, Met) as well as pre-formed di- and tripeptides, indicating the plausibility of peptide bond formation between two non-canonical nucleosides. Abiotic nucleobase modification with amino acids could represent the earliest mechanism of RNA-mediated peptide synthesis prior to the advent of aminoacylation, which persisted till the modern days.

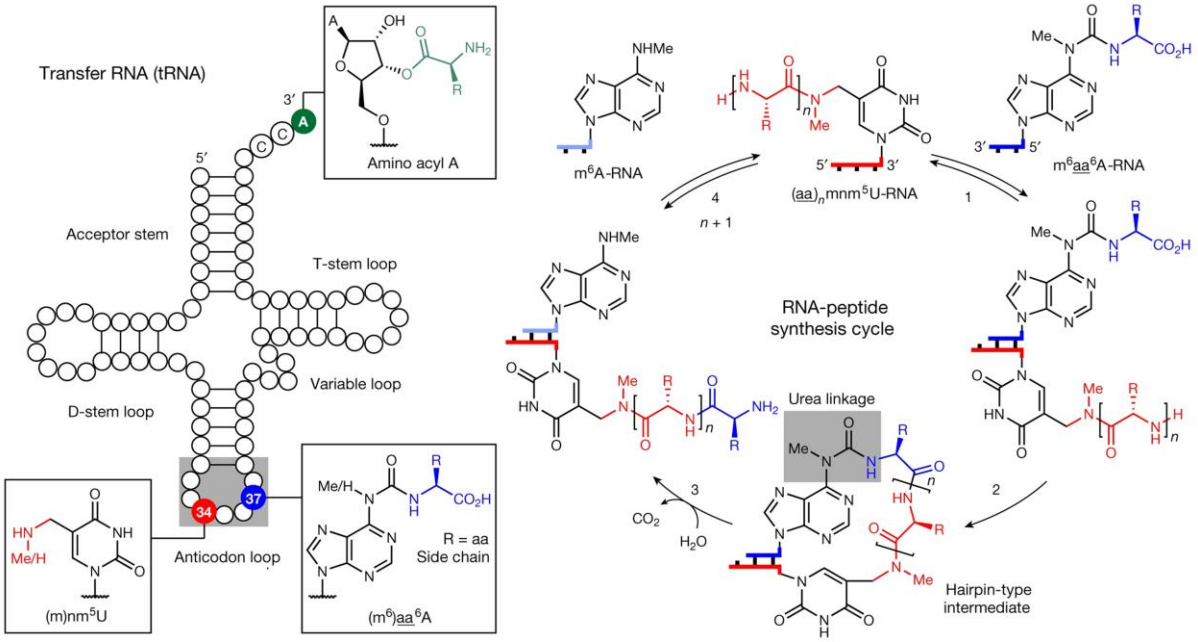


Fig. 1.9. A possible mechanism for peptide bond synthesis mediated through nucleobase modifications commonly found in tRNAs. Adopted from Müller *et al.* (2022).

Despite the seemingly common evolutionary origin of the ancestral PTC and proto-tRNAs, the PTC later evolved into a separate structural and catalytic scaffold enabling the peptidyl transfer between two spatially approximated tRNA substrates. Historically, the earliest crystallographic studies inferred that PTC acts as an acid-base catalyst, providing the invariant catalytic adenosine residue that abstracts a proton from the amino group of aminoacyl-tRNA for its subsequent nucleophilic attack on peptidyl-tRNA (Nissen *et al.* 2000). However, numerous biochemical studies indicate that the PTC acts as an entropic catalyst that accommodates and orients two substrates (aminoacyl-tRNA and peptidyl-tRNA) in the position that facilitates nucleophilic attack and formation of the transient state (Sievers *et al.* 2004; Bieling *et al.* 2006). The model uncatalysed reaction between free amino acids and aminoacyl adenosine esters is possible but occurs at a very slow rate ($<10^{-6}$ s^{-1}); the substrate positioning provided by the PTC accelerates the reaction up to 10^7-fold (Sievers *et al.* 2004; Polacek *et al.* 2005). The entropic catalysis requires the well-fixed three-dimensional structure of the PTC, which could be achieved by the accretion of rRNA and ribosomal proteins around the ancestral PTC core leading to formation of the large ribosomal subunit (Hsiao *et al.* 2009; Petrov *et al.*

2015). In contrast to rRNA, ribosomal proteins are not involved in catalysis and are less conserved across three domains of life and thus most likely appeared later in the evolutionary history of the ribosome (Cech 2000; Lecompte *et al.* 2002; Ban *et al.* 2014). In the absence of ribosomal proteins, the ancestral PTC could still have catalysed non-specific and non-coded condensation of prebiotic amino acids leading to the emergence of short peptides that predated ribosomal proteins (Davidovich *et al.* 2009; Hsiao *et al.* 2009; Petrov *et al.* 2015; Lupas and Alva 2017). Short peptides could have bound back onto the ancestral PTC and provided a structural framework for the subsequent accretion of rRNA as well as other evolutionary advantages necessary for the early ribosome evolution, which will be addressed in this dissertation.

1.5. Evolution of the genetic code

The further evolution of the ribosome and the other components of translation machinery ultimately led to the emergence of genetic coding. The genetic code is the set of rules universally employed by all living organisms to translate the nucleotide sequence of mRNAs to the amino acid sequence of synthesised proteins. The structure of the genetic code is non-random, and each of 20 canonical amino acids used for protein synthesis is assigned to one or several specific sequences of three nucleotides (codons) in mRNAs. Several evolutionary theories have been proposed in order to explain the origin and structure of the genetic code and will be discussed below.

1.5.1. The frozen accident theory of the genetic code evolution

The frozen accident theory, proposed by Crick (1968), suggests that the genetic code was fully established by the time of the LUCA, and any change to the codon assignment would be deleterious to protein functions and selected against over the course of evolution. The frozen accident theory explains the universality of the genetic code across all living organisms with a few exceptions observed for the codon assignment in non-plant mitochondria and nuclei of certain organisms, including the stop codon reassignments to incorporate 21. (selenocysteine) and 22. (pyrrolysine) proteinogenic amino acids (Jukes and Osawa 1990; Ambrogelly *et al.* 2007). The codon reassignments leading to the deviations from the standard genetic code were suggested to have occurred after divergence of life from the LUCA via several mechanisms. Codon capture hypothesis assumes that certain codons could have temporarily disappeared from genomes of ancestral organisms during GC content optimization (Osawa and Jukes 1989). Some of these codons could have later reappeared and been assigned to different tRNAs (Osawa and Jukes 1989). According to the ambiguous intermediate hypothesis, tRNA

mutations could have led to the transient assignment of a codon to two amino acids by wild-type and mutant tRNAs with the subsequent evolutionary fixation of mutant tRNA (Schultz and Yarus 1994; Knight *et al.* 2001). Lastly, the codon reassignment could have also occurred via the loss and gain of tRNA-encoding genes through deletion and duplication events over the course of evolution, in particular, during genome size reduction of mitochondrial genomes (Andersson and Kurland 1991; Sengupta *et al.* 2007).

1.5.2. The adaptation theory of the genetic code evolution

According to the adaptation theory, proposed by Woese (1965) and Epstein (1966), the structure of the genetic code was shaped in a way that allows to minimise the effect of point mutations and translation errors on the structure and function of synthesised proteins. In fact, many amino acids are assigned to synonymous codons that differ only in the third nucleotide, thus preventing the changes in amino acid sequences caused by single point mutations in mRNA (Epstein 1966). The high degeneracy of the third nucleotide in synonymous codons is compensated by wobble-pairing with the first nucleotide of anticodons in tRNAs (Crick 1966). Synonymous codons encoding the same amino acid are not used with the same frequency by living organisms, and several lines of evidence inferred the higher usage frequency for the codons that minimise the impact of errors from mutation or mistranslation (Archetti 2004), albeit this observation is not consistent across all organisms (Zhu *et al.* 2003; Marquez *et al.* 2005). Related codons that differ only in one nucleotide in any position often encode for amino acids with similar physico-chemical properties as evidenced by the examples of negatively charged Asp (GAU/C) and Glu (GAA/G); positively charged Lys (AAA/G) and Arg (AGA/G); aromatic Phe (UUU/C) and Tyr (UAU/C); and so on (Woese 1965; Saier 2019). The genetic code is less robust towards two nucleotide substitutions, however, all codons with uridine in the second position are strictly assigned to hydrophobic amino acids (Val, Leu, Ile, Phe, Ile, Met) (Koonin and Novozhilov 2009; Saier 2019). The adaptive theory is supported by several statistical analyses elucidating that the probability of finding an alternative to the genetic code with a different amino acid-codon distribution but the same robustness towards point mutations and translation errors is less than 10^{-8} (Gilis *et al.* 2001; Goodarzi *et al.* 2004).

1.5.3. The stereochemical theory of the genetic code evolution

The stereochemical theory, proposed by Woese (1967), implies that the specific assignment of amino acids to codons was dictated by physico-chemical affinity of amino acids to the respective codons. So far, no direct experimental evidence has been provided to observe preferential binding of single amino acids to trinucleotide RNA sequences. However, indirect support towards the stereochemical theory comes from the analyses of amino acid-binding sites

of riboswitches and RNA aptamers evolved to bind amino acids *in vitro* (Yarus *et al.* 2005; Yarus *et al.* 2009). Six (Arg, His, Ile, Phe, Tyr, Trp) proteinogenic amino acids were shown to have binding sites that are statistically enriched in the cognate triplets (Yarus *et al.* 2009). Yarus *et al.* (2005) speculated that those triplets could have been later acquired by translation machinery to become the respective codons and anticodons during transition from the RNA-world to modern DNA/RNA/protein world. Another support comes from the more recent analysis of the spatial adjacency of amino acids and nucleotides in the structures of bacterial and archaeal ribosomes (Johnson and Wang 2010). Johnson and Wang (2010) elucidated that 8 amino acids (Ile, Leu, The, Glu, Gly, Arg, Lys, Tyr) are found in close proximity (within 5 Å) to their respective codons, and 10 amino acids (Phe, Trp, Asp, Ile, Gln, Tyr, Leu, His, Arg, Met) are found near their respective anticodons with a statistically high frequency in ribosome structures. These observations indicate that the genetic code could have originated very early in the evolutionary history of life, presumably in the "RNA world" stage, and still remains imprints of ancestral RNA-amino acid interactions. However, many simpler proteinogenic amino acids lack side chains that are sufficiently complex for interactions with RNA, implying that the physico-chemical affinity of amino acids to RNA may have played a role at the later stages of genetic code evolution (Johnson and Wang 2010).

1.5.4. The co-evolution theory of the genetic code evolution

According to the coevolution theory, proposed by Wong (1975), the early evolution of the genetic code was limited by the prebiotic availability of amino acids. The coevolution theory implies that the genetic code evolved from a smaller set of abiotically synthesizable amino acids that gradually expanded to include more complex amino acids after establishment of their corresponding biosynthetic pathways. The strongest support towards the stepwise addition of amino acids to the genetic code comes from the pre-translational modifications of aminoacyl-tRNAs that convert the amino acid moiety into another amino acid. The most known example is selenocysteine (Sec), which does not have its cognate aminoacyl-tRNA synthetase. Instead, seryl-tRNA synthetase catalyses formation of seryl-tRNA^{Sec} that is converted into selenocysteinyl-tRNA^{Sec} by selenocysteine synthase (in bacteria) or O-phosphoseryl-tRNA^{Sec} kinase and O-phosphoseryl-tRNA^{Sec}:Sec synthase (in archaea and eukaryotes) (Turanov *et al.* 2011; Meng *et al.* 2022). Many bacteria and archaea do not have aminoacyl-tRNA synthetases for Asn, Gln, and Cys, which instead are introduced to protein synthesis indirectly by enzymatic modifications of aspartyl-tRNA^{Asn}, glutamyl-tRNA^{Gln}, and phosphoseryl-tRNA^{Cys}, respectively (Sheppard *et al.* 2008; Meng *et al.* 2022). The derivation of both asparagine synthetase and asparaginyl-tRNA synthetase from aspartyl-tRNA synthetase as well as

glutaminyl-tRNA synthetase from glutamyl-tRNA synthetase suggest that pre-translational modifications of aminoacyl-tRNAs preceded the direct incorporation of Asn and Gln into proteins (Francklyn 2003; O'Donoghue *et al.* 2011). The cognate codons for Asn (AAU/C) and Gln (CAA/G) are related to the cognate codons for their biosynthetic precursors, Asp (GAU/C) and Glu (GAA/G), respectively, differing only in the purine nucleotide at the first position, which hints at plausible codon reassignments from Asp/Glu to Asn/Gln in the early history of the genetic code (Wong 1975). In fact, many other proteinogenic amino acids (Leu, Ile, Pro, Thr, Lys, Arg, His, Met, Cys, Trp, Tyr) share related codons, i.e. the codons that differ by only one nucleotide, with their respective amino acid precursors, providing the general link between the structure of the genetic code and amino acid biosynthesis in modern living organisms (Wong 1975).

Another evidence to the gradual genetic code expansion comes from ancestral sequence reconstructions of aminoacyl-tRNA synthetases that are considered one of the most ancient proteins predating the emergence of the LUCA (Ribas de Pouplana 2020). Aminoacyl-tRNA synthetases divided into two structurally and functionally different classes are generally believed to have descended from two common ancestors that were encoded by the opposite strands of the same ancestral gene (Rodin and Ohno 1995; Martinez-Rodrigez *et al.* 2015). The common evolutionary ancestors of aminoacyl-tRNA synthetases presumably operated with a smaller amount of amino acids that later expanded to the current amino acid set by gene duplication and mutation events (Rodin and Ohno 1995; Martinez-Rodrigez *et al.* 2015). Isoleucyl- and valyl-tRNA synthetases share a high level of structural and sequence similarity, and their pre-LUCA paralog ancestor, inferred by Fournier *et al.* (2011), is enriched in both Val and Ile residues that are highly preserved in contemporary isoleucyl- and valyl-tRNA synthetases, indicating the simultaneous presence of Val and Ile in the ancestral genetic code. However, the similar ancestral sequence reconstruction performed later by Fournier and Alm (2015) for tyrosyl- and tryptophanyl-tRNA synthetases revealed that their pre-LUCA paralog ancestor was completely devoid of Trp residues, indicating the later recruitment of Trp to the genetic code. The later inclusion of Trp (and other aromatic amino acids) is also supported by the sequence conservation analyses of the set of LSU and SSU ribosomal proteins that are universally present across three domains of life. Two independent analyses inferred the universally conserved sites (dating back to the LUCA) of those ribosomal proteins that were depleted in Glu, Lys, Phe, Tyr, Trp, Cys and enriched in Gly and Asn at the time of the LUCA, providing hints at the plausible chronology of amino acid incorporation into the genetic code (Brooks and Fresco 2002; Fournier and Gogarten 2007).

The temporality of amino acid recruitment into the genetic code has been elucidated by several researchers using different metrics. Wong (2005) inferred three general stages of amino acid recruitment to the genetic code based on the experimental reports on their abiotic formation and stability (Fig. 1.10). The comparison between two lines of experimental evidence implying either terrestrial or extraterrestrial origin of amino acids leads to the consensus set of 10 proteinogenic amino acids (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu) that were presumably available in the prebiotic environment and constituted the early version of the genetic code (stage 1) (Wong 2005; McDonald and Storrie-Lombardi 2010; Longo and Blaber 2012; Kirschning 2022). The later addition of the other 10 proteinogenic amino acids (Asn, Gln, Lys, Arg, His, Phe, Tyr, Trp, Cys, Met) in stage 2 is supported by their absence/scarcity in abiotic synthesis simulations and/or their instability under prebiotic conditions. Despite the chemical and structural similarity to Asp and Glu, the prebiotic abundance of Asn and Gln is unlikely due to their thermal instability as well as spontaneous deamidation into respective Asp and Glu via intramolecular cyclization into succinimide (Asn) or glutarimide (Gln) that occurs both *in vitro* and *in vivo* (Geiger and Clarke 1987; Robinson 2002). Positively charged amino acids (Lys, Arg, His) have not been produced from atmospheric gas mixtures or identified (except Lys) in carbonaceous chondrites (McDonald and Storrie-Lombardi 2010; Longo and Blaber 2012; Cobb and Pudritz 2014; Kirschning 2022). Instead, shorter non-proteinogenic homologs of Lys, namely 2,3-diaminopropionic (Dap), 2,4-diaminobutyric acid (Dab), and ornithine (Orn), have been found in substantial amounts in both terrestrial and extraterrestrial sources (Ring *et al.* 1972; Wolman *et al.* 1972; Meierhenrich *et al.* 2004). Given the central role of His in enzyme catalysis, several attempts have been made to abiotically produce imidazole and its derivatives from more convoluted precursors (Shen *et al.* 1990; Vázquez-Salazar *et al.* 2017; Ribeiro *et al.* 2020). His was shown to be produced in four consecutive reactions from highly concentrated solutions of erythrose and formamidine, however, the prebiotic relevance of the proposed synthetic route remains unknown (Shen *et al.* 1990; Vázquez-Salazar *et al.* 2017). Aromatic (Phe, Tyr) amino acids were not synthesised from atmospheric gas mixtures but have been consistently identified along with alkylbenzenes and polycyclic aromatic hydrocarbons in carbonaceous chondrites and the Ryugu asteroid (Burton *et al.* 2012; Glavin *et al.* 2018; Potiszil *et al.* 2023; Zeichner *et al.* 2023). The more convoluted synthetic route to Phe and Tyr was proposed to occur via H_2S-mediated derivatization of phenylacetylene, which can be formed in small quantities by pyrolysis of simple hydrocarbons (Friedmann and Miller 1969). Inclusion of H_2S into reduced gas mixtures during electric discharge experiments also leads to formation of Met, which is

otherwise absent in carbonaceous chondrites and asteroids despite the high abundance of sulphur oxides and various alkylsulphonic acids in carbonaceous chondrites (Parker *et al.* 2011; Naraoka *et al.* 2023). However, both aromatic (Phe, Tyr, Trp) and sulphur-containing (Cys, Met) amino acids rapidly decompose under UV irradiation, which would exclude their prebiotic accumulation despite the plausibility of their either extraterrestrial or terrestrial origin (Wong 2005). The recent quantum-chemical calculations of the chemical reactivity of amino acids imply that their later recruitment into the genetic code could have occurred in response to the oxidative damage of early polypeptides caused by the introduction of molecular oxygen in the biosphere (Granold *et al.* 2018). Lastly, the non-universal distribution of selenocysteine (21.) and pyrrolysine (22.) across three domains of life implies their latest inclusion into the genetic code after the divergence of life from the LUCA (stage 3) (Wong 2005).

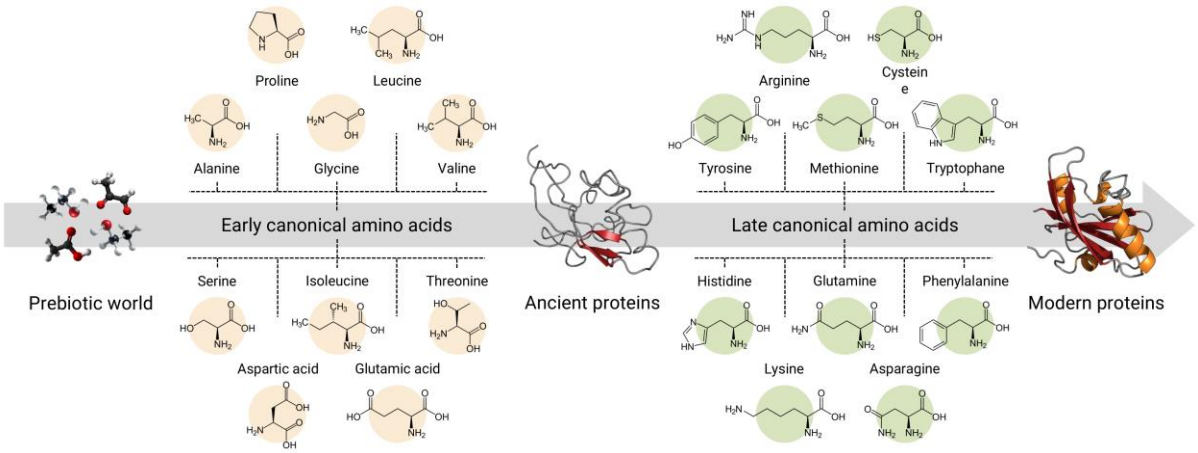


Fig. 1.10. Co-evolution theory of the genetic code (Wong 2005).

Two meta-analyses have been implemented to elucidate the exact order of amino acid incorporation into the genetic code. Higgs and Pudritz (2009) inferred the order of the amino acid recruitment based on the relative concentrations of amino acids that were cumulatively reported in various carbonaceous chondrites analyses, electric discharge experiments, interstellar ice irradiation and hydrothermal synthesis simulations (Fig. 1.12). The inferred order was shown to strongly correlate with the thermodynamic calculations on amino acid formation from CO_2, NH_4, and H_2 in surface seawater conditions (18 °C, 1 atmosphere), implying the chemical complexity of amino acids as one of the critical factors of their recruitment into the genetic code (Amend and Shock 1998). In fact, there is a general negative correlation between the chemical complexity of amino acids and the number of their respective codons in the genetic code with Trp, the most complex canonical amino acid, having only one codon (Koonin and Novozhilov 2009). The proposed order was shown to be in good agreement with the metabolic costs of amino acid biosyntheses in *Escherichia coli* expressed in ATP

equivalents (Akashi and Gojobori 2002; Higgs and Pudritz 2009). The amino acid chronology proposed by Higgs and Pudritz (2009) strongly correlates with the meta-ranking of amino acids incorporation, inferred by Trifonov (2008), based predominantly on different theories of the genetic code evolution (Fig. 1.11). The ancestral version of the genetic code presumably consisted of a smaller number of codons that could have also been smaller in size. The minimal codon size of three nucleotides (4^3=64) is required for the unambiguous coding of 20 (or 21) canonical amino acids, however, a smaller number of amino acids can be assigned to single nucleotides (4^1=4), dinucleotides (4^2=16) or a lesser number of related trinucleotides. Thus, numerous mono-, di- and trinucleotide models of the genetic code expansion have been proposed implying various sets of the cognate amino acids to be the first components of the genetic code. Alternatively, the plausible chronology of amino acid incorporation was also inferred from compositional, structural and phylogenetic analyses of tRNAs and aminoacyl-tRNA synthetases. Trifonov (2008) elucidated the consensus temporal order of amino acid inclusion into the genetic code based on 101 theories considering the genetic code expansion from smaller sets of codons, evolution of translation machinery components, abiotic formation of amino acids and their physico-chemical properties. Both meta-rankings are in good agreement with the classification of amino acids into early and late proposed by Wong (2005).

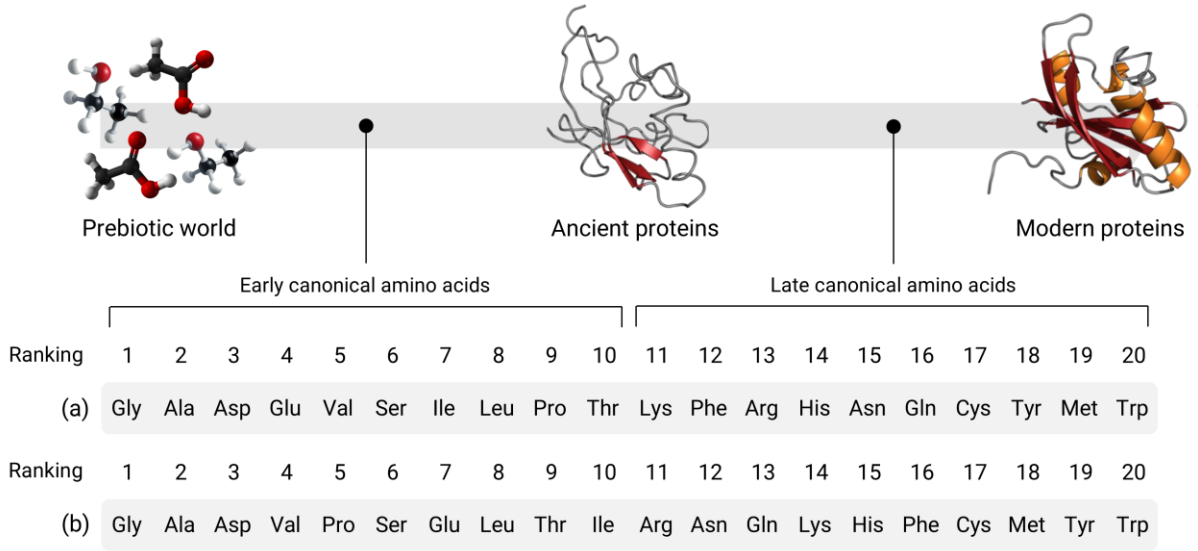


Fig. 1.11. The temporality of inclusion of canonical amino acids in the genetic code deduced from (a) their prebiotic availability; and (b) the consensus among various theories of the genetic code evolution (Trifonov 2008; Higgs and Pudritz 2009).

All the aforementioned theories are not mutually exclusive and can be viewed as complementary since they capture different aspects of the genetic code evolution (Yarus *et al.* 2005; Koonin and Novozhilov 2009). However, none of the evolutionary theories addresses the lack of evolutionary continuity between the amino acid repertoire presumably available on

prebiotic Earth and the amino acid composition of the genetic code. Both terrestrial and extraterrestrial scenarios of abiotic amino acid synthesis indicate the high prebiotic abundance of diverse non-canonical amino acids that were not recruited to the genetic code over prebiotically scarce proteinogenic amino acids. The exclusion of prebiotically available non-canonical amino acids from the genetic code will be addressed in this dissertation.

2. Aims of the thesis

The main objective of this dissertation was to elucidate the selection factors that led to (i) the exclusion of prebiotically available non-proteinogenic alternatives of early proteinogenic amino acids from the genetic code and (ii) the inclusion of late proteinogenic amino acids into the genetic code. Numerous lines of evidence support the plausibility of the non-specific formation of random short peptide sequences mediated first abiotically by diverse physicochemical mechanisms and subsequently by ancestral translation machinery after the emergence of the proto-ribosome. These random peptide sequences were most likely enriched in numerous non-polar, positively and negatively charged non-canonical amino acids, which, however, were not included in the genetic code despite their high prebiotic availability. Instead, many origin-of-life researchers agree that the early protein alphabet consisted of a smaller amino acid set that included aliphatic non-polar and negatively charged proteinogenic amino acids. The early amino acid alphabet later expanded to include positively charged and aromatic proteinogenic amino acids. This work aimed to experimentally elucidate the structural, biophysical, and functional constraints that led to the current amino acid composition of the standard genetic code.

Specifically, the following individual aims were addressed in this dissertation:

Aim I. Exploration of biophysical and structural properties of early, alternative early, and modern amino acid alphabets.

Aim II. Exploration of biophysical and structural properties of differently charged amino acids alphabets.

Aim III. Characterization of RNA-binding propensity of ancestral ribosomal peptides enriched in late proteinogenic amino acids.

3. Materials and methods

This section provides a detailed description of the experimental protocols established and implemented by the author of the dissertation for the two published research articles and one submitted manuscript. The author performed all the experiments listed below except for the chemical synthesis and quality control of combinatorial peptide libraries (sections 3.1.2 and 3.2.2) and ancestral ribosomal peptides (section 3.3.2). If not stated otherwise, all chemicals were purchased from Sigma Aldrich. All buffers were prepared and filtered prior to analyses using 0.22 μm MF-Millipore membrane filters (Merck).

3.1. Biophysical and structural analysis of early, alternative early, and late combinatorial peptide libraries

3.1.1. Design of combinatorial peptide libraries

For this study, we designed six 25-mer combinatorial peptide libraries composed of amino acid subsets that represent different or alternative stages of the amino acid alphabet evolution. The amino acid subsets were the following:

1. 10E: early canonical amino subset (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu).
2. 10U: alternative early amino acid subset, in which branched Val, Leu, Ile are replaced by unbranched 2-aminobutyric acid (Aba), norvaline (Nva), and norleucine (Nle).
3. 11R: early canonical amino acid subset supplemented with positively charged Arg.
4. 11D: early canonical amino acid subset supplemented with positively charged Dab.
5. 11Y: early canonical amino acid subset supplemented with Tyr.
6. 19F: full canonical amino acid set except Cys.

25-mer combinatorial peptide libraries were designed to include equimolar amounts of amino acids at every position.

3.1.2. Chemical synthesis and quality control of combinatorial peptide libraries

Peptide libraries were commercially synthesised by Spyder Institute Praha (Czech Republic) using isokinetic mixtures of Fmoc-protected amino acids and standard Fmoc-based solid-phase methodology. After synthesis, peptide libraries were lyophilized three times with 1 mM HCl to establish chloride as a counter ion. The quality of synthesised peptide libraries was confirmed using high-performance liquid chromatography (HPLC) amino acid analysis to assess the amino acid composition and matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry analysis to determine the molecular weight distribution. The HPLC amino acid analysis was performed by Dr. Radko Soucek (Drug Discovery group, IOCB, Czech Republic) on Agilent 1260 HPLC (Agilent Technologies)

equipped with an ultraviolet-visible (UV-Vis) detector using automated ninhydrin or o-phthalaldehyde (OPA) derivatization. Before the amino acid analysis, the library samples were hydrolysed in 6 M HCl at 110 °C for 20 h, and the hydrolysate was subsequently evaporated and reconstituted with 0.1 M HCl. The MALDI-TOF analysis was performed by Dr. Katerina Novakova (Josef Cvačka Group, IOCB, Czech Republic) on UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics), according to the standard procedure. Before the MALDI-TOF analysis, 10 mg/ml stock solutions of peptide libraries were prepared by dissolving them in 50 % (v/v) acetonitrile supplemented with 0.1 % (v/v) acetic acid. 19F peptide library was dissolved in 50 % (v/v) acetonitrile. The samples were prepared from stock solutions using 2,5-dihydroxybenzoic acid (DHB) as a matrix, according to the standard protocol.

3.1.3. Solubility analysis of combinatorial peptide libraries

Solubility and aggregation analyses were carried out in a series of universal Britton-Robinson (ABP) buffers that were prepared from a stock solution containing 20 mM acetic acid, 20 mM phosphoric acid, and 20 mM boric acid. The pH was adjusted with 5 M NaOH to 3.0, 5.0, 7.4, 9.0, or 11.0. The ionic strength was adjusted with NaCl to the conductivity that corresponds to the conductivity of 50 mM (low ionic strength) or 500 mM (high ionic strength) NaCl (Tab. 3.1). The conductivities of ABP buffers and standard NaCl solutions were measured on NGC Quest Plus System (Bio-Rad).

Tab. 3.1. The NaCl concentrations in 20 mM ABP buffers

Ionic strength	Low (eq. 50 mM NaCl)					High (eq. 500 mM NaCl)				
pH	3.0	5.0	7.4	9.0	11.0	3.0	5.0	7.4	9.0	11.0
c (NaCl), mM	38	30.6	15.6	8.8	0	488	480.6	465.6	458.8	450

Before analysis, 5 mg/ml stock suspensions of peptide libraries were prepared by resuspending them in autoclaved MilliQ water. Stock suspensions were diluted into the corresponding ABP buffer to the final concentration of 0.5 mg/ml and incubated for 30 min at room temperature with mild shaking (500 rpm). After incubation, the insoluble fraction was removed by centrifugation at maximum speed (21,300 *g*) for 30 min at 4 °C, and the supernatant (cleared sample) was subsequently used for solubility and aggregation analyses.

The relative solubility of all peptide libraries (except 19F) was estimated by UV-Vis absorption spectroscopy. UV-Vis absorption spectroscopy analysis was carried out by measuring the absorbance at 215 nm in the cleared sample on Nanodrop 2000 (Thermo Fisher Scientific) using 2 µl of the cleared sample. The standard curves were plotted from 2-fold serial

dilutions (0.5, 0.25, 0.125, 0.063, and 0.031 mg/ml) of 0.5 mg/ml peptide library solutions in 20 mM ABP buffer (pH 11) at low ionic strength. The measurements were performed in technical sextuplicates, and the overall analysis was performed three times.

The relative solubility of 19F peptide library was estimated using Quantitative Fluorometric Peptide Assay kit (Thermo Fisher Scientific). 10 µl of the cleared sample was mixed with 70 µl of Fluorometric Peptide Assay Buffer and 20 µl of Fluorometric Peptide Assay Reagent in a 96-well black plate (Thermo Fisher Scientific). The reaction mixture was incubated for 30 min at room temperature in the dark, and the fluorescence intensity was measured on Spark multimode plate reader (Tecan), according to the following setup: excitation at 390 nm, emission at 475 nm. The standard curve was plotted from 2-fold serial dilutions (0.5, 0.25, 0.125, 0.063, and 0.031 mg/ml) of 0.5 mg/ml 19F peptide library solution in dimethyl sulfoxide (DMSO). The measurements were performed in technical quadruplicates, and the overall analysis was conducted three times.

3.1.4. Aggregation propensity analysis of combinatorial peptide libraries

Aggregation propensity analysis was carried out using analytical size-exclusion chromatography. The 100 µl aliquot of cleared sample was loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva) pre-equilibrated with two bed volumes of the corresponding ABP buffer. The isocratic elution of peptides was run with one bed volume of 10 mM ABP buffer (pH 7.4) at 0.5 ml/min flow rate at room temperature, and the progress of elution was monitored by absorbance of peptide bonds at 215 nm. The relative amount of soluble aggregates was calculated by dividing the area under the aggregate (~8-ml elution volume) peak by the total area under all the peaks from 5- to 20-ml elution volumes using ChromLab Software v6.1 (Bio-Rad). Size-exclusion chromatography runs were performed once.

3.1.5. Secondary structure analysis of combinatorial peptide libraries

Secondary structure analysis was performed by far-UV CD spectroscopy. The far-UV CD spectra were collected on Chirascan-plus spectrophotometer (Applied Photophysics) in the 190–260 nm spectral range using a 1 mm path-length quartz cell. The CD measurements were carried out for 160 µl aliquots of cleared samples at 0.2 mg/ml nominal concentration, according to the following setup: 1 nm step, 1 nm/s scanning speed, room temperature. The CD signal was obtained as ellipticity in units of millidegrees (mdeg). The resulting spectra were averaged from two scans, buffer-spectrum subtracted, and baseline-corrected. The overall analysis was performed twice, and the ratio of ellipticities at 222 and 200 nm was averaged and used to determine the secondary structure content in peptide libraries.

The effect of pH on secondary structure was determined by diluting 5 mg/ml suspensions of peptide libraries into 10 mM ABP buffers at pH 3.0, 5.0, 7.4, 9.0, and 11.0 to the final concentration of 0.2 mg/ml. The effect of 2,2,2-trifluoroethanol on the secondary structure was determined by diluting 5 mg/ml stock suspensions of peptide libraries into a series of 10 mM ABP buffers (pH 7.4) supplemented with 0–90 % (v/v) 2,2,2-trifluoroethanol (TFE) to the final concentration of 0.2 mg/ml. The peptide samples were incubated for 30 min at room temperature with gentle shaking (500 rpm) and then cleared by centrifugation at maximum speed (21,300 *g*) for 15 min at 4 °C.

3.2. Biophysical and structural analysis of differently charged combinatorial peptide libraries

3.2.1. Design of combinatorial peptide libraries

For this study, we designed eight 25-mer combinatorial peptide libraries that were composed of amino acid subsets with various combinations of positively and negatively charged amino acids. The amino acid subsets were the following:

1. 8E: the subset of early uncharged canonical amino acids (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr).
2. 8E+Aad,Api: the 8E subset supplemented with negatively charged non-canonical amino acids (2-aminoadipic acid (Aad), 2-aminopimelic acid (Api));
3. 8E+Lys,Arg: the 8E subset supplemented with positively charged canonical amino acids (Lys, Arg);
4. 8E+Dap,Dab: the 8E subset supplemented with positively charged non-canonical amino acids (Dap, Dab);
5. 8E+Lys,Arg,Asp,Glu: the 8E subset supplemented with positively charged canonical amino acids (Lys, Arg) and negatively charged canonical amino acids (Asp, Glu).
6. 8E+Lys,Arg,Aad,Api: the 8E subset supplemented with positively charged canonical amino acids (Lys, Arg) and negatively charged non-canonical amino acids (Aad, Api).
7. 8E+Dap,Dab,Asp,Glu: the 8E subset supplemented with positively charged non-canonical amino acids (Dap, Dab) and negatively charged canonical amino acids (Asp, Glu).
8. 8E+Dap,Dab,Aad,Api: the 8E subset supplemented with positively charged non-canonical amino acids (Dap, Dab) and negatively charged non-canonical amino acids (Aad, Api).

The combinatorial peptide libraries were designed to include equimolar amounts of amino acids at every position.

3.2.2. Chemical synthesis and quality control of combinatorial peptide libraries

The synthesis and quality control were performed, according to the methodology described in 3.1.2. The experimental protocol for HPLC amino acid analysis was modified by Dr. Radko Soucek for the additional identification of non-canonical amino acids (Dap, Dab, Aad, Api).

3.2.3. Solubility analysis of combinatorial peptide libraries

The relative solubility of all peptide libraries (except 8E) was analysed by UV-Vis spectroscopy, according to the methodology described in 3.1.3., except the analysis of the peptide libraries was carried out twice. The relative solubility of 8E peptide library was determined using fluorescamine assay (Udenfriend *et al.* 1972). In brief, 10 µl of the cleared sample was mixed with 80 µl of the corresponding ABP buffer and 30 µl of 3 mg/ml fluorescamine solution in DMSO in a 96-well black plate (Thermo Fisher Scientific). The reaction mixture was kept for 30 min at room temperature in the dark, and the fluorescence intensity was recorded on Spark multimode plate reader (Tecan), according to the following setup: excitation at 365 nm, emission at 470 nm. The standard curve was plotted from 2-fold serial dilutions (0.5, 0.25, 0.125, 0.063, and 0.031 mg/ml) of 0.5 mg/ml 8E peptide library solution in DMSO. The measurements were performed in technical quadruplicates, and the overall analysis was conducted two times.

3.2.4. Aggregation propensity analysis of combinatorial peptide libraries

The aggregation propensity analysis was conducted, according to the methodology described in 3.1.4.

3.2.5. Secondary structure analysis of combinatorial peptide libraries

The secondary structure analysis was performed, according to the methodology described in 3.1.5.

3.2.6. Isolation of monomer, aggregate, and total fractions from 10E peptide library

Before analysis, 5 mg/ml stock suspension of 10E peptide library was diluted with 10 mM ABP buffer (pH 7.4) to the final concentration of 0.5 mg/ml and then incubated for 30 min at room temperature with mild shaking (500 rpm). After incubation, the sample was cleared by centrifugation at maximum speed (21,300 *g*) for 30 min at 4 °C. The 800 µl aliquot of cleared sample was loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva) pre-equilibrated with two bed volumes of 10 mM ABP buffer (pH 7.4). The isocratic elution of peptides was run with one bed volume of the 10 mM ABP buffer (pH 7.4) at 0.5 ml/min flow rate at room

temperature, and the progress of elution was monitored by absorbance of peptide bonds at 215 nm. The aggregate (~8-ml elution volume) and monomer (~14-ml elution volume) fractions were isolated and concentrated to 0.2 mg/ml by centrifugation (10,000 *g*, 4 °C) in Amicon Ultra 0.5 ml centrifugal units, MWCO 3000 (Millipore). The total fraction was prepared by diluting the cleared sample with 10 mM ABP (pH 7.4) to 0.2 mg/ml.

3.2.7. Compositional, biophysical, and structural analysis of monomer, aggregate, and total fractions from 10E peptide library

The HPLC amino acid analysis and secondary structure analysis of monomer, aggregate, and total fractions from 10E peptide library were performed, according to the methodologies described in 3.1.2 and 3.1.5, respectively. The isolation and HPLC amino acid analysis of individual fractions were performed three times, and the secondary structure analysis was conducted twice.

The molecular weight distribution was determined by mass photometry analysis using TwoMP system (Refeyn). Before analysis, fraction samples were diluted with MP buffer (10 mM Tris (pH 7.4), 50 mM NaCl) to 20 µg/ml. The mass photometry measurements were performed in silicone gaskets (Grace Bio-Labs) placed on MassGlass CC sample carrier slides (Refeyn) that were cleaned with MilliQ water and isopropanol before the measurements. 17.5 µl of MP buffer was placed in a silicone gasket well to focus the objective and subsequently mixed with 2.5 µl of the sample to achieve the final concentration of 2.5 µg/ml. The images were acquired for 60 s using AcquireMP software (Refeyn), and molecular weight distributions were obtained using DiscoverMP software (Refeyn). Mass calibration was performed using 10 nM bovine serum albumin (BSA) in MP buffer as a standard. The mass photometry analysis was performed twice.

3.3. Characterization of ancestral PTC rRNA-rPeptide interactions

3.3.1. Preparation of ancestral PTC rRNA constructs

In this study, we used two previously described ancestral PTC rRNA constructs: the wild-type big PTC (WT bPTC) rRNA derived from Hsiao *et al.* (2013) and the wild-type small PTC (WT sPTC) rRNA inferred by Bose *et al.* (2022). The sequences of both ancestral PTC rRNA constructs and single-stranded DNA (ssDNA) oligonucleotides used for their synthesis are listed in Tab. 3.2. WT_bPTC ssDNA oligonucleotide was purchased from Integrated DNA Technologies, and the remaining ssDNA oligonucleotides were purchased from Sigma Aldrich.

Tab. 3.2. The sequences of ancestral PTC rRNAs and ssDNA oligonucleotides

Name	Sequence
WT bPTC rRNA	GGUGCCUAUUGAAGCAUGAGCCGGCGACUCAGCCGUAAGGCUGGACCCGAAACCG GGCGAGCUAGCCCUGGCCAGCCGUAAGGCGAACCGGUGGGGGAUGCAAACCCCUC GGAUGAGCUGGGGCUAGGAGUGAAAAGCUAACCGAGCCCGGAGAUAGCUGGUUCU GCCGUAAGGCAGCGUUGCCGUAAGGCAAGUGCGAAUGCCGGCAUGAGUAACGAGC GGGAGAACCCUCGCCAAGGAACUCUGCAAGCCGUAAGGCCUCUGGCGACUGUUUA CCAAAAACACAGCGCCGUAAGGCGCGAGCCGUAAGGCUAACGACCGGAGCGCUGU CUCGGCGAGGGACCCGGUGAAAUUGAACUGGCCGUGAAGAUGCGGCCUACCCGUG GCAGGACGAAAAGACCCCGUGGAGCUUUACUGCCGUAAGGCAGUUUGACUGGGGC GGUCGGCCGUAAGGCUAAAAGUUACCCCGGGGAUAACAGGCUGAUCGCCGUAAGG CGGUUUGGCACCUCGAUGUCGGCUCGUCGCGCCGUAAGGCUUGGGCUGUUCGCCC AUUAAAGCGGCACGCGAGCUGGGUUCAGAACGUCGUGAGACAGUUCGGUCUCUAU CCGCCACGGGCA
WT_bPTC ssDNA	AAAAAGTGGCCTGTAATACGACTCACTATAGGTGCCTATTGAAGCATGAGCCGGC GACTCAGCCGTAAGGCTGGACCCGAAACCGGGCGAGCTAGCCCTGGCCAGCCGTA AGGCGAACCGGTGGGGGATGCAAACCCCTCGGATGAGCTGGGGCTAGGAGTGAAA AGCTAACCGAGCCCGGAGATAGCTGGTTCTGCCGTAAGGCAGCGTTGCCGTAAGG CAAGTGCGAATGCCGGCATGAGTAACGAGCGGGAGAACCCTCGCCAAGGAACTCT GCAAGCCGTAAGGCCTCTGGCGACTGTTTACCAAAAACACAGCGCCGTAAGGCGC GAGCCGTAAGGCTAACGACCGGAGCGCTGTCTCGGCGAGGGACCCGGTGAAATTG AACTGGCCGTGAAGATGCGGCCTACCCGTGGCAGGACGAAAAGACCCCGTGGAGC TTTACTGCCGTAAGGCAGTTTGACTGGGGCGGTCGGCCGTAAGGCTAAAAGTTAC CCCGGGGATAACAGGCTGATCGCCGTAAGGCGGTTTGGCACCTCGATGTCGGCTC GTCGCGCCGTAAGGCTTGGGCTGTTCGCCCATTAAAGCGGCACGCGAGCTGGGTT CAGAACGTCGTGAGACAGTTCGGTCTCTATCCGCCACGGGCA
bPTC_F primer	AAAAAGTGGCCTGTAATACGACTCA
bPTC_R primer	TCTATCCGCCACGGGCA
WT sPTC rRNA	GGAAGACCCCGUGGAGCUCUUCGGAGUUACCCCGGGGAUAACAGGCUGAUCUCUU CGGAGGUUUGGCACCUCGAUGUCGGCUCGUCGCUUCGGCGGCACGCGAGCUGGGU UCAGAACGUCGUGAGACAGUUCGGUC
WT_sPTC_F ssDNA	GCGTAATACGACTCACTATAGGAAGACCCCGTGGAGCTCTTCGGAGTTACCCCGG GGATAACAGGCTGATCTCTTCGGAGGTTTGGCAC
WT_sPTC_R ssDNA	GACCGAACTGTCTCACGACGTTCTGAACCCAGCTCGCGTGCCGCCGAAGCGACGA GCCGACATCGAGGTGCCAAACCTCCGAAGAG

WT bPTC rRNA was prepared from the double-stranded DNA (dsDNA) template obtained by polymerase chain reaction (PCR) amplification of WT_bPTC ssDNA. The PCR amplification was performed with bPTC_F (forward) and bPTC_R (reverse) primers using Q5 High-Fidelity DNA Polymerase (NEB). The reaction mixture was prepared by mixing 50 ng of WT_bPTC ssDNA, 1 µl of 10 mM dNTP Mix (SERVA), 2.5 µl of 10 mM bPTC_F primer, 2.5 µl of 10 mM bPTC_R primer, 10 µl of 5x Q5 Reaction Buffer (NEB) and water to adjust the volume to 50 µl. After 0.5 µl of Q5 High-Fidelity DNA Polymerase (NEB) was added, the reaction mixture was placed into a thermocycler, and PCR was run according to the following

program: initial denaturation at 98 °C for 30 s; denaturation at 98 °C for 10 s; annealing at 69 °C for 30 s; extension at 72 °C for 60 s; final extension at 72 °C for 20 s; 32 cycles.

WT sPTC rRNA was prepared from the dsDNA template obtained by annealing of WT_sPTC_F and WT_sPTC_R ssDNA oligonucleotides and subsequent DNA synthesis using Klenow Fragment of DNA Polymerase I (NEB). The reaction mixture was prepared by mixing 15 pmol of WT_sPTC_F ssDNA, 15 pmol of WT_sPTC_R ssDNA, 5 µl NEB2 Buffer (NEB), and water to adjust the volume to 47 µl. For annealing, the reaction mixture was heated at 80°C for 5 min and gradually cooled to room temperature at 0.1 °C/s rate. After 1 µl of 5 U/µl Klenow Fragment of DNA Polymerase I (NEB) and 2 µl of 10 mM dNTP Mix (SERVA) were added, the reaction mixture was incubated at 25 °C for 1 h followed by heat inactivation of the enzyme at 50 °C for 15 min.

The resulting dsDNA templates were purified from 2% TAE-agarose gel using Monarch Gel Dissolving Buffer (NEB) and Monarch PCR&DNA Cleanup Kit (NEB), according to the manufacturer's instructions. After purification, concentrations of dsDNA templates were determined spectrophotometrically on Nanodrop 2000 (Thermo Fisher Scientific). The purified dsDNA templates were transcribed *in vitro* using HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB). 500 ng of dsDNA template was mixed with 2 µl of T7 RNA Polymerase Mix (NEB), 10 µl of NTP Buffer Mix (NEB), and nuclease-free water to adjust the volume to 20 µl. The reaction mixture was subsequently incubated for 3 h at 37 °C and then mixed with 2 µl of 2 U/µl DNase I (RNase-free) (NEB) and nuclease-free water to adjust the volume to 50 µl. The reaction mixture was incubated for 15 min at 37 °C, and the resulting ancestral PTC rRNA was isolated from the reaction mixture by Monarch RNA Cleanup Kit (NEB), according to the manufacturer's protocol. The concentrations of ancestral PTC rRNAs were determined spectrophotometrically on Nanodrop 2000 (Thermo Fisher Scientific), and their size and purity were confirmed by denaturing urea polyacrylamide gel electrophoresis (urea-PAGE). Before electrophoresis, 200 ng of ancestral PTC rRNA was mixed with 2 µl of RNA Gel Loading Dye (2X) (Thermo Fisher Scientific) and water to adjust the total volume to 4 µl. The sample was subsequently denatured at 70 °C for 10 min and loaded on 5% (for WT bPTC rRNA) or 8% (for WT sPTC rRNA) urea-TBE polyacrylamide gel. The electrophoresis was run at 450 V for approximately 30 min using Nucleic Acid Precast Mini Format Electrophoresis System (Bio-Rad). RiboRuler Low Range RNA Ladder (Thermo Fisher Scientific) was used as a reference. After electrophoresis, gels were soaked in TBE buffer supplemented with GelRed Nucleic Acid Gel Stain (Biotium) for 10 min and visualised by InGenius3 gel documentation system (Syngene). RNA samples were stored at –80 °C before use.

3.3.2. Fluorescent labelling of ancestral PTC rRNA constructs

The 3'-end of ancestral PTC rRNA constructs was labelled with fluorescein-5-thiosemicarbazide (FTSC) using a published protocol (Zearfoss and Ryder 2012) with slight modifications. For oxidation, 500 pmol of ancestral PTC rRNA was mixed with 10 µl of 1 mM $NaIO_4$ (20 equivalents), 10 µl of 500 mM potassium acetate (pH 5.2), and nuclease-free water to adjust the volume to 50 µl. The reaction mixture was incubated for 90 min at room temperature in the dark. The oxidised PTC rRNA was purified from the reaction mixture with Monarch RNA Cleanup Kit (NEB), according to the manufacturer's protocol, and immediately mixed with 10 µl of 15 mM FTSC (300 equivalents) in N,N-dimethylformamide (DMF), 10 µl of 500 mM potassium acetate (pH 5.2) and nuclease-free water to adjust the volume to 50 µl. The reaction mixture was incubated overnight at 4 °C in the dark, and fluorescein-labelled PTC rRNA was subsequently isolated using Monarch RNA Cleanup Kit (NEB), according to the manufacturer's protocol.

3.3.3. Design and chemical synthesis of rPeptides

The sequences of ancestral ribosomal peptides (hereinafter referred to as rPeptides) were taken from the proto-ribosome model described in Hsiao *et al.* (2013) and were modified for this study (Tab. 3.3). The rPeptides were commercially synthesized by Spyder Institute Praha (Czech Republic), according to the standard Fmoc-based solid-phase methodology, and their identity was confirmed by MALDI-TOF analysis, according to the protocol described in 3.1.2.

Tab. 3.3. The sequences of rPeptides

Name	Sequence
pL2	GRRPHVRGAAMNPVDHPHGGGEGRAPRGR
pL3	RHPGSIGNRKTPGRVYKG
pL4	YSGRKIWPQKHTGRARHGDIGAPIFV
pL15	PGSGHGKTATRGHKGQKSRSGGLKD
pL22	LKRVLPRARGRADIIKKR

3.3.4. Ancestral PTC rRNA-rPeptide interaction analysis by microscale thermophoresis

The interaction of five rPeptides with ancestral PTC rRNA constructs was characterized by microscale thermophoresis. Prior to analysis, ancestral PTC rRNAs were denatured at 85 °C for 30 s, gradually cooled to room temperature for over 30 min, and diluted in (i) 20 mM Tris-HCl (pH 7.5); or (ii) 20 mM Tris-HCl (pH 7.5), 2 mM KCl, 2 mM $MgCl_2$, and 2 mM $CaCl_2$ to the working concentration of 100 nM. The stock solutions of rPeptides were prepared by dissolving them in 20 mM Tris-HCl (pH 7.5) to the working concentration of 1 mM.

MST measurements were performed on a Monolith NT.115 system (NanoTemper Technologies). 15 rPeptide solutions were prepared by 2-fold serial dilution of the rPeptide stock solution with the rPeptide concentrations ranging from 1 mM to 61.04 nM. 10 µl of rPeptide solutions were subsequently mixed with 10 µl of 100 nM fluorescein-labelled PTC rRNA, and the resulting samples were incubated for 15 min at room temperature in the dark. After incubation, the samples were placed into standard-treated capillaries (NanoTemper Technologies) and subjected to MST analysis, according to the following setup: 1 % LED power (green channel), 40 % MST power, laser-on time 20 s, and laser-off time 5 s. The normalised fluorescence values were derived by T-jump analysis and plotted against the rPeptide concentrations. The resulting curves were fitted into the one-site binding and cooperative binding models to calculate the dissociation constants and apparent binding stoichiometries. The analysis was conducted twice.

4. Results

4.1. Exploration of biophysical and structural properties of early, alternative early, and modern amino acid alphabets

The results in this section were included in the attached article 1 - Makarov M., Sanchez Rocha A.C., Krystufek R., Cherepashuk I., Dzmitruk V., Charnavets T., Faustino A.M., Lebl M., Fujishima K., Fried S.D., Hlouchova K. Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. J. Am. Chem. Soc. ***145****:9, 5320–9 (2023).*

In this study, we assessed the biophysical and structural properties of six 25-mer combinatorial peptide libraries. The amino acid composition of individual peptide libraries is depicted in Fig. 4.1.

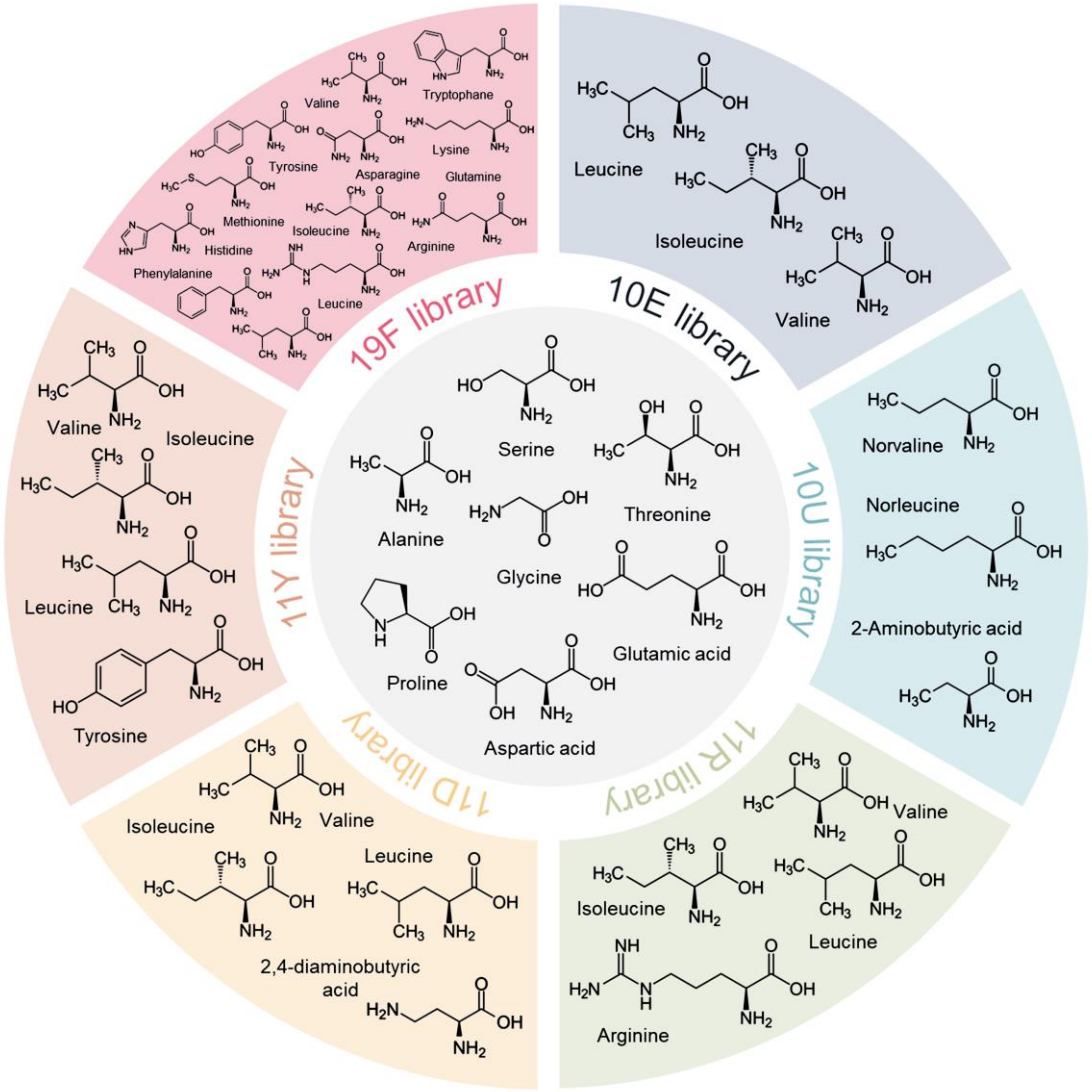


Fig. 4.1. The amino acid composition of 25-mer combinatorial peptide libraries.

The specific aim of this study was to compare the solubility, aggregation propensity, and secondary structure conferred by (i) the early canonical amino acid alphabet (10E library);

(ii) alternative early versions of the amino acid alphabet containing prebiotic non-canonical amino acids (10U, 11D libraries); and (iii) evolutionary later versions of the amino acid alphabet containing late canonical amino acids (11Y, 11R, 19F libraries). Unbranched aliphatic non-proteinogenic Aba, Nva, and Nle (10U library) were chosen as prebiotic alternatives of branched aliphatic proteinogenic Val, Leu, and Ile (10E library). Positively charged non-canonical Dab (11D library) was selected as a prebiotic alternative to positively charged canonical Arg (11R library), which was not presumably available in the prebiotic environment. Tyr was chosen as a representative of canonical aromatic amino acids recruited in the later stages of the genetic code evolution (11Y library). Cys was excluded from the modern amino acid subset (19F library) due to potential complications in solid-phase synthesis and subsequent analyses caused by its oxidation.

The peptide libraries were commercially synthesized using the Fmoc-based solid-phase approach from isokinetic mixtures of Fmoc-protected amino acids. The quality of peptide libraries was subsequently assessed by HPLC amino acid analysis and mass spectrometry analysis. The MALDI-TOF spectra were compared with the theoretical molecular weight distributions generated by MSc. Ivan Cherepashuk for 100,000 random peptide sequences (Fig. 4.2). The comparison confirmed the correct distribution and range of molecular weights in all synthesized libraries.

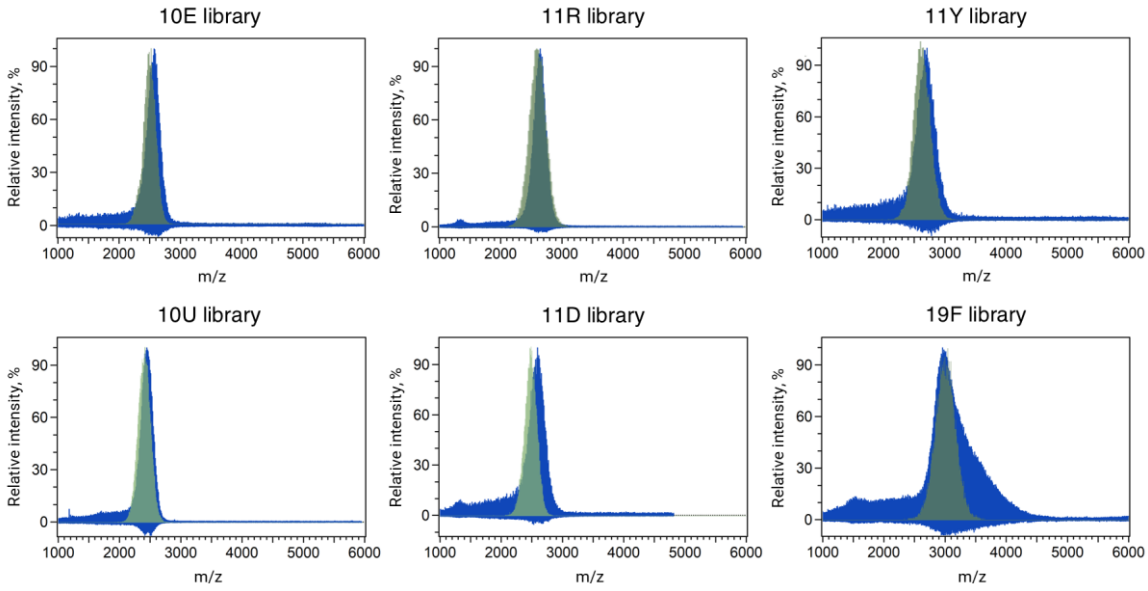


Fig. 4.2. MALDI-TOF spectra of 25-mer combinatorial peptide libraries. The spectra were obtained for peptide libraries dissolved in 50 % (v/v) acetonitrile + 0.1 % (v/v) acetic acid using DHB as a matrix. 19F peptide library was dissolved in 50 % (v/v) acetonitrile. Experimentally obtained m/z distribution is shown in blue, and theoretical mass distribution is shown as a green overlay.

HPLC amino acid analysis revealed close-to-equimolar amino acid compositions for all peptide libraries (Fig. 4.3). The moderate deviations from the equimolar amino acid ratios are most likely caused by the different reactivity of Fmoc-protected amino acids in the isokinetic mixtures during synthesis. The content of Asn, Gln, and Trp in 19F peptide library was not estimated due to the specifics of the standard sample preparation. Prior to the analysis, peptides and proteins are hydrolysed by concentrated HCl into individual amino acids, which are subsequently derivatised with OPA or ninhydrin for fluorescent or spectrophotometric detection, respectively. The acidic hydrolysis leads to the irreversible deamidation of Asn and Gln into respective Asp and Glu and oxidative degradation of Trp (Shih and Kalmar 1984; la Cour *et al.* 2019). The significantly high Asp content in 11D library is plausibly caused by its co-elution with Dab, which was not detected in the analysis. The presence of Dab in 11D library was additionally confirmed by fluorescamine assay that was performed by MSc. Robin Krystufek. The assay is based on the reaction of amine-reactive non-fluorescent fluorescamine with primary amino groups present at the N-terminus and side chains of peptides and proteins, which leads to the formation of a fluorescent product (Udenfriend *et al.* 1972). 11D library showed a higher fluorescence intensity in comparison with the 10E library in the fluorescamine assay, indicating the presence of Dab (data are not included in the thesis).

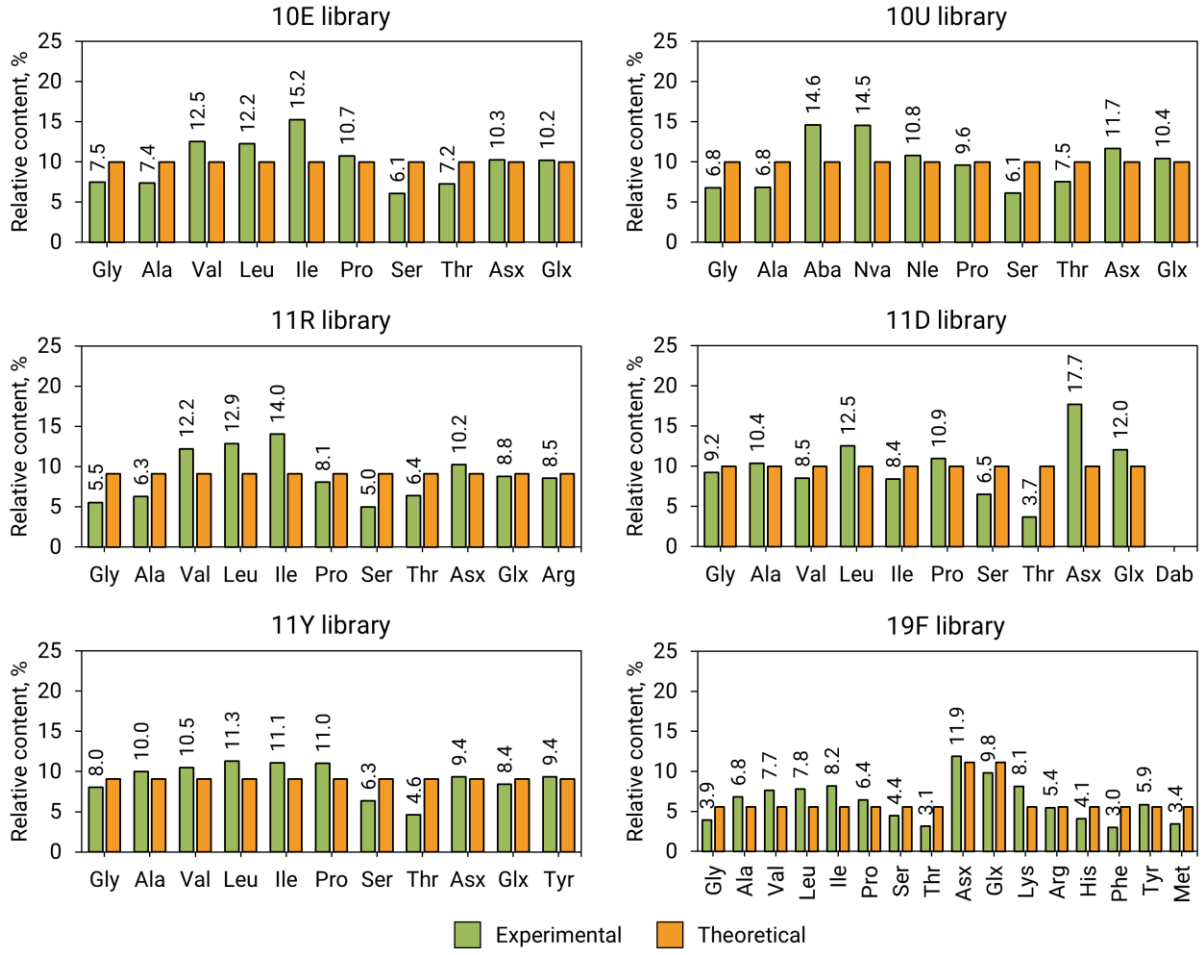


Fig. 4.3. Amino acid composition of 25-mer combinatorial peptide libraries. The experimentally obtained amino acid ratios (green) were determined by HPLC amino acid analysis. The theoretical amino acid ratios (orange) represent the equimolar distribution of amino acids in peptide libraries. Asp and Asn are denoted as Asx, and Glu and Gln are denoted as Glx.

4.1.1. Biophysical analysis of combinatorial peptide libraries

Following the synthesis and quality control, peptide libraries were characterized in terms of their biophysical properties, namely solubility and aggregation propensity. Given the considerable uncertainty surrounding the prebiotic environment conditions (salinity, pH, temperature), the relative solubility and aggregation propensity of peptide libraries were tested in a wide range of pH (3–11) and ionic strengths (eq. 50 mM or 500 mM NaCl) (Kempe and Kazmierczak 2002; Maruyama *et al.* 2013; Halevy and Bachan 2017).

For the biophysical analysis, the water suspensions of peptide libraries were diluted into a series of 20 mM ABP buffers to 0.5 mg/ml, incubated at room temperature and centrifuged to separate soluble and insoluble fractions. The relative solubility of all peptide libraries (except 19F library) was assessed by measuring the absorption of peptide bonds in soluble fractions of peptide libraries by UV-Vis absorption spectroscopy. The peptide bonds display an absorption maximum near 190 nm; however, absorption of peptide bonds was measured at 215 nm due to

the possible interference of buffer components at lower wavelengths (Rosenheck and Doty 1961; Tong *et al.* 2020). Due to extremely low solubility, the relative solubility of 19F peptide library was estimated by fluorescence spectroscopy using an amine-reactive fluorogenic reagent. The chemical composition of the reagent is not specified by the manufacturer. Prior to analysis, the calibration curves were established from 2-fold serial dilutions of the peptide library samples dissolved in 20 mM ABP (pH 11) at low ionic strength to ensure the linear response of absorbance at 215 nm on the peptide concentration at the selected concentration range (0–0.5 mg/ml) (Fig. 4.4). The calibration curve for 19F library was established from a 2-fold serial dilution of 19F library dissolved in DMSO to confirm the linear response of fluorescence intensity on the peptide concentration at the selected concentration range (0–0.5 mg/ml).

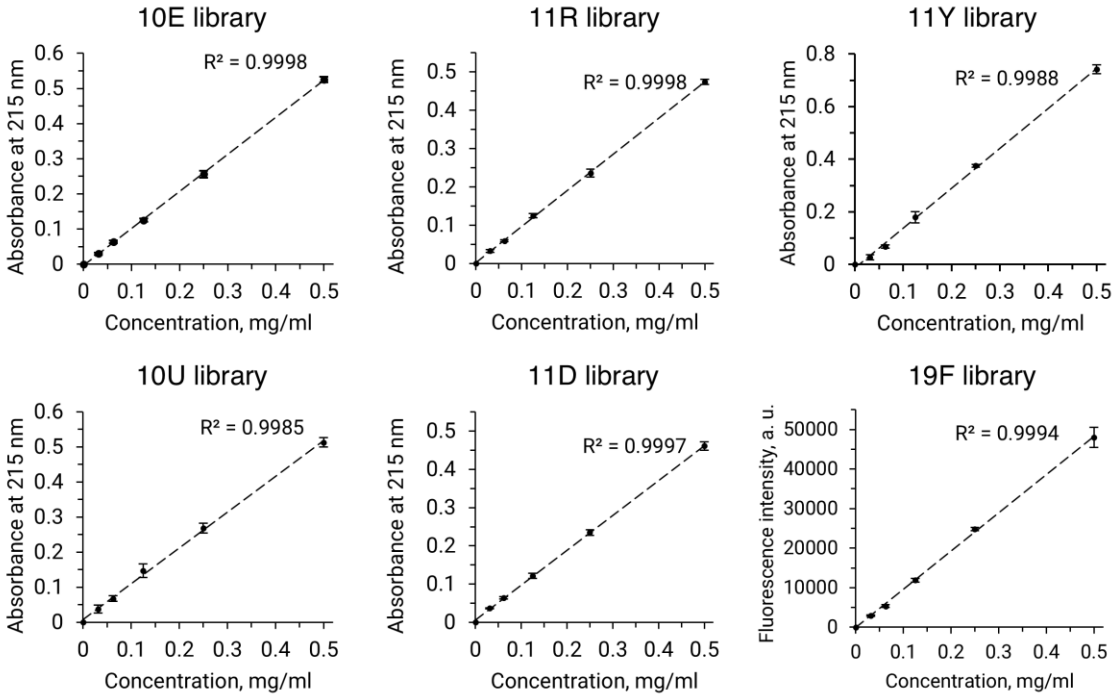


Fig. 4.4. Calibration curves for 25-mer combinatorial peptide libraries. The calibration curves were obtained from 2-fold serial dilutions of peptide libraries dissolved in 20 mM ABP buffer (pH 11) at low ionic strength by UV-Vis absorption spectroscopy. The calibration curve for 19F peptide library was obtained from 2-fold serial dilutions of the library solution in DMSO by fluorescence spectroscopy. Error bars correspond to standard deviations derived from technical quadruplicates (19F) or (10E, 10U, 11R, 11D, 11Y) sextuplicates.

The solubility analysis results are depicted in Fig. 4.5. The relative solubility values were obtained by dividing the absorbance at 215 nm measured at the given conditions (pH, ionic strength) by the highest absorbance at 215 nm observed across the sample set of a specific library. The relative solubility values for 19F peptide library were obtained by dividing the fluorescence intensity measured at the given conditions (pH, ionic strength) by the fluorescence

intensity measured for 0.5 mg/ml 19F library solution in DMSO. The Student's t-test was used to estimate the statistically significant differences in the relative solubilities of peptide libraries at low and high ionic strengths.

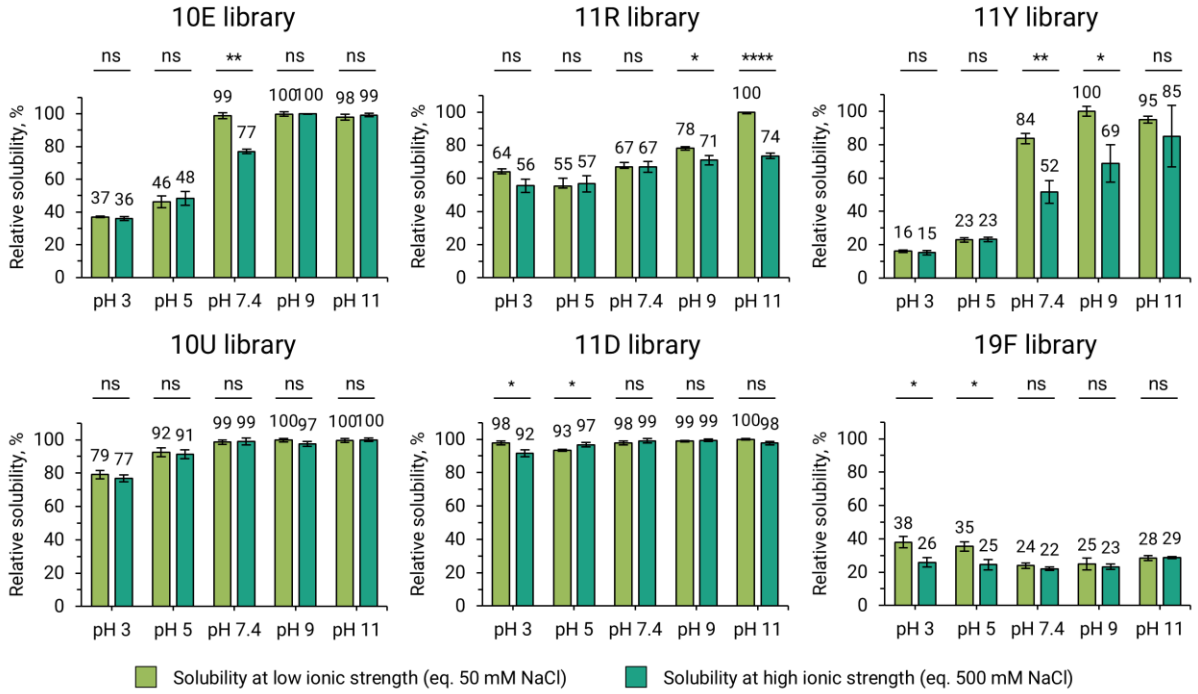


Figure 4.5. Relative solubility of 25-mer combinatorial peptide libraries at different ionic strengths and pHs. Solubility was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at low and high ionic strengths. Solubility of 19F peptide library was determined by fluorescence spectroscopy, and solubility of the other peptide libraries was determined by UV–Vis absorption spectroscopy. Error bars correspond to standard deviations derived from three analyses. The statistical difference was assessed by Student's t-test (ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$).

The solubility analysis results demonstrate the opposite effects of pH and ionic strength on the solubility of peptide libraries. Increasing ionic strength generally decreases the solubility of peptide libraries, whereas increasing pH improves the solubility of all peptide libraries except for 19F library. The most drastic influences of pH and ionic strength were observed for 11Y library, whereas the solubility of 11D library is almost invariant on pH and ionic strength. The comparison of individual peptide libraries indicates that 19F library composed of the full amino acid set is drastically less soluble than the rest of the peptide libraries that consist of reduced amino acid subsets. Peptide libraries containing prebiotic non-proteinogenic amino acids (10U and 11D) show significantly higher solubility than peptide libraries containing their proteinogenic alternatives (10E and 11R, respectively). On the other hand, peptide libraries that include late proteinogenic amino acids (11R, 11Y, and 19F) are less soluble than the peptide library composed solely of early proteinogenic amino acids (10E).

The propensity to form higher-order soluble aggregates across the same range of pH and ionic strengths was tested by analytical size-exclusion chromatography. The soluble fractions of peptide libraries were loaded onto the chromatography column, and the progress of elution was monitored by absorption of peptide bonds at 215 nm. The resulting chromatograms are depicted in Fig. 4.6 and Fig. 4.7.

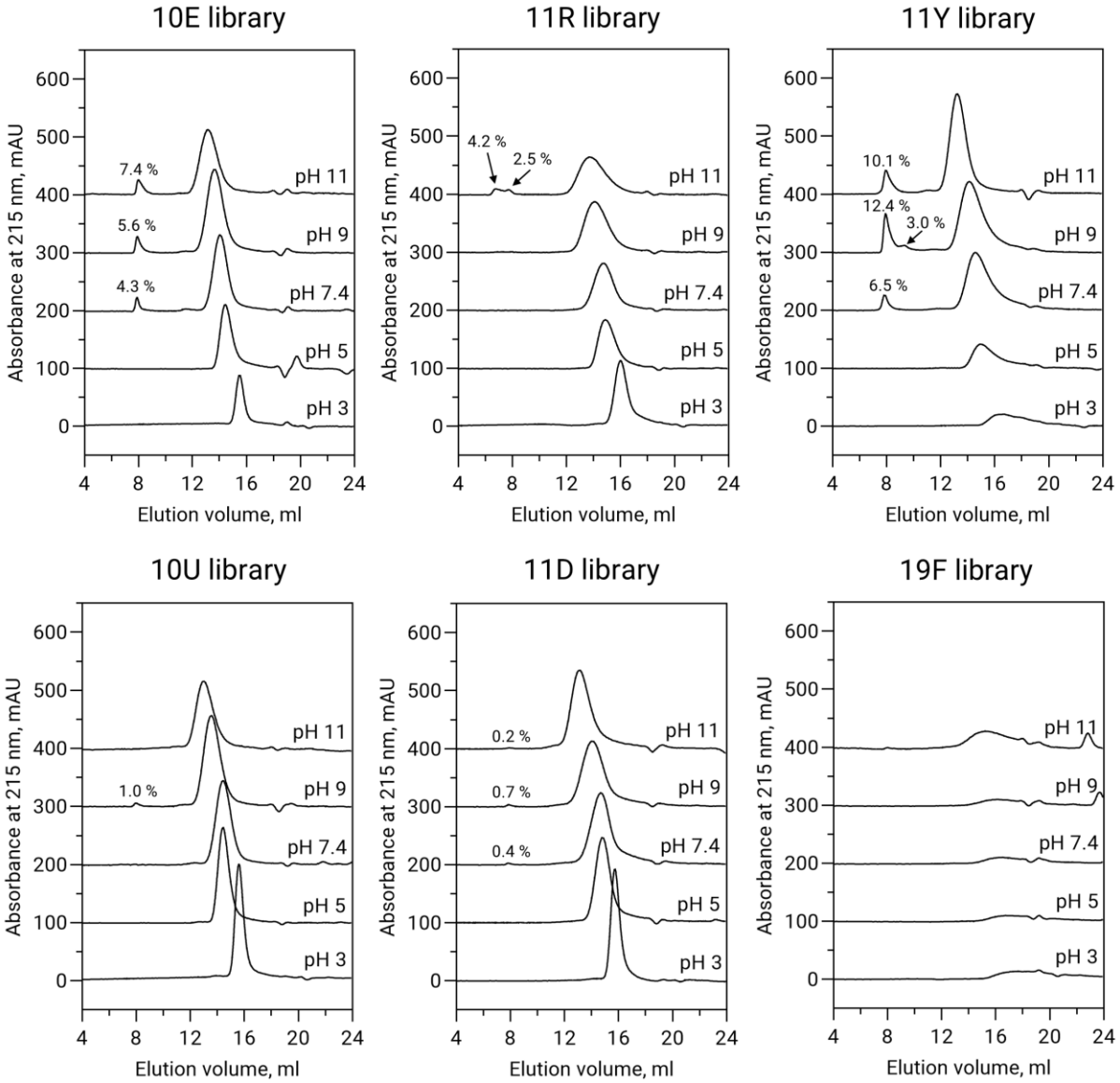


Fig. 4.6. Aggregation propensity of 25-mer combinatorial peptide libraries. Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at low ionic strength by analytical size-exclusion chromatography. Chromatograms report absorbance at 215 nm as a function of elution volume.

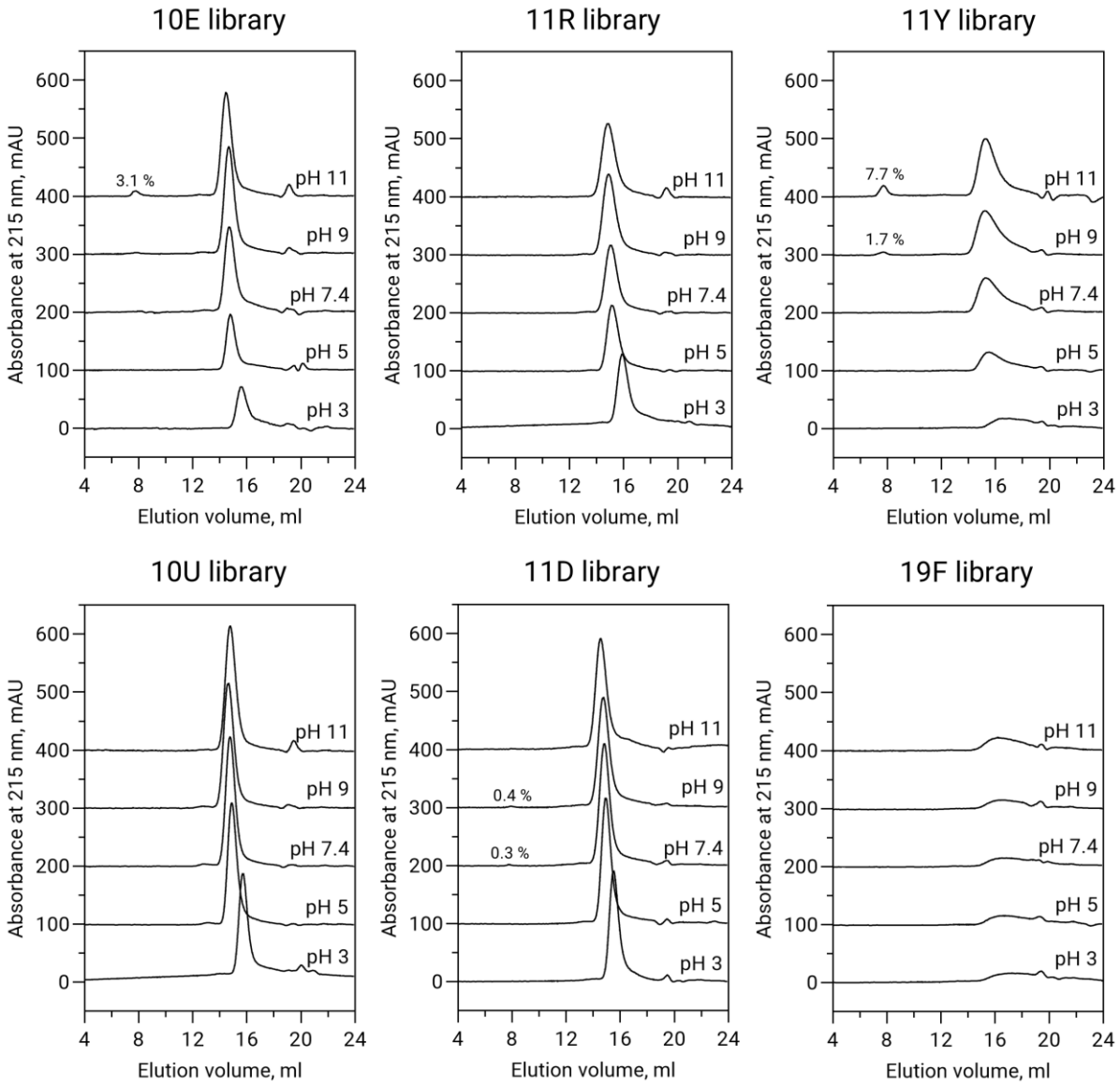


Fig. 4.7. Aggregation propensity of 25-mer combinatorial peptide libraries. Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at high ionic strength by analytical size-exclusion chromatography. Chromatograms report absorbance at 215 nm as a function of elution volume.

In general, chromatograms for all peptide libraries (except for 19F) are characterized by the presence of one major elution peak between 13- and 17-ml elution volumes, which corresponds to the monomer fraction of peptide libraries. The intensities of monomer peaks are proportional to the solubilities of peptide libraries measured by UV-Vis absorption spectroscopy. Due to extremely low solubility, the monomer peaks for 19F library are not well-defined on the corresponding chromatograms. At low ionic strength, the monomer peaks shift toward lower elution volumes and show a broader distribution with increasing pH. Increasing ionic strength reduces the elution volume shift and leads to narrower and more well-defined monomer peaks. Most of the chromatograms obtained at near-neutral and alkaline pH are

characterized by the presence of a second minor peak at ~8 ml elution volume, which corresponds to the aggregate fraction of peptide libraries. The aggregate peaks are not present on the chromatograms obtained at acidic pH. Significant (more than 1 %) formation of higher-order soluble aggregates is detected for peptide libraries composed entirely of canonical amino acids (10E, 11R, and 11Y). The aggregation propensity conferred by canonical amino acid subsets is significantly reduced at high ionic strength. In contrast, peptide libraries containing non-canonical amino acids (10U and 11D) demonstrate low (less than 1 %) aggregation propensity.

4.1.2. Structural analysis of combinatorial peptide libraries

Following the biophysical analysis, peptide libraries were analysed by far-UV CD spectroscopy to determine their secondary structure content. To estimate the influence of pH on the secondary structure, the water suspensions of peptide libraries were diluted into a series of 10 mM ABP buffers (pH 3–11), incubated at room temperature and centrifuged in order to remove insoluble fraction. The CD spectra of soluble fractions were subsequently recorded in the far-UV range, from 190 to 260 nm (Fig. 4.8). The effect of ionic strength on the secondary structure was not determined in this study due to the strong absorption of chloride ions below 200 nm (Tong *et al.* 2020). Given the high degeneracy and the unknown distribution of individual sequences in peptide libraries, the obtained CD signals could not be converted into molar residue ellipticities, which is used for the deconvolution analysis of CD spectra. Instead, the secondary structure content was estimated by the ratios of the negative CD signal at 222 nm arising from α-helices and β-sheets and the negative CD signal at 200 nm arising from unstructured regions (random coils) of peptides and proteins (Kelly *et al.* 2005).

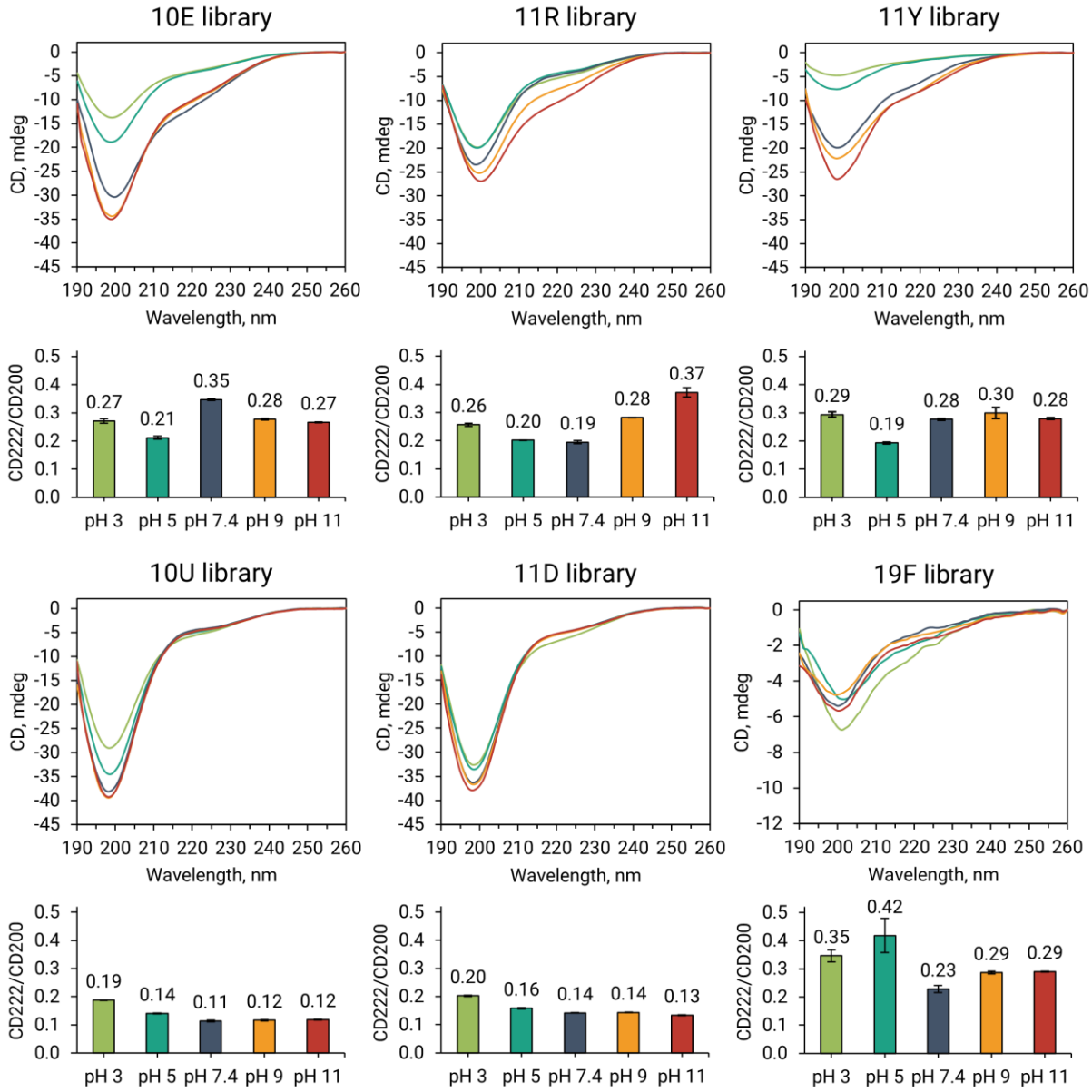


Fig. 4.8. Effect of pH on the secondary structure of 25-mer combinatorial peptide libraries. Far-UV CD spectra were measured for 0.2 mg/ml nominal peptide libraries solutions in 10 mM ABP buffers (pH 3–11). The inset graphs show the ratios of CD signals at 222 to 200 nm. Errors bars correspond to standard deviations derived from two analyses.

The intensities of CD signals correlate with the solubilities of peptide libraries obtained by UV-Vis absorption spectroscopy. All far-UV CD spectra are characterised by a pronounced minimum at 198 nm, indicating the large degree of disorder in analysed peptide libraries. However, the far-UV CD spectra of peptide libraries composed solely of proteinogenic amino acids (10E, 11R, 11Y) demonstrate a broad distinct negative signal between 210 and 240 nm, confirming the partial presence of peptides in α-helical and/or β-sheet conformation. The peptide libraries containing non-proteinogenic amino acids (10U and 11D) are less structured than libraries containing their respective proteinogenic alternatives (10E and 11R) across the

entire pH range, as indicated by CD222/CD200 ratios. At respective pH, the CD222/CD200 ratios are 1.4–3.2 times lower for 10U library relative to 10E library and 1.3–2.8 times lower for 11D library compared to 11R library. The CD222/CD200 ratios of peptide libraries with late canonical amino acids (11R and 11Y) are comparable to those of the peptide library composed solely of early canonical amino acids (10E) with notable deviations at pH 7.4. and 11. The reliable assessment of secondary structure content in 19F library is hindered by the low intensity of CD signals resulting from the low solubility of 19F library. Overall, there is no general correlation between the secondary structure content in peptide libraries and pH.

Given the high prevalence of unstructured sequences in peptide libraries, far-UV CD spectra were additionally measured at different TFE concentrations (Fig. 4.9). TFE is an organic solvent commonly used in CD and nuclear magnetic resonance (NMR) studies to induce the secondary structure formation in peptides (Buck 1998). To estimate the effect of TFE on the secondary structure, the water suspensions of peptide libraries were diluted into a series of 10 mM ABP buffers (pH 7.4) supplemented with 0–90 % (v/v) TFE to 0.2 mg/ml, incubated at room temperature and centrifuged in order to remove insoluble fraction. The far-UV CD spectra of soluble fractions were subsequently recorded in the 190–260 nm wavelength range. Similarly to the pH-dependence analysis, the secondary structure content was estimated by the ratios of CD signals at 222 to 200 nm.

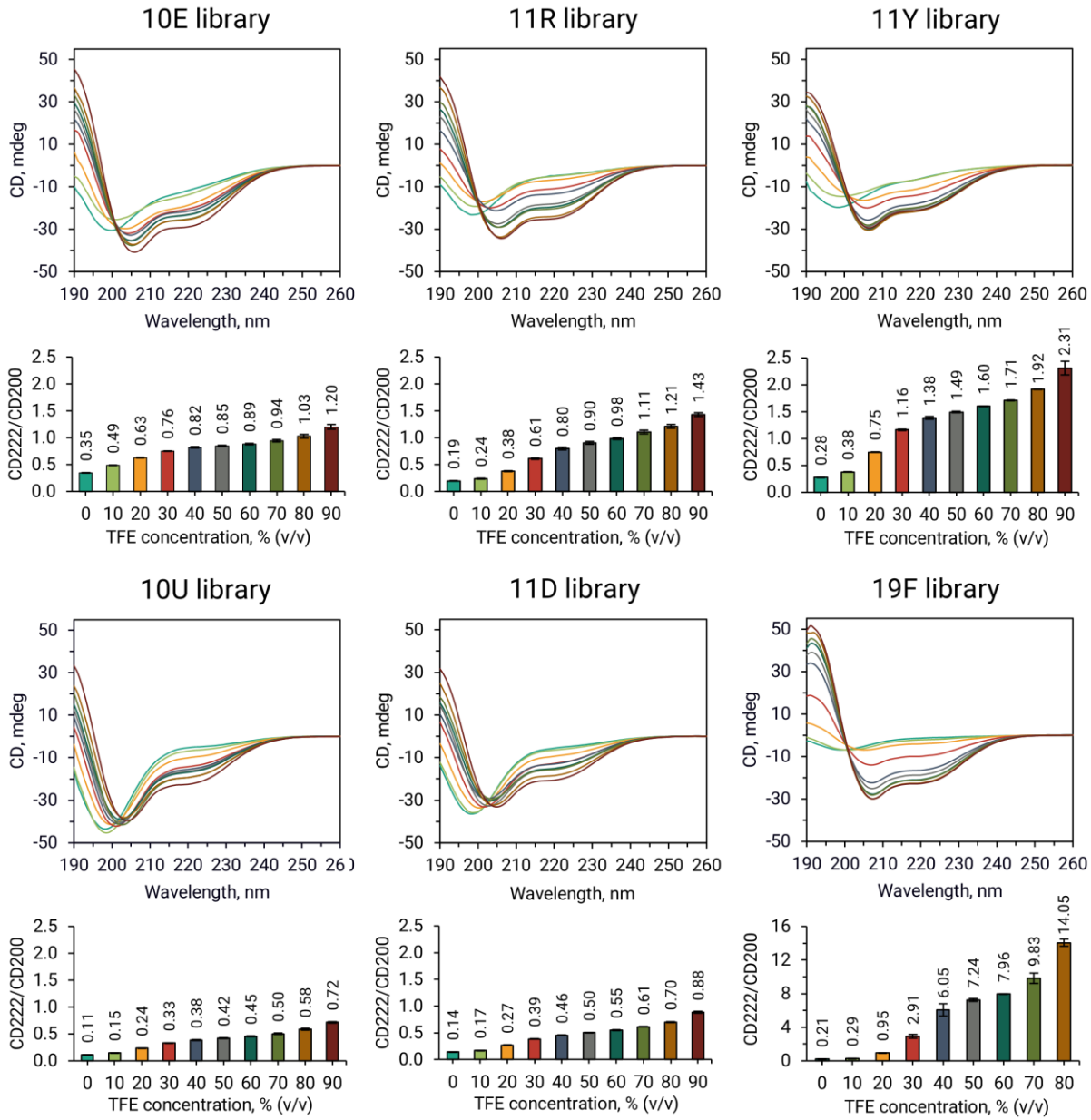


Fig. 4.9. Effect of TFE on the secondary structure of 25-mer combinatorial peptide libraries. Far-UV CD spectra were measured for 0.2 mg/ml nominal peptide libraries solutions in 10 mM ABP buffers (pH 7.4) containing 0–90% (v/v) TFE. The inset graphs show the ratios of CD signals at 222 and 200 nm. Errors bars correspond to standard deviations derived from two analyses.

Upon TFE addition, all peptide libraries adopt predominantly α-helical conformation as indicated by the gradual shift of the minimum from 198 nm to 205–210 nm, increase in the intensity of the broad negative CD signal between 210 and 240 nm, and formation of the distinct maximum around 190 nm (Kelly *et al.* 2005). The overall increase in the intensity of CD signals is additionally caused by the dissolution of partially insoluble peptide libraries (11R, 11Y, 19F) in TFE. Peptide libraries containing non-proteinogenic amino acids (10U and 11D) demonstrate a lower secondary structure-forming propensity than peptide libraries with respective proteinogenic amino acids (10E and 11R) even upon additional stabilisation by TFE,

which is consistent with the observations from pH-dependence analysis. However, in contrast to changes in pH, TFE titration reveals drastic differences in the secondary structure-forming propensity of peptide libraries containing late canonical amino acids (11R, 11Y, 19F). Upon TFE addition, the CD222/CD200 ratios of 10E library are comparable to those of 11R library but significantly lower than those of 11Y and 19F libraries. 19F library shows the highest secondary structure-forming propensity out of all tested peptide libraries.

4.2. Exploration of biophysical and structural properties of differently charged amino acid alphabets

The results in this section were included in the attached manuscript 2 – Makarov M., Kryštůfek R., Manriquez-Sandoval E., Dutta S., Kalvoda T., Kormaník J.M., Chernavets T., Lébl M., Brown S.M., Konvalinka J., Rulíšek L., Singharoy A., Fried S.D., Hlouchová K. Self-assembly of random primordial acidic peptides as a stepping stone to foldable polymers. Submitted to Chem. (2024).

In this study, we analysed the biophysical and structural properties of eight 25-mer combinatorial peptide libraries. The amino acid composition of individual peptide libraries is illustrated in Fig. 4.10. The 8E+Asp,Glu library mentioned in this study corresponds to 10E library previously characterised in section 4.1.

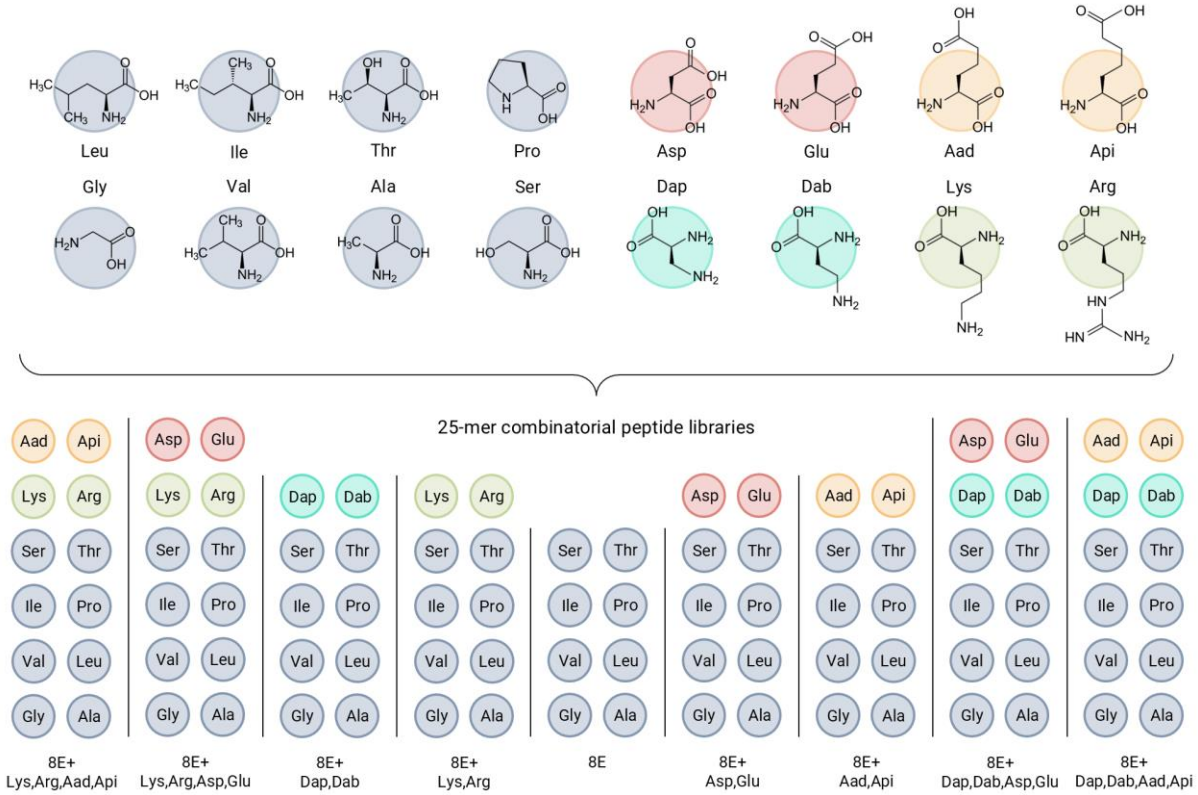


Fig. 4.10. The amino acid composition of 25-mer combinatorial peptide libraries. The three letter codes for non-canonical amino acids are as follows: Dap: 2,3-diaminopropionic acid; Dab: 2,4-diaminobutyric acid; Aad: 2-aminoadipic acid; Api: 2-aminopimelic acid.

The specific goal of this study was to determine the effect of positively and negatively charged amino acids on the solubility, aggregation propensity, and secondary structure of 25-mer combinatorial peptide libraries. All analysed peptide libraries share the same subset of eight early uncharged canonical amino acids (8E: Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr). 8E+Lys,Arg library is supplemented with two basic canonical amino acids, whereas 8E+Dap,Dab library includes their two prebiotically available alternatives with shorter side chains. 8E+Asp,Glu (10E) library contains two acidic canonical amino acids, whereas

8E+Aad,Api library includes their two homologs with longer side chains. The remaining peptide libraries contain various combinations of two basic and two acidic amino acids.

The peptide libraries were commercially synthesized using standard Fmoc-based solid phase peptide synthesis from the isokinetic mixtures of Fmoc-protected amino acids. Following the chemical synthesis, HPLC amino acid analysis and mass spectrometry analysis were used to establish the quality of peptide libraries. The MALDI-TOF spectra were compared with the theoretical mass distributions produced by MSc. Sean M. Brown for 1,000,000 random peptide sequences (Fig. 4.11). The comparison indicated the close-to-correct range and distribution of molecular weights in all synthesized libraries.

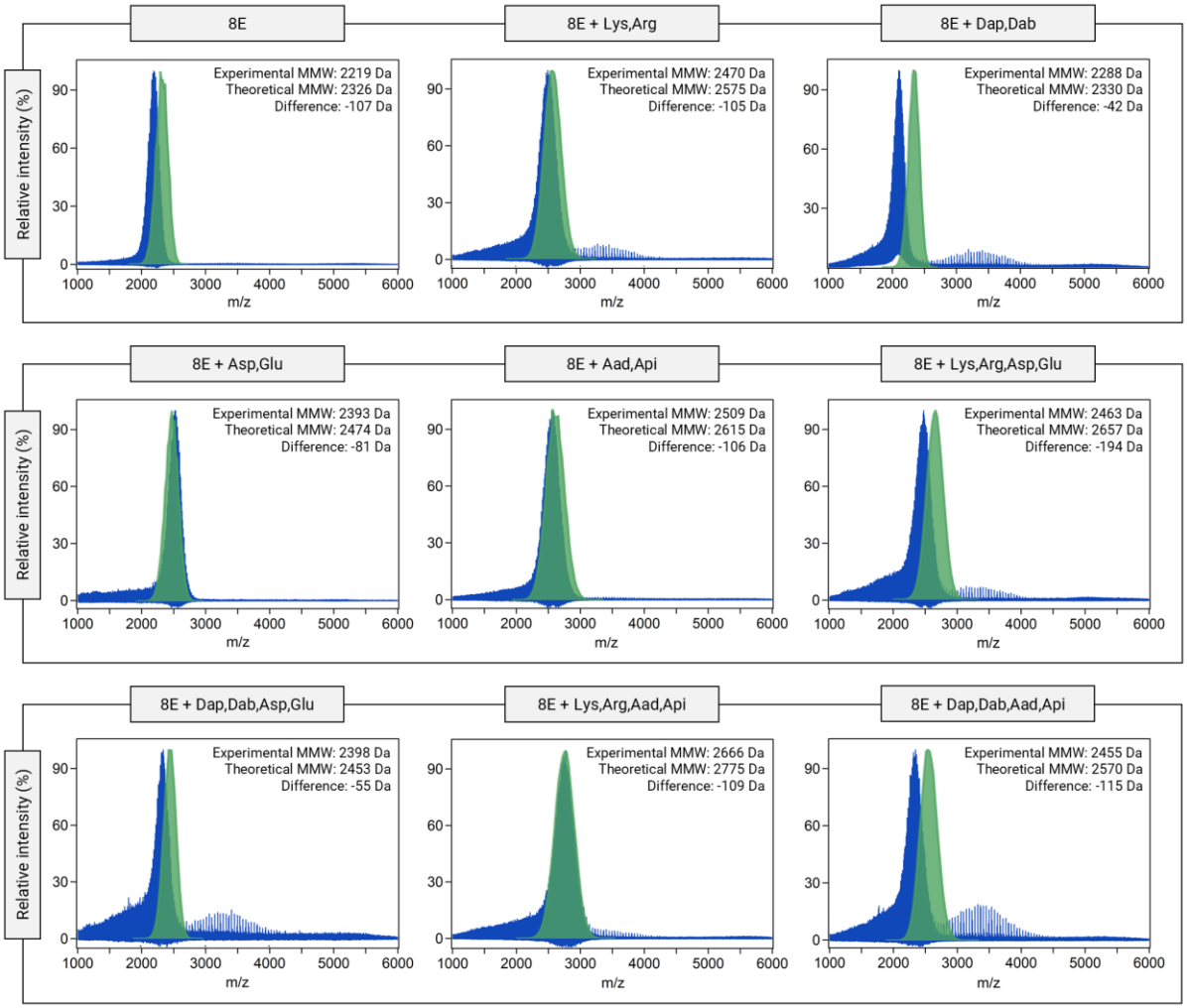


Fig. 4.11. MALDI-TOF spectra of 25-mer combinatorial peptide libraries. The spectra were obtained for peptide libraries dissolved in 50 % (v/v) acetonitrile using DHB as a matrix. Experimentally obtained m/z distribution is shown in blue, and theoretical mass distribution is shown as a green overlay. Mean molecular weights (MMW) were calculated from both distributions. The data for 8E+Asp,Glu (10E) peptide library were taken from the previous study (Makarov *et al.* 2023).

The experimental protocol for HPLC amino acid analysis was modified by Dr. Radko Soucek for the additional identification of non-canonical amino acids (Dap, Dab, Aad, Api).

For methodological consistency, 8E+Asp,Glu (10E) peptide library was re-analysed using the modified protocol. The analysis results demonstrate nearly uniform amino acid distribution in all peptide libraries (Fig. 4.12). The moderate deviations from the equimolar amino acid ratios most likely result from the different reactivity of Fmoc-amino acids in the isokinetic mixtures during synthesis.

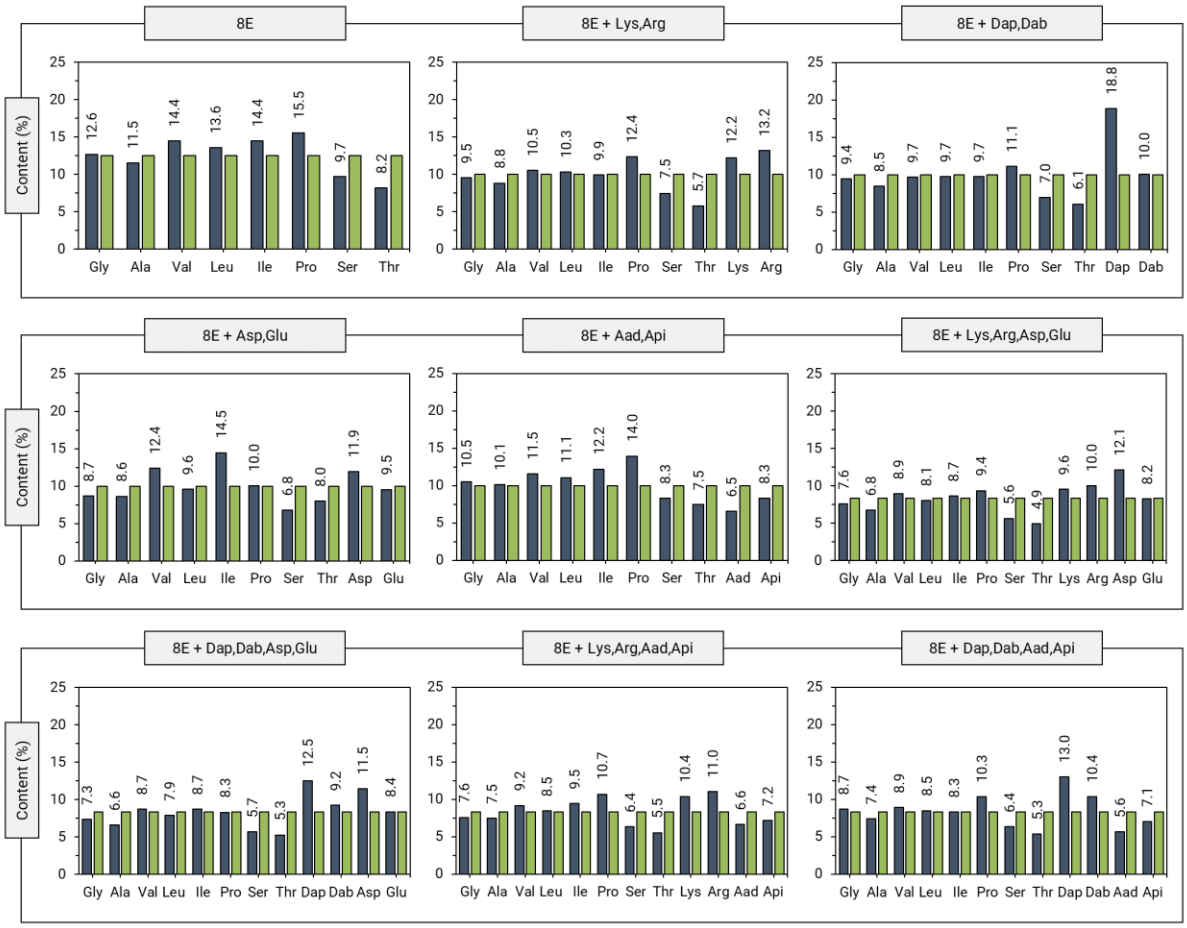


Fig. 4.12. Amino acid composition of 25-mer combinatorial peptide libraries. The experimentally obtained amino acid ratios (blue) were determined by HPLC amino acid analysis. The theoretical amino acid ratios (green) represent the equimolar distribution of amino acids in peptide libraries. Asp and Asn are denoted as Asx, and Glu and Gln are denoted as Glx.

4.2.1. Biophysical analysis of combinatorial peptide libraries

After the chemical synthesis and quality control, peptide libraries were tested for their solubility and aggregation propensity. Similarly to the previous study, the relative solubility and aggregation propensity of peptide libraries were characterized in a wide range of pH (3–11) and ionic strengths (eq. 50 or 500 mM NaCl).

Solubility analysis of peptide libraries was performed, according to the experimental setup previously described in 4.1.1. The relative solubility of all peptide libraries (except 8E) was analysed by UV-Vis absorption spectroscopy. Due to relatively low solubility, 8E library

was analysed using fluorescamine assay. The calibration curves established prior to analysis are illustrated in Fig. 4.13.

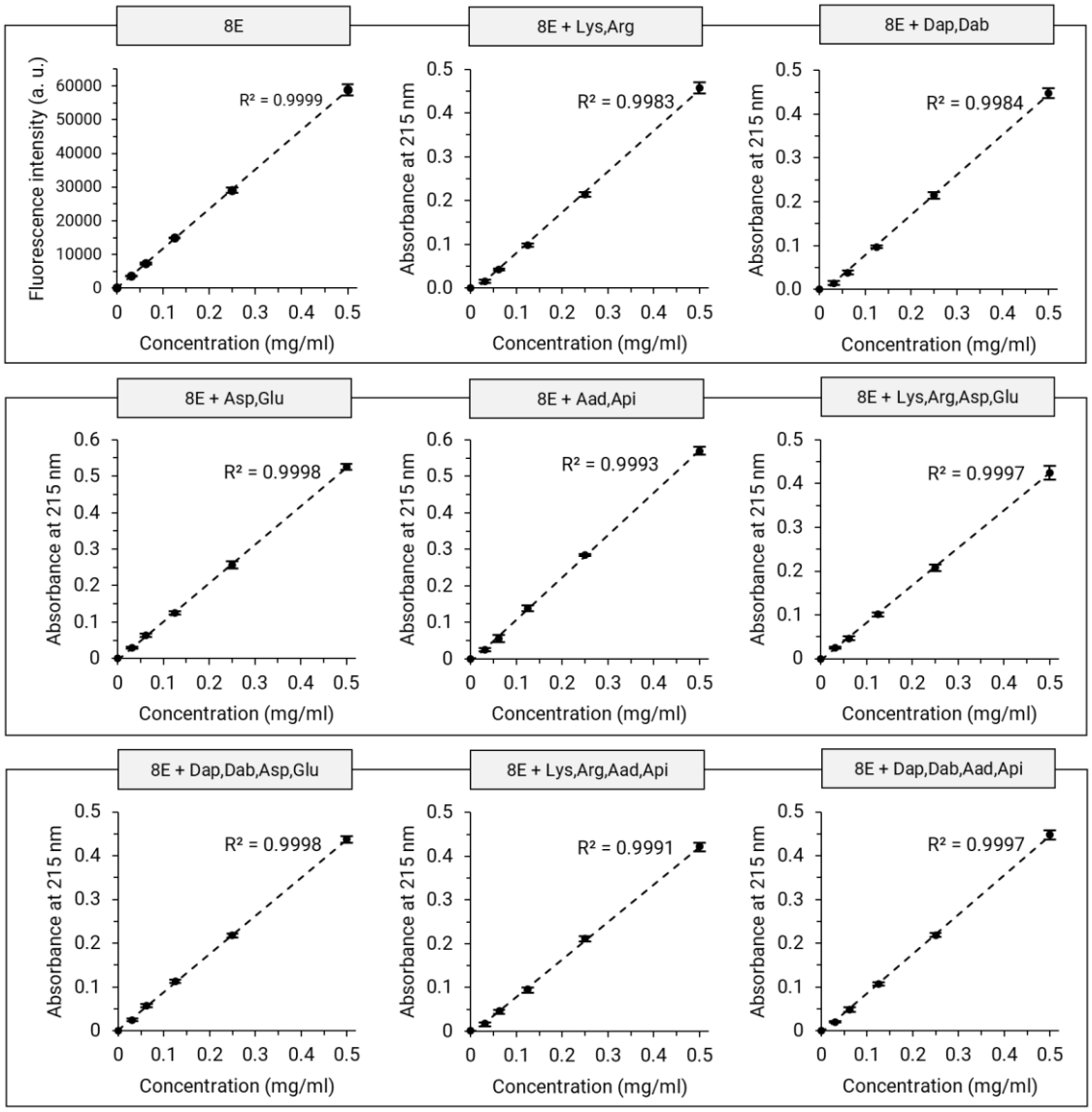


Fig. 4.13. Calibration curves for 25-mer combinatorial peptide libraries. The calibration curves were obtained from 2-fold serial dilutions of peptide libraries dissolved in 20 mM ABP buffer (pH 11) at low ionic strength by UV-Vis absorption spectroscopy. The calibration curve for 8E peptide library was obtained from 2-fold serial dilutions of the library solution in DMSO by fluorescence spectroscopy. Error bars correspond to standard deviations derived from technical quadruplicates (8E) or sextuplicates for the remaining peptide libraries. The data for 8E+Asp,Glu (10E) peptide library were taken from the previous study (Makarov *et al.* 2023).

The relative solubilities of peptide libraries are depicted in Fig. 4.14.

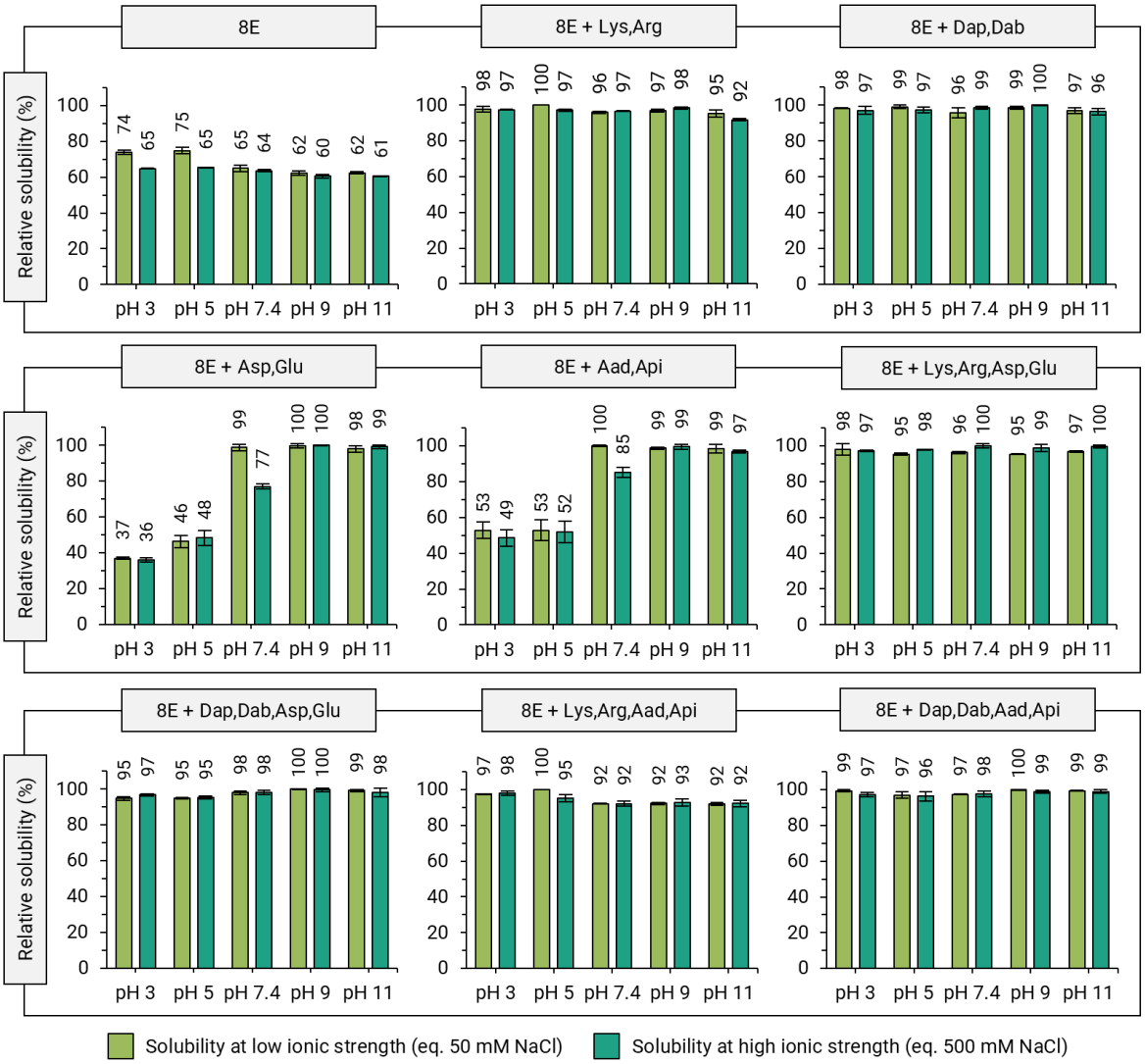


Fig. 4.14. Relative solubility of 25-mer combinatorial peptide libraries at different ionic strengths and pHs. Solubility was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at low and high ionic strengths. Solubility of 8E peptide library was determined by fluorescamine assay, and solubility of the other peptide libraries was determined by UV–Vis absorption spectroscopy. Error bars correspond to standard deviations derived from two analyses. The data for 8E+Asp,Glu (10E) library were taken from the previous study (Makarov *et al.* 2023).

The significant influence of pH on the solubility is observed only for peptide libraries supplemented with two negatively charged amino acids (8E+Asp,Glu (10E) and 8E+Aad,Api) that show poor solubility at acidic pH but are fully solubilised at near-neutral and alkaline pH. High ionic strength has minimal effect on their solubility, with a moderate decrease in solubility at near-neutral pH (7.4). The peptide library with long-chain acidic amino acids (8E+Aad,Api) demonstrates relatively higher solubility than the library containing their shorter-chain homologs (8E+Asp,Glu (10E)). The uncharged peptide library (8E) remains moderately insoluble across all tested conditions, and its solubility slightly decreases with increasing pH and ionic strength. In contrast, peptide libraries containing two positively charged amino acids

(8E+Lys,Arg and 8E+Dap,Dab) remain fully soluble across the tested range of pH and ionic strengths. The same extremely high solubility invariant on pH and ionic strength is also observed for all mixed-charge peptide libraries (8E+Lys,Arg,Asp,Glu; 8E+Dap,Dab,Asp,Glu; 8E+Lys,Arg,Aad,Api; 8E+Dap,Dap,Aad,Api).

Following the solubility analysis, the aggregation propensity of peptide libraries was analysed by analytical size-exclusion chromatography, according to the methodology previously described in 4.1.1. The resulting chromatograms are depicted in Fig. 4.15 and Fig. 4.16.

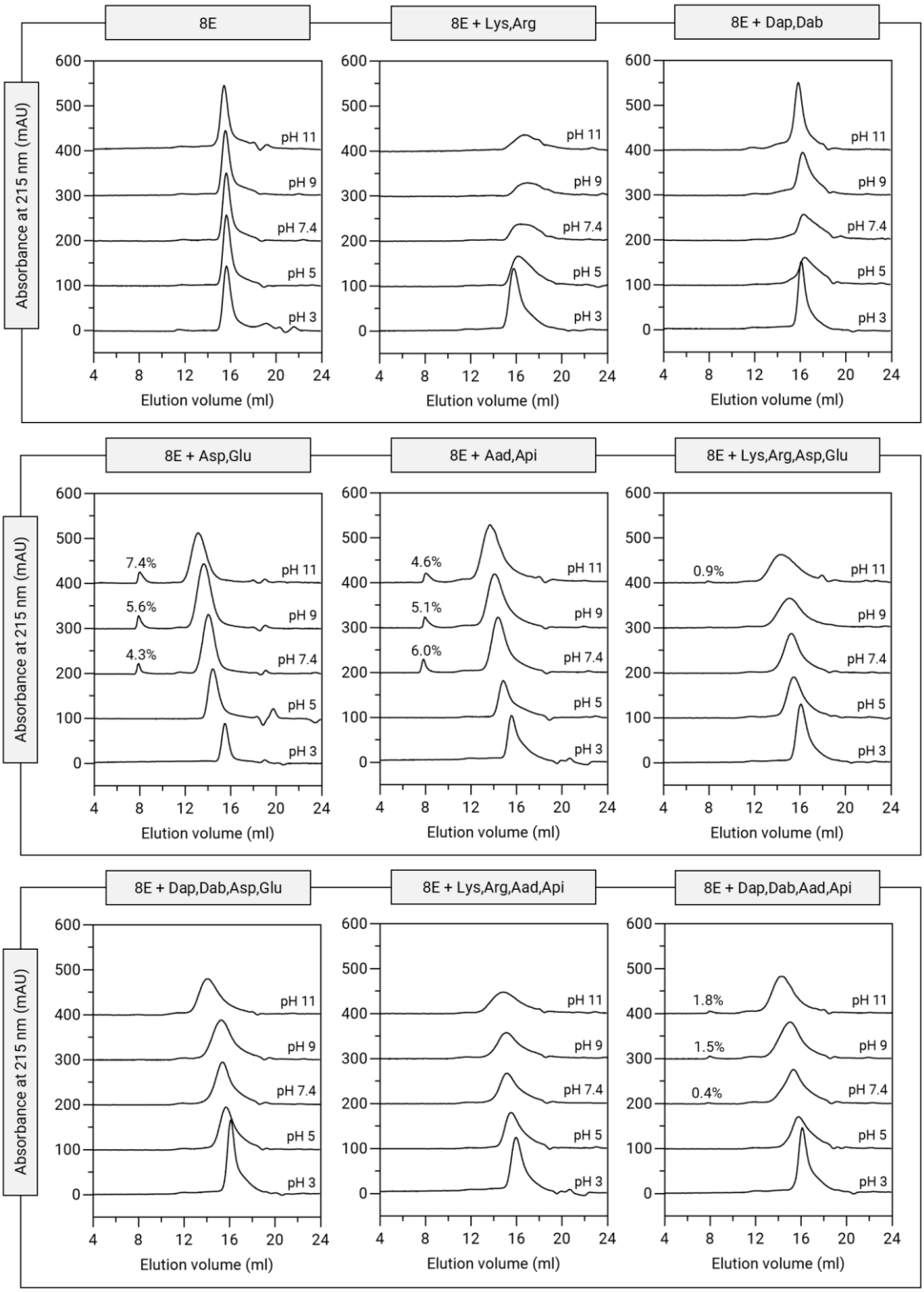


Fig. 4.15. Aggregation propensity of 25-mer combinatorial peptide libraries. Aggregation propensity of the libraries was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at low ionic strength by analytical size-exclusion chromatography. Chromatograms report absorbance at 215 nm as a function of elution volume. The data for 8E+Asp,Glu (10E) library were taken from the previous study (Makarov *et al.* 2023).

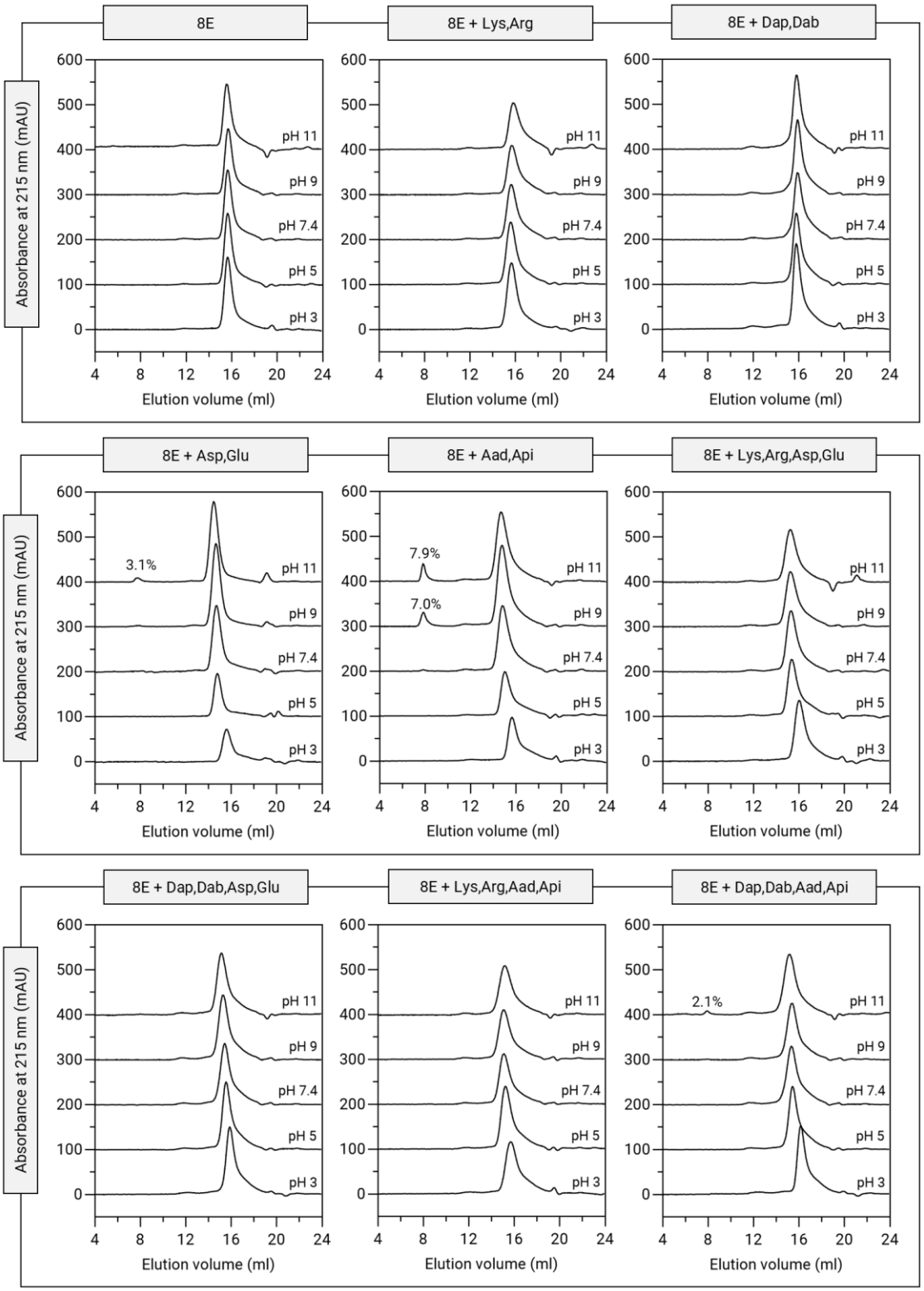


Fig. 4.16. Aggregation propensity of 25-mer combinatorial peptide libraries. Aggregation propensity of the libraries was measured for 0.5 mg/ml nominal peptide library solutions in 20 mM ABP buffers (pH 3–11) at high ionic strength by analytical size-exclusion chromatography. Chromatograms report absorbance at 215 nm as a function of elution volume. The data for 8E+Asp,Glu (10E) library were taken from the previous study (Makarov *et al.* 2023).

Similarly to the chromatograms obtained in the previous study, chromatograms for all peptide libraries are characterised by one major elution peak between 13- and 17-ml elution volumes, corresponding to the monomer fraction of peptide libraries. The intensities of monomer peaks correlate with the solubilities of peptide libraries measured by UV-Vis absorption spectroscopy. The monomer peaks of the uncharged peptide library (8E) are well-defined and have the same elution volume across the entire range of pH and ionic strengths. At low ionic strength, the monomer peaks of positively charged peptide libraries (8E+Lys,Arg and 8E+Dap,Dab) slightly shift toward higher elution volumes with increasing pH. In contrast, the monomer peaks of the remaining peptide libraries shift in the opposite direction and show a broader distribution with increasing pH. High ionic strength decreases the elution volume shifts and yields narrower and more well-defined monomer peaks. The substantial (>1 %) formation of higher-order soluble aggregates (~8-ml elution volume) is observed only on the chromatograms of two negatively charged peptide libraries (8E+Asp,Glu (10E) and 8E+Aad,Api) and 8E+Dap,Dab,Aad,Api library obtained at near-neutral and alkaline pH. Their aggregation propensity is significantly decreased at high ionic strength. The aggregate peaks are absent on the chromatograms of the remaining peptide libraries.

4.2.2. Structural analysis of combinatorial peptide libraries

After biophysical analysis, the secondary structure content in peptide libraries was estimated by far-UV CD spectroscopy as previously described in 4.1.2. First, the effect of pH on the secondary structure was analysed by collecting the far-UV CD spectra in the 190–260 nm wavelength range at different pH. The resulting far-UV CD spectra and calculated CD222/CD200 ratios are presented in Fig. 4.17. Given the high degree of disorder in peptide libraries, the secondary structure of peptide libraries was additionally induced by TFE. The far-UV CD spectra and calculated CD222/CD200 ratios obtained at different TFE concentrations are illustrated in Fig. 4.18.

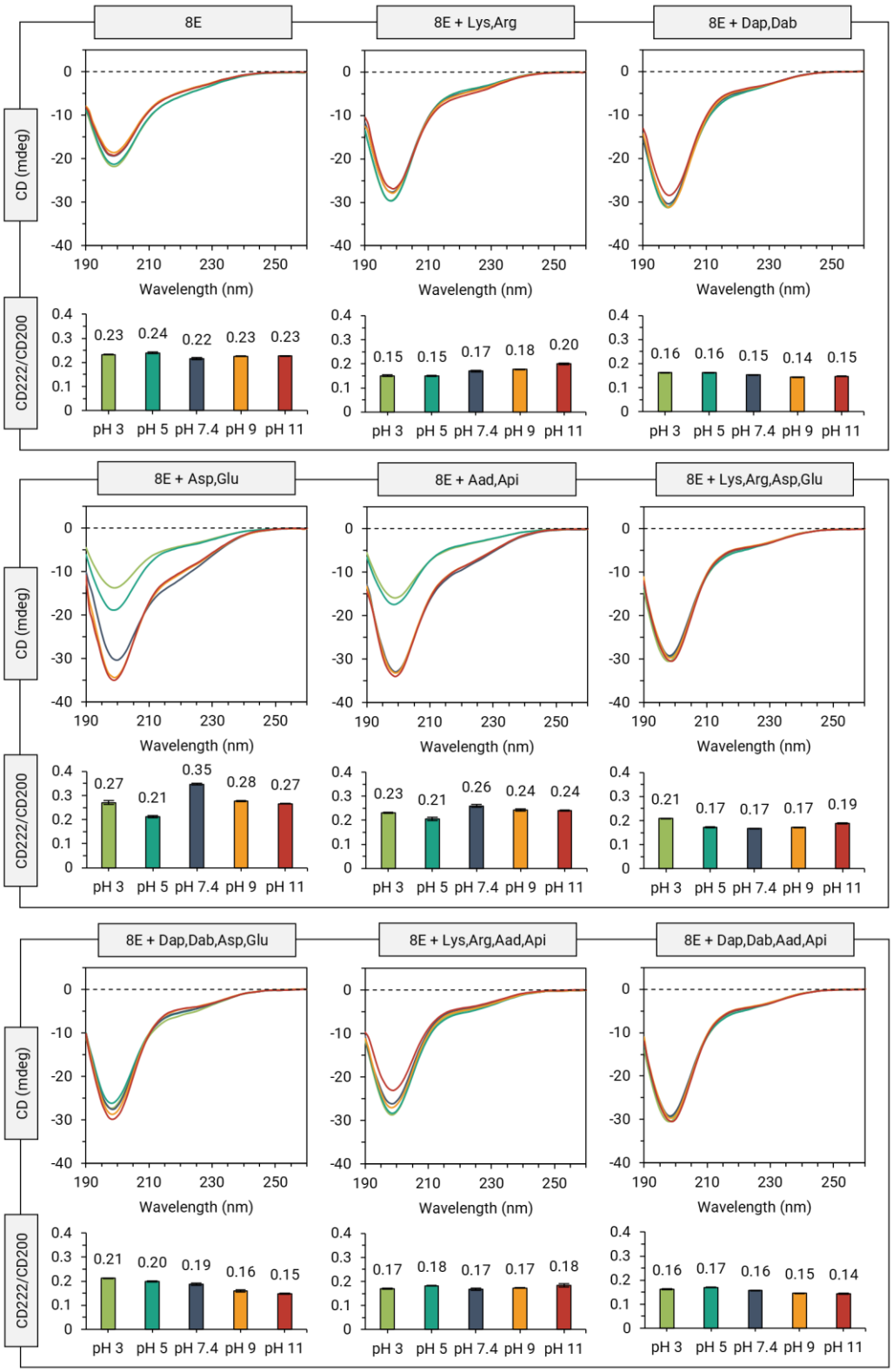


Fig. 4.17. Effect of pH on the secondary structure of 25-mer combinatorial peptide libraries. Far-UV CD spectra were measured for 0.2 mg/ml nominal peptide libraries solutions in 10 mM ABP buffers (pH 3–11). The graphs show the ratios of CD signals at 222 to 200 nm. Errors bars correspond to standard deviations derived from two analyses. The data for 8E+Asp,Glu (10E) library were taken from the previous study (Makarov *et al.* 2023).

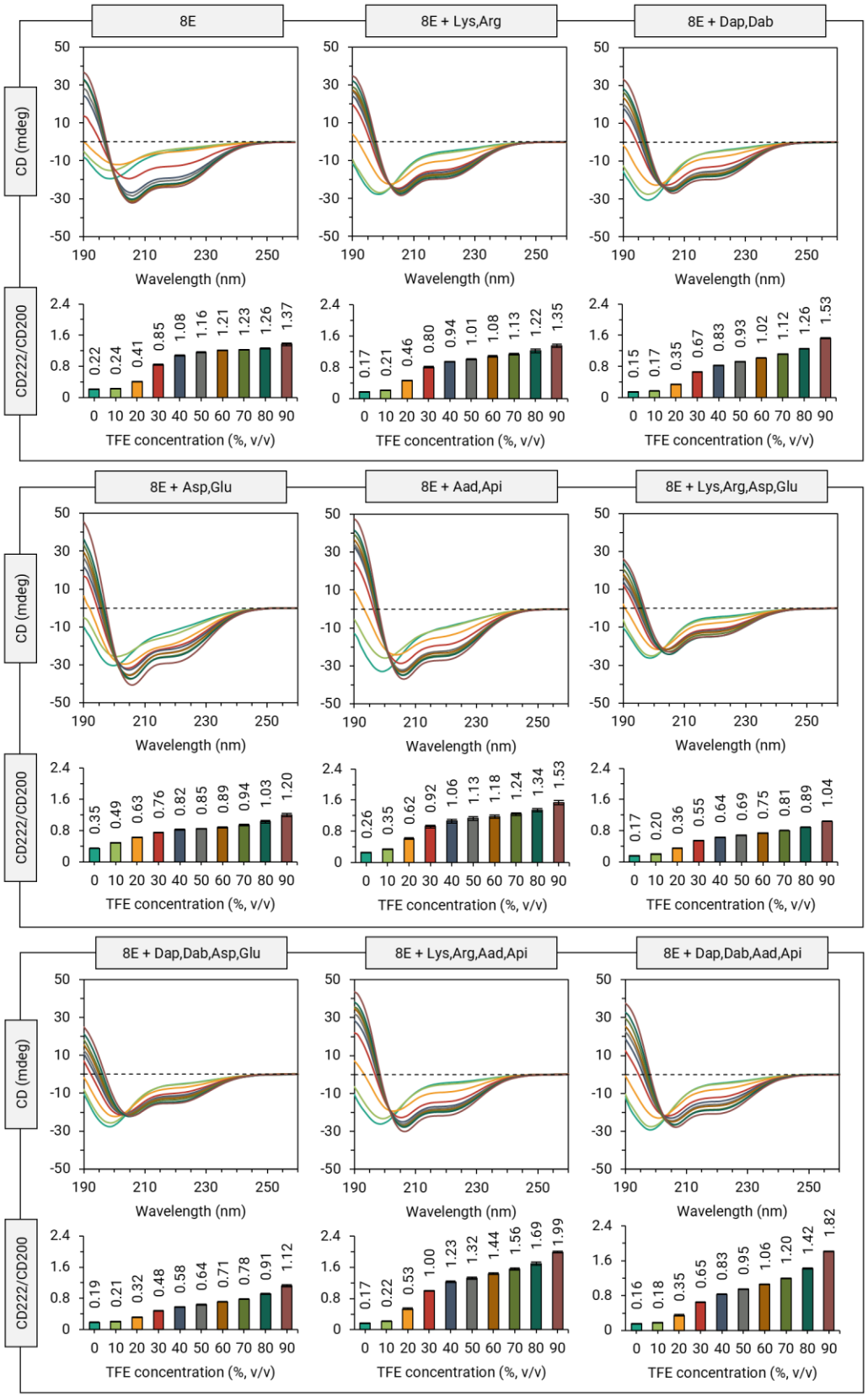


Fig. 4.18. Effect of TFE on the secondary structure of 25-mer combinatorial peptide libraries. Far-UV CD spectra were measured for 0.2 mg/ml nominal peptide libraries solutions in 10 mM ABP buffers (pH 7.4) containing 0–90% (v/v) TFE. The graphs show the ratios of CD signals at 222 to 200 nm. Errors bars correspond to standard deviations derived from two analyses. The data for 8E+Asp,Glu (10E) library were taken from the previous study (Makarov *et al.* 2023).

In general, the intensities of CD signals are proportional to the solubilities of peptide libraries obtained by UV-Vis absorption spectroscopy. Similarly to far-UV CD spectra obtained in the previous study, all far-UV CD spectra have a pronounced minimum at 198 nm, indicating the predominance of unstructured sequences in analysed peptide libraries. The overall comparison of CD222/CD200 ratios reveals that negatively charged peptide libraries (8E+Asp,Glu (10E) and 8E+Aad,Api) are more structured than the positively charged and mixed-charge peptide libraries across the entire pH range (3–11). The peptide library with short-chain acidic amino acids (8E+Asp,Glu) demonstrates the highest secondary structure content across all libraries, whereas the secondary structure content in the peptide library with long-chain acidic amino acids (8E+Aad,Api) is relatively lower and comparable to that of the uncharged peptide library (8E). Both positively charged libraries (8E+Lys,Arg and 8E+Dap,Dab) are significantly less structured than the uncharged peptide library (8E), with the peptide library containing basic canonical amino acids (8E+Lys,Arg) having higher secondary structure content than the peptide library with respective non-canonical prebiotic alternatives (8E+Dap,Dab) at alkaline pH. Mixed-charge peptide libraries with acidic canonical amino acids (8E+Lys,Arg,Asp,Glu and 8E+Dap,Dab,Asp,Glu) are moderately more structured than mixed-charged peptide libraries containing their long-chain homologs (8E+Lys,Arg,Aad,Api and 8E+Dap,Dab,Aad,Api, respectively) at acidic pH. Overall, there is no general dependency of secondary structure content in peptide libraries on pH.

Upon TFE addition, all peptide libraries adopt predominantly α-helical conformation, similarly to the peptide libraries previously described in 4.1.2. The overall increase in the intensity of CD signals for 8E peptide library is additionally caused by its dissolution in TFE. 8E+Asp,Glu (10E) peptide library, which demonstrated the highest secondary structure content in aqueous solutions, shows an inferior secondary structure-forming propensity upon TFE titration. An even lower secondary structure-forming propensity is observed only for two mixed-charged libraries supplemented with the same negatively charged canonical amino acids (8E+Lys,Arg,Asp,Glu and 8E+Dap,Dab,Asp,Glu). Upon TFE addition, all peptide libraries that contain long-chain negatively charged amino acids (8E+Aad,Api; 8E+Lys,Arg,Aad,Api; and 8E+Dap,Dab,Aad,Api) are characterized by higher secondary-structure forming propensity than the respective peptide libraries with their short-chain homologs (8E+Asp,Glu; 8E+Lys,Arg,Asp,Glu; and 8E+Dap,Dab,Asp,Glu, respectively). A similar but less pronounced trend is observed when comparing the secondary structure-forming propensities of peptide libraries supplemented with long-chain positively charged amino acids (8E+Lys,Arg;

8E+Lys,Arg,Asp,Glu; and 8E+Lys,Arg,Aad,Api) and their short-chain prebiotic alternatives (8E+Dap,Dab; 8E+Dap,Dab,Asp,Glu; and 8E+Dap,Dab,Aad,Api, respectively).

4.2.3. Compositional, biophysical, and structural analysis of monomer, aggregate, and total fractions from 10E peptide library

To understand the nature of higher-order soluble aggregates from 8E+Asp,Glu (10E) peptide library, we performed a more detailed characterisation of its monomer, aggregate, and total fractions. Prior to analyses, the individual fractions of 8E+Asp,Glu (10E) peptide library were isolated by preparative size-exclusion chromatography and concentrated to the same concentration (0.2 mg/ml) in 10 mM ABP buffer (pH 7.4). The total fraction was prepared by diluting 8E+Asp,Glu (10E) peptide library to 0.2 mg/ml in 10 mM ABP buffer (pH 7.4). The prepared fractions were subsequently characterised by HPLC amino acid analysis and far-UV CD spectroscopy. The resulting amino acid compositions and far-UV CD spectra of individual fractions are presented in Fig. 4.19. The amino acid enrichment demonstrates the differences in the amino acid content calculated for monomer/aggregate fraction and total fraction of 8E+Asp,Glu (10E) peptide library.

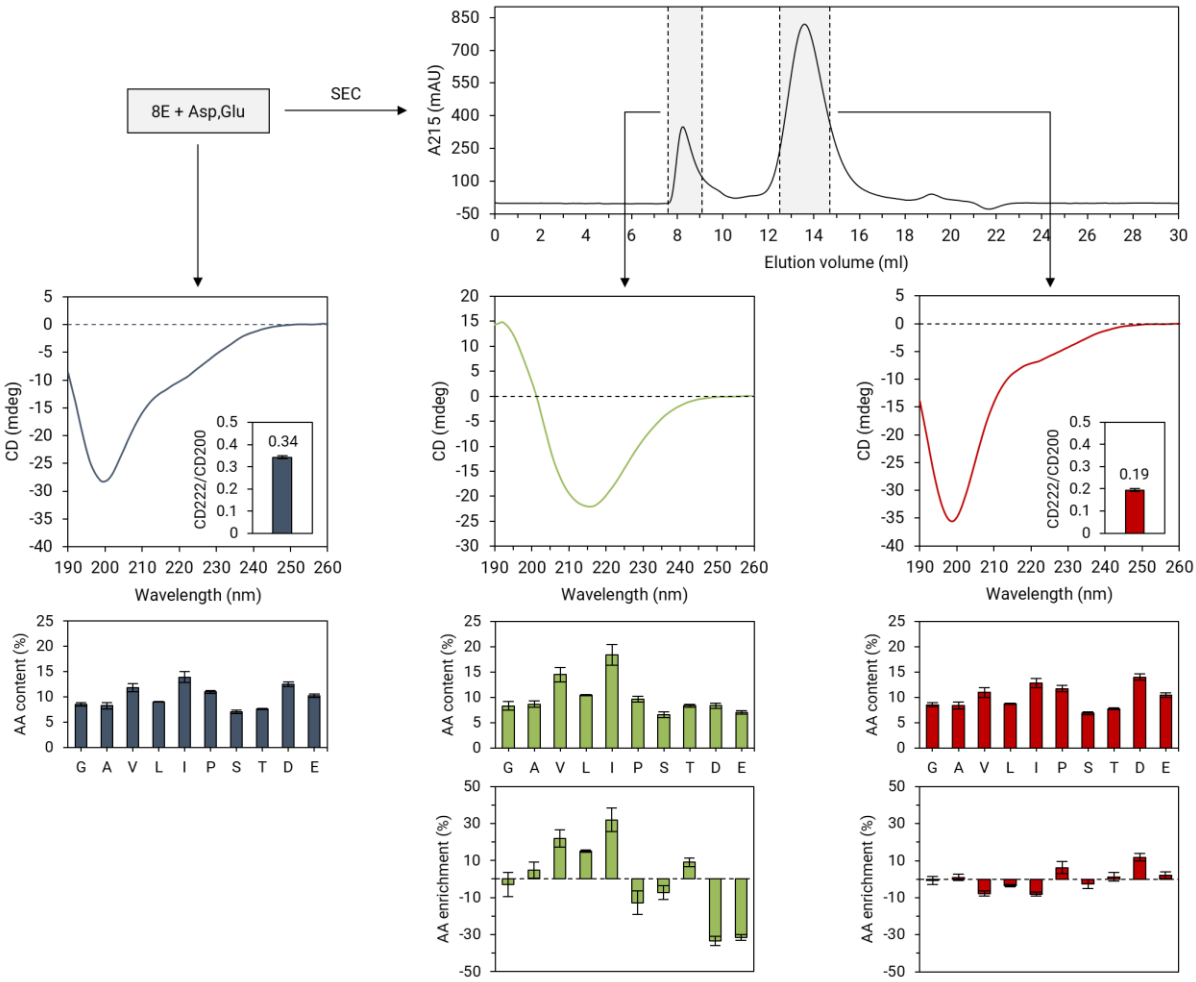


Fig. 4.19. The amino acid composition and secondary structure of monomer (red), aggregate (green) and total (blue) fractions of 8E+Asp,Glu (10E) peptide library. The peptide library was dissolved in 10 mM ABP buffer (pH 7.4) and resolved by preparative size-exclusion chromatography (SEC). The monomer fraction at 13.5 ml and aggregate fraction at 8.5 ml were isolated and analysed by HPLC amino acid analysis and far-UV CD spectroscopy. Errors bars for the amino acid (AA) content and enrichment correspond to standard deviations derived from three analyses. Error bars for CD222/CD200 ratios correspond to standard deviations derived from two analyses.

Far-UV CD spectroscopy reveals that the aggregate fraction of 8E+Asp,Glu (10E) peptide library is characterised by distinct β-sheet conformation, as indicated by the pronounced minimum at 215 nm and the maximum at 192 nm (Kelly *et al.* 2005). The HPLC amino acid analysis reveals the statistically significant enrichment of the aggregate fraction in branched aliphatic amino acids (Val, Leu, Ile) and depletion in Pro and negatively charged amino acids (Asp, Glu). In contrast, the monomer and total fractions have a similar amino acid composition and are predominantly unstructured, as evidenced by the pronounced minimum at 198 nm (Kelly *et al.* 2005). The total fraction has a higher secondary structure content than the monomer fraction, as evidenced by the comparison of CD222/CD200 ratios.

Following the compositional and structural analyses, the molecular weight distribution in individual fractions of 8E+Asp,Glu (10E) peptide library was estimated by mass photometry. Mass photometry is a rapid and straightforward technique for the determination of the molecular weight of biomolecules in solution. During the analysis, the biomolecule adsorbed on the surface of a microscope coverslip is exposed to a beam of light, which leads to light scattering by the biomolecule and light reflection at the coverslip-solution interface (Young *et al.* 2018). Mass photometry measures the interference between the light scattered by the biomolecule and the light reflected at the coverslip-solution interface. The resulting signal, called interferometric contrast, is directly proportional to the molecular mass of the biomolecule. During the analysis, the monomer, aggregate, and total fractions of 8E+Asp,Glu (10E) peptide library were diluted to 2.5 μg/ml in 10 mM Tris (pH 7.4), 50 mM NaCl. The mass calibration was performed with the monomer (66 kDa), dimer (132 kDa) and trimer (198 kDa) fractions present in 10 nM BSA solution in 10 mM Tris (pH 7.4), 50 mM NaCl. The resulting mass histograms are presented in Fig. 4.20.

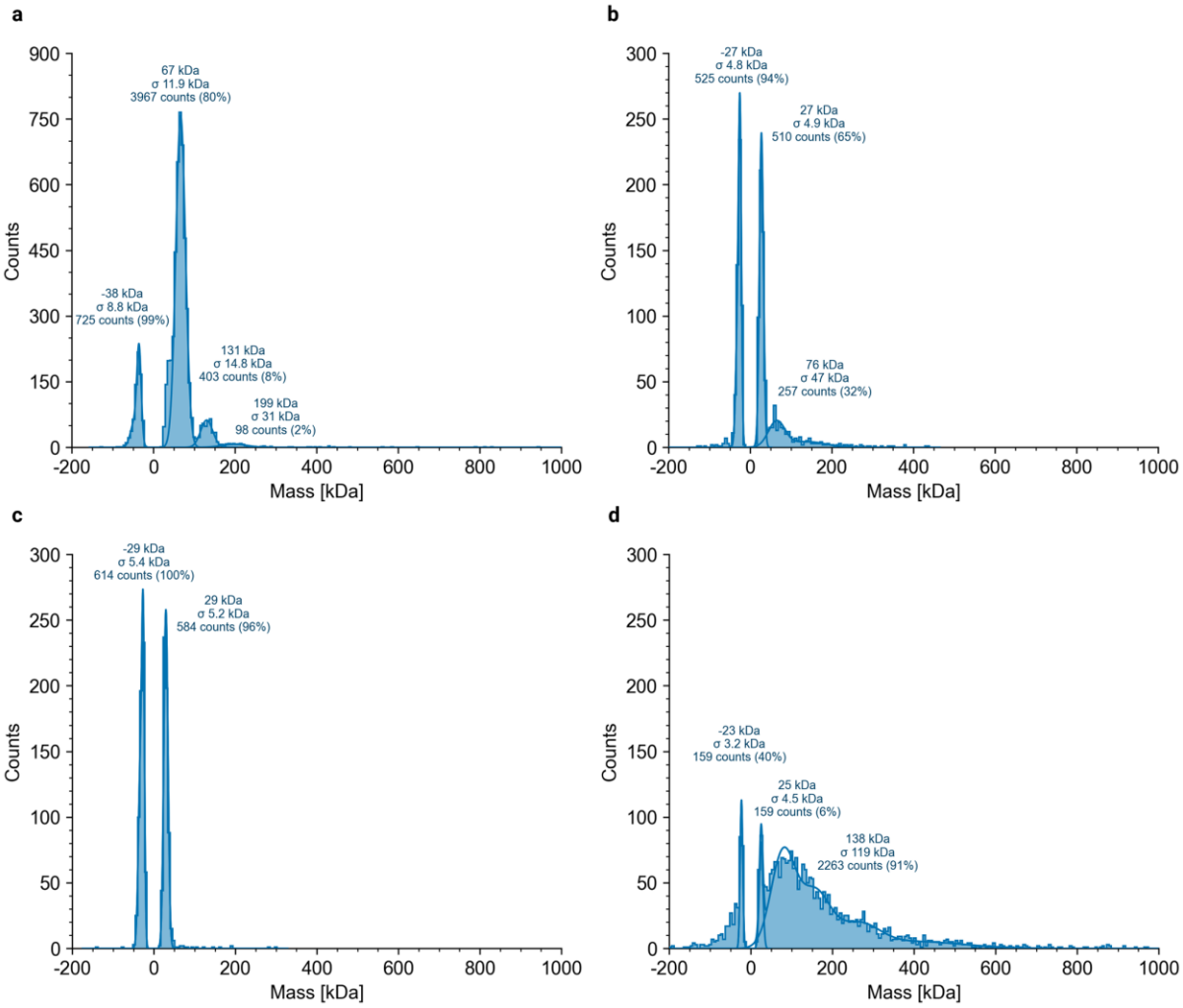


Fig. 4.20. The molecular weight distributions in total (b), monomer (c), and aggregate (d) fractions of 8E+Asp,Glu (10E) peptide library. The mass histograms were obtained for 2.5 μg/ml peptide library fractions in 10 mM Tris (pH 7.4), 50 mM NaCl using mass photometry. Mass calibration was performed using 10 nM BSA standard (a).

All mass histograms are characterised by the presence of two symmetric peaks in the negative and positive molecular weight ranges resulting from the detection limit of mass photometry (~30 kDa) (Young *et al.* 2018). The mass histograms of aggregate and total fractions are characterised by a broad non-Gaussian distribution of molecular weights, indicating the high heterogeneity of soluble aggregates formed in 8E+Asp,Glu (10E) peptide library. The mean molecular weight in the aggregation fraction is approximately 138 kDa, whereas the mean molecular weight in the total fraction is shifted towards lower values due to the lower content of soluble aggregates. In contrast, no measurable peaks are present on the mass histogram for the monomer fraction of 8E+Asp,Glu (10E) peptide library.

4.3. Characterization of RNA-binding propensity of ancestral ribosomal peptides enriched in late proteinogenic amino acids

The results in this section were included in the attached article 3 – Codispoti S., Yamaguchi T., Makarov M., Giacobelli V.G., Mašek M., Kolář M.H., Sanchez Rocha A.C., Fujishima K., Zanchetta G., Hlouchová K. The interplay between peptides and RNA is critical for protoribosome compartmentalization and stability. Nucleic Acids Res. ***28****, gkae823 (2024).*

For this study, we adopted two previously described models of the proto-ribosome (Hsiao *et al.* 2013; Bose *et al.* 2022). The proto-ribosome model inferred by Bose *et al.* (2022) consists of 136 nucleotides long ancestral PTC rRNA construct (WT sPTC rRNA) derived from *Thermus thermophilus* 23S rRNA. The proto-ribosome model inferred by Hsiao *et al.* (2013) is composed of (i) 617 nucleotides long ancestral PTC rRNA construct (WT bPTC rRNA), derived from *Thermus thermophilus* 23S rRNA; and (ii) presumably ancestral fragments of five ribosomal proteins (uL2, uL3, uL4, uL15, and uL22 from *Thermus thermophilus*), denoted as pL2, pL3, pL4, pL15 and pL22 rPeptides, respectively. The three-dimensional structures of rPeptides (as part of ribosomal proteins from *Thermus thermophilus*) are presented in Fig. 4.21. The three-dimensional structures of ancestral PTC RNA constructs (with associated rPeptides) are illustrated in Fig. 4.22.

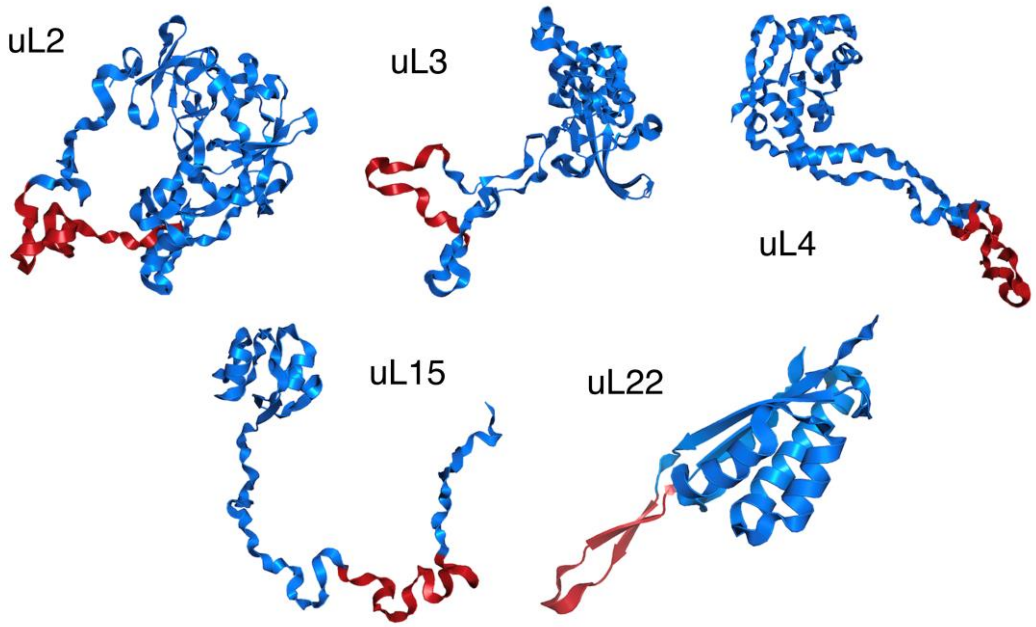


Fig. 4.21. The three-dimensional structures of ribosomal proteins uL2, uL3, uL4, uL15, and uL22 from *Thermus thermophilus*. The structures were extracted from *Thermus thermophilus* 70S ribosome (PDB ID: 4V51) and visualised using PyMOL-2.5.7. The ancestral fragments of ribosomal proteins (rPeptides) derived by Hsiao *et al.* (2013) are highlighted in red.

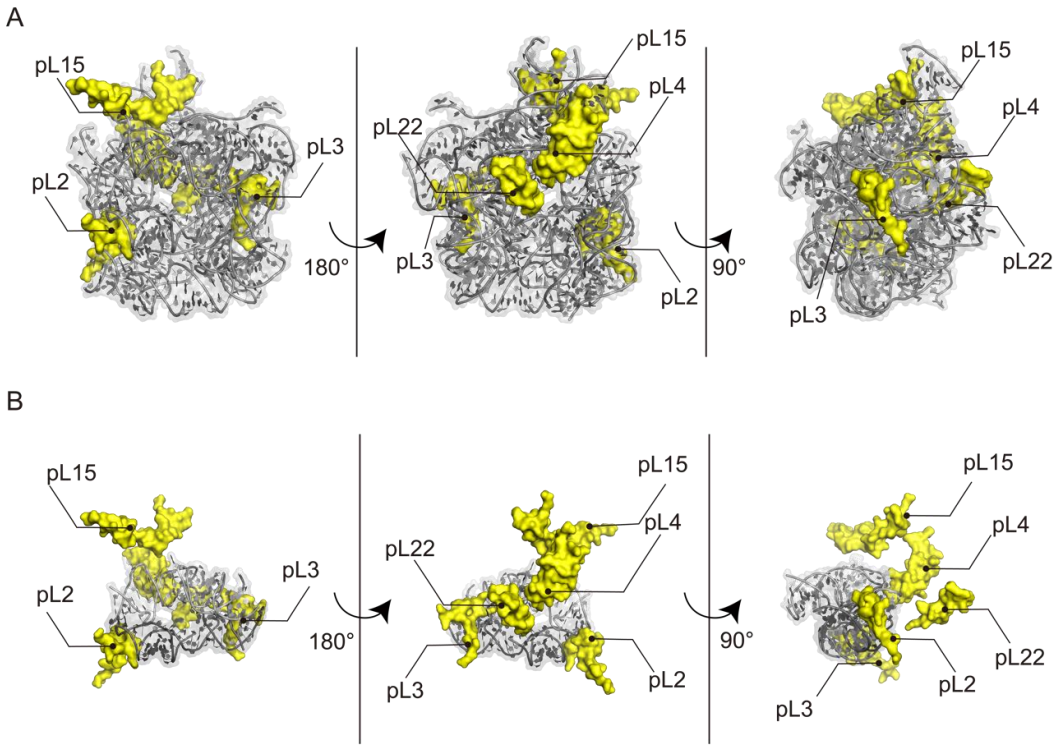


Fig. 4.22. The three-dimensional structures of WT bPTC (A) and WT sPTC (B) rRNAs (in grey) with associated rPeptides (in yellow). The structures were extracted from *Thermus thermophilus* 70S ribosome (PDB ID: 4V51) and visualised using PyMOL-2.5.7. The sequences of WT bPTC rRNA and rPeptides were taken from Hsiao *et al.* (2013), and the sequence of WT sPTC rRNA was taken from Bose *et al.* (2022).

The aim of this study was to characterise interactions between rPeptides and two ancestral PTC rRNA constructs in order to elucidate the plausible involvement of ancestral peptides enriched in late proteinogenic amino acids in the early ribosome evolution. The effect of rPeptides on the coacervation and stability of ancestral PTC rRNA was subsequently characterised by two principal authors of this study, MSc. Simone Codispoti and Tomoko Yamaguchi, PhD, respectively, and their results will be briefly discussed in this dissertation.

The original sequences of three rPeptides from Hsiao *et al.* (2013) were preliminary modified for this study. In particular, the sequences of pL4 and pL15 rPeptides were truncated due to the failure of Fmoc-based solid-phase synthesis of the full-length peptides. The truncated variants of pL4 and pL15 rPeptides were designed by Alma Carolina Sanchez Rocha, PhD, based on the best structural preservation of their interactions with WT bPTC rRNA. The sequence of pL2 rPeptide was extended by three amino acid residues from both N- and C-termini due to the reported absence of its interaction with WT bPTC rRNA (Hsiao 2013). The rPeptides were commercially synthesised by Fmoc-based solid-phase peptide synthesis, and their correct molecular weights were subsequently confirmed by mass spectrometry analysis (Fig. 4.23).

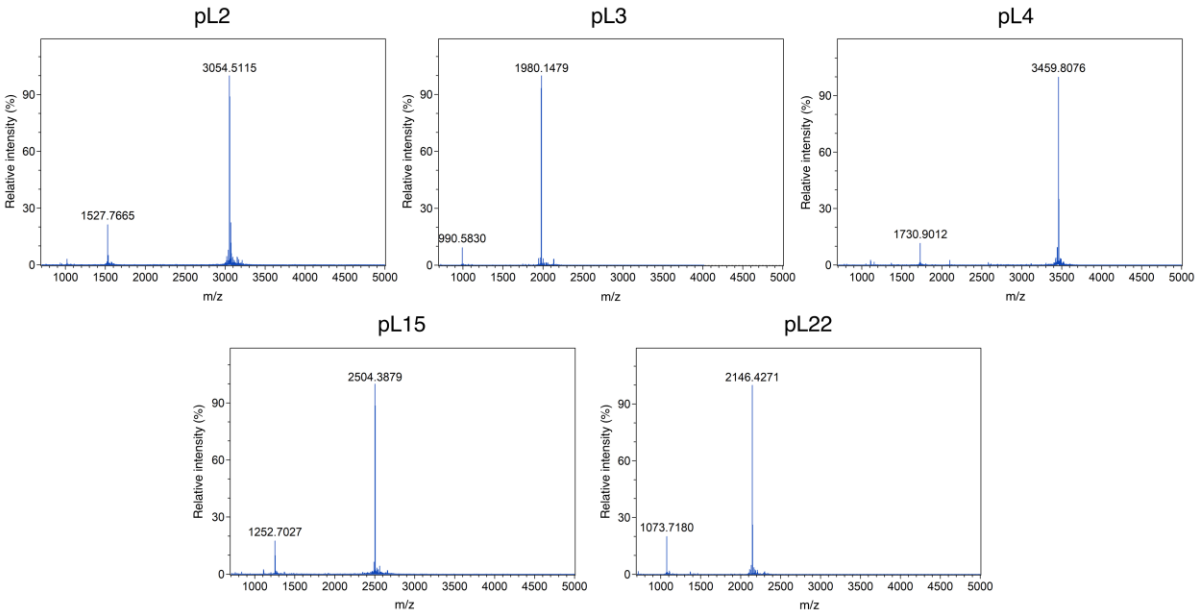


Fig. 4.23. MALDI-TOF spectra of rPeptides. The spectra were obtained for rPeptides dissolved in 50% (v/v) acetonitrile using DHB as a matrix. The main m/z peaks correspond to the theoretical molecular weights of rPeptides.

Both ancestral PTC rRNAs were synthesized *in vitro* by T7 RNA polymerase-mediated transcription of dsDNA templates. The dsDNA templates were prepared by PCR amplification of a long synthetic ssDNA oligonucleotide (in the case of WT bPTC rRNA) or annealing of two short synthetic ssDNA oligonucleotides followed by DNA synthesis by Klenow fragment of DNA polymerase I (in the case of WT sPTC rRNA). The correct sizes of *in vitro* transcribed ancestral PTC rRNA constructs were confirmed by denaturing urea-PAGE (Fig. 4.24).

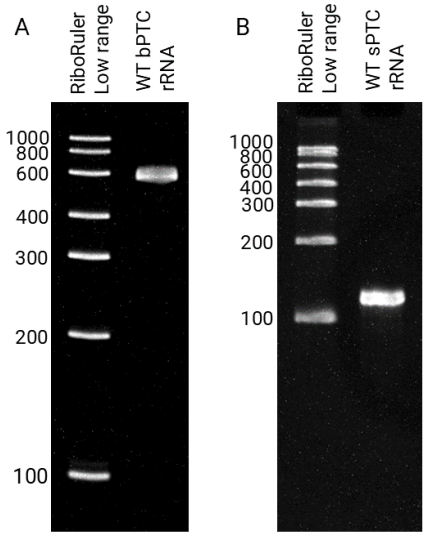


Fig. 4.24. Denaturing polyacrylamide gels of the ancestral PTC rRNA constructs. (A) 5% TBE-urea polyacrylamide gel of WT bPTC rRNA; (B) 8% TBE-urea polyacrylamide gel of WT sPTC rRNA. RiboRuler Low Range RNA Ladder (Thermo Fisher Scientific) was used as a reference.

Following the synthesis, ancestral PTC rRNA constructs were fluorescently labelled by periodate-mediated oxidation of 3'-terminal nucleotide followed by the spontaneous reaction of oxidised 3'-terminal nucleotide with the thiosemicarbazide derivative of fluoresceine, FTSC (Fig. 4.25).

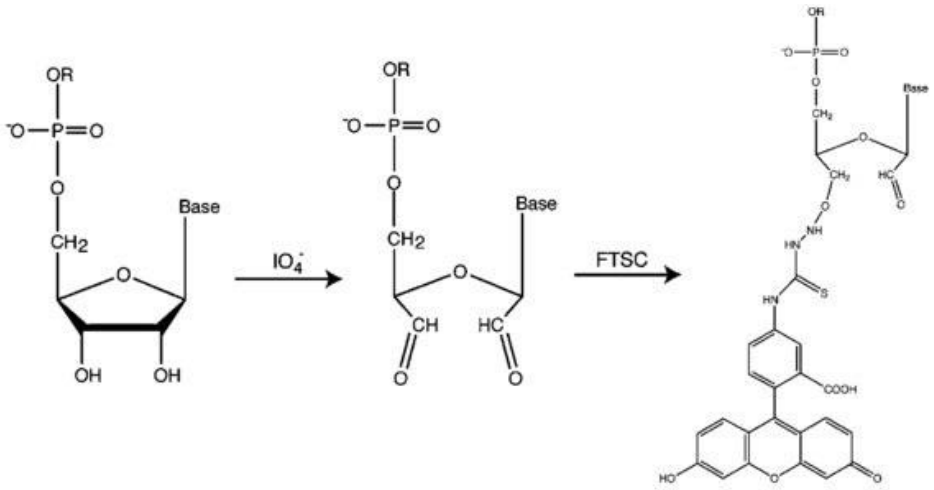


Fig. 4.25. The reaction scheme for 3'-end labelling of RNA. Sodium periodate selectively oxidises and cleaves the vicinal diol moiety of 3'-terminal nucleotide, producing two carbonyl groups. The resulting carbonyl groups spontaneously react with FTSC, leading to the covalent attachment of fluorescein to the 3'-end of RNA through thiosemicarbazone linkage. Adopted from Pagano *et al.* (2011).

The interaction of fluorescein-labelled ancestral PTC rRNA constructs with rPeptides was characterised by microscale thermophoresis in the absence and presence of metal ions. Microscale thermophoresis is a technique used for the quantitative estimation of biomolecular interactions in solution (Seidel *et al.* 2013). A fluorescently labelled biomolecule is mixed with a serially diluted ligand, placed into a glass capillary and irradiated with an infrared laser with an emission wavelength of 1480 nm. The infrared laser creates a microscopic temperature gradient (0–6 °C), which causes the movement of the fluorescently labelled biomolecule along the temperature gradient (thermophoresis). The thermophoresis is detected as the decrease (or sometimes increase) in the fluorescence intensity and depends on the size, charge, and solvation shell of the biomolecule affected by ligand binding. The resulting MST signal is expressed as the normalised fluorescence calculated from the decrease/increase in the initial fluorescence shortly after the infrared laser irradiation (T-jump analysis) or later (thermophoresis analysis) and plotted against ligand concentration. The resulting binding curves are used to estimate the dissociation constant of the biomolecule-ligand complex. During the analysis, the fluorescein-labelled ancestral PTC rRNAs were kept at a constant final concentration (50 nM), whereas the final concentration of 2-fold serially diluted rPeptides varied from 30.52 nM to 500 μM. To estimate the effect of metal ions, the MST measurements were performed in (i) 20 mM Tris

(pH 7.5) or (ii) 20 mM Tris (pH 7.5), 1 mM KCl, 1 mM $MgCl_2$, and 1 mM $CaCl_2$. The resulting binding curves and the initial fluorescence traces are presented in Fig. 4.26. The dissociation constants of ancestral PTC rRNA-rPeptide complexes were obtained by fitting the binding curves into the single site-binding model (WT sPTC rRNA and WT bPTC rRNA) or Hill cooperative binding model (WT bPTC rRNA). The fitting of the binding curves was performed by MSc. Simone Codispoti. The dissociation constants obtained using the single-site binding model are presented in Fig. 4.27.

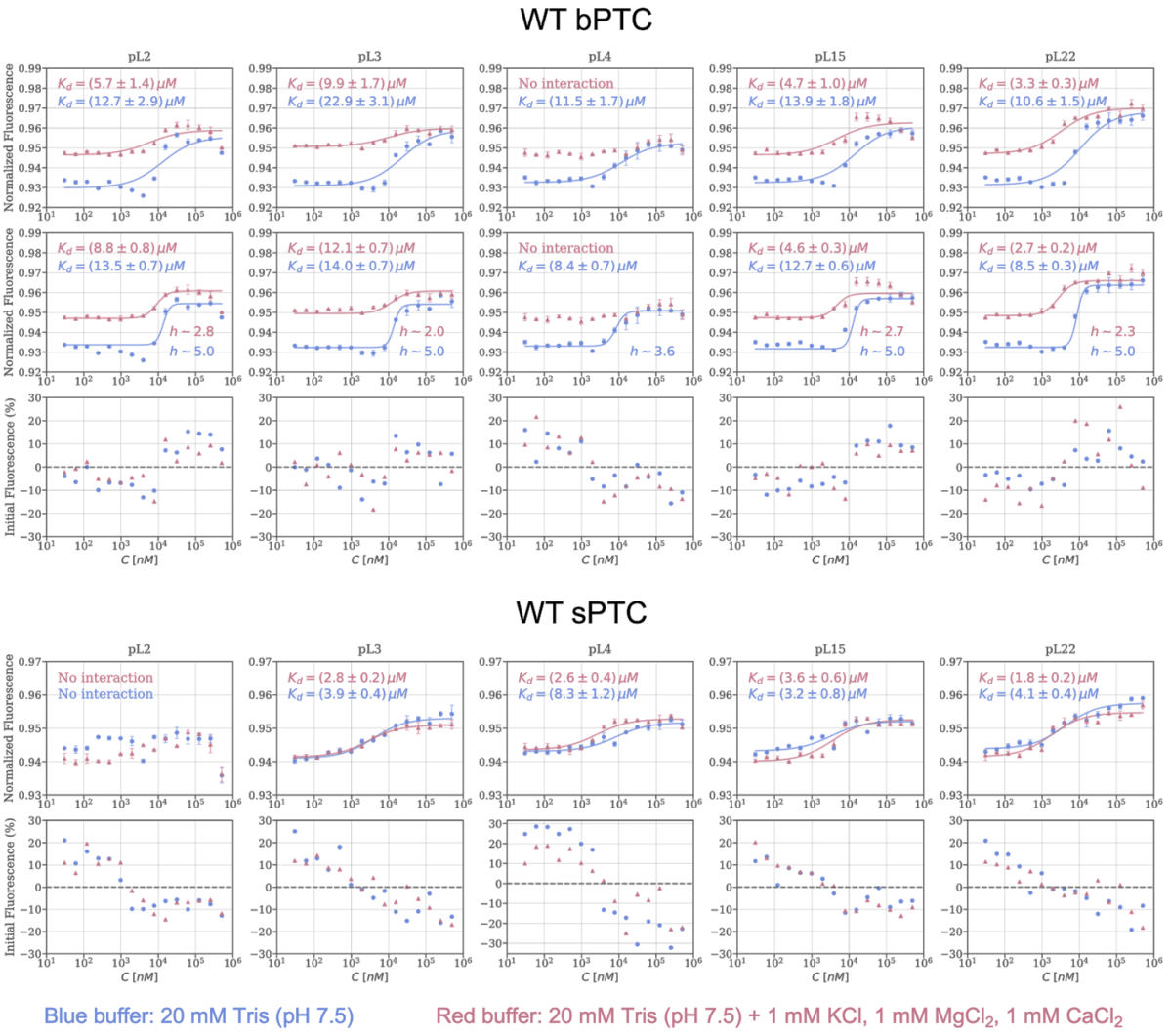


Fig. 4.26. The binding curves (upper rows) and fluorescence traces (lower rows) for WT bPTC (top panel) and WT sPTC (bottom panel) rRNAs with pL2, pL3, pL4, pL15, and pL22 rPeptides. The MST measurements were performed with (red) and without (blue) metal ions (1 mM KCl, 1 mM $MgCl_2$, and 1 mM $CaCl_2$) supplemented in 20 mM Tris (pH 7.5). The dissociation constants were derived by fitting the binding curves to the single site-binding model (WT sPTC rRNA and WT bPTC rRNA) or Hill cooperative binding model (WT bPTC rRNA).

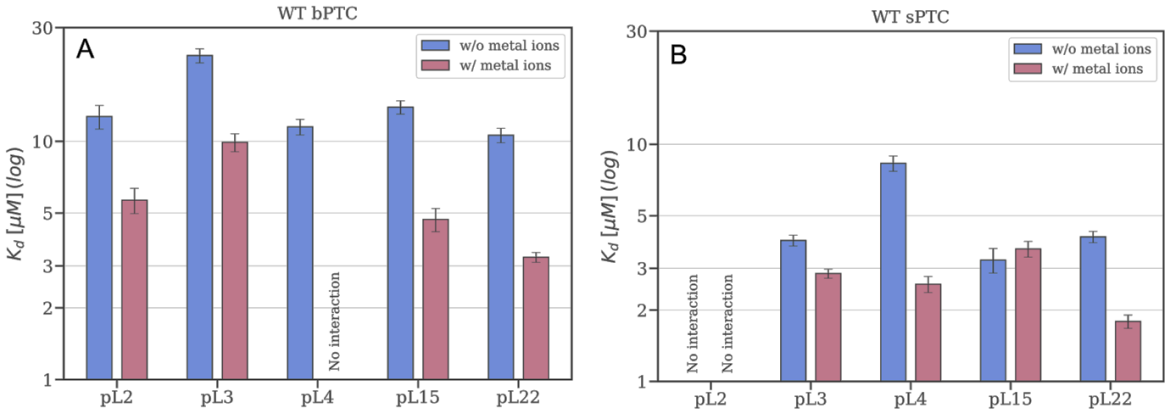


Fig. 4.27. The estimated dissociation constants of ancestral PTC rRNA-rPeptide complexes in the absence and presence of metal ions. The dissociation constants were obtained by fitting the binding curves to the single-site binding model. "No interaction" indicates the cases where the estimated dissociation constant is outside the 30.52 nM to 500 μM concentration range or the dissociation curve could not be fitted. Error bars represent the standard deviations from two replicates.

The MST measurements demonstrate that both ancestral PTC rRNAs interact with most rPeptides with micromolar dissociation constants. Dissociation constants could not be derived only for pL4 rPeptide with WT bPTC rRNA in the presence of metal ions and for pL2 rPeptide with WT sPTC rRNA in both conditions. The presence of metal ions does not significantly affect the shapes of binding curves of rPeptides to WT sPTC rRNA but significantly alters the interaction of rPeptides with WT bPTC rRNA. Based on the comparison of the apparent Hill coefficients, the binding stoichiometries of WT bPTC rRNA and rPeptides (except for pL4) significantly decrease in the presence of metal ions. For both WT bPTC rRNA and WT sPTC rRNA, the dissociation constants of the complexes with rPeptides moderately decrease upon the addition of metal ions.

5. Discussion

5.1. Foldability defined the selection of early canonical amino acid alphabet

The vast majority of scientifically rationalised hypotheses on the origin of life assumes that cellular life developed from environmentally available organic compounds (Dose 1986; Lal 2008; Cleaves 2012). The extensive research on the abiotic formation of building blocks of life implies the prebiotic availability of more than 90 different amino acids of terrestrial or extraterrestrial origin (section 1.1). However, only 10 of them (known as early canonical amino acids) were selected to build proteins that constitute all modern terrestrial life. The early set is represented by hydrophobic (Gly, Ala, Val, Leu, Ile, Pro), hydroxy (Ser, Thr) and negatively charged (Asp, Glu) amino acids, all of which have their respective abiotically synthesisable isomers and homologs with similar physicochemical properties (Fig. 1.1). The most striking feature of the genetic code is the absence of four-carbon aliphatic α-amino acids (Aba and 2-aminoisobutyric acid) that are consistently identified in both abiotic synthesis simulations and extraterrestrial bodies in much higher concentrations than their higher (C5–C6) homologs (Ring *et al.* 1972; Lawless and Peterson 1975; Weber and Miller 1981; Potiszil *et al.* 2023). An increase in the number of carbon atoms leads to the exponential increase in possible structural isomers for aliphatic α-amino acids, all of which can be plausibly produced under prebiotic conditions but cannot be identified in a single analysis. However, some of them, namely unbranched Nva and Nle, are frequently found in comparable or even higher quantities than their branched isomers (Val and Leu) that were selectively recruited over them to the genetic code (Ring *et al.* 1972; Wolman *et al.* 1972; Lawless and Peterson 1975; Weber and Miller 1981; Schlesinger and Miller 1983; Potiszil *et al.* 2023). Given the scarcity or absence of experimental evidence on the abiotic formation of late canonical amino acids that constitute the other half of the genetic code, prebiotic availability can certainly be excluded as the selection factor that defined the amino acid composition of the extant proteome. Moreover, positively charged late canonical amino acids (Lys, Arg) have their respective analogous amino acids (Dap, Dab, Orn) that can be abiotically synthesised and even biosynthesised at lower metabolic costs due to lower chemical complexity (Nigam and Ressler 1966; Wolman *et al.* 1972; Weber and Miller 1981; Meierhenrich *et al.* 2004; Kobylarz *et al.* 2014). All these observations invoke the question of what selection factors shaped the standard genetic code, which was addressed in this study.

For this study, the early canonical subset (10E) of amino acids was partially substituted (10U) or supplemented (11D, 11R, 11Y, 19F) with late canonical and prebiotically available

non-canonical amino acids to elucidate the selection factors that led to their respective inclusion and exclusion from the genetic code. The biophysical (solubility, aggregation propensity) and structural (secondary structure) properties conferred by the resulting amino acid alphabets were screened on the vast sequence space occupied by 25-mer fully randomised peptide libraries. The fundamental properties (solubility, aggregation, and folding) of peptides and proteins largely depend on both the content (amino acid composition) and the specific arrangement (amino acid sequence) of individual amino acids. Given the extremely high degeneracy of tested combinatorial peptide libraries, the observed differences in their biophysical and structural properties can be solely attributed to their amino acid composition. The length of 25 residues was chosen as a compromise between the minimal length required for the formation of stable intramolecular (monomeric) peptide folds and the limitations of standard Fmoc-based solid phase synthesis (Baker *et al.* 2017). The selected length exceeds the maximum lengths of peptides obtained in most studies on the abiotic non-specific condensation of single amino acids mediated by various physico-chemical mechanisms (section 1.2). However, the longer peptide sequences could have still been synthesised by abiotic ligation of shorter peptide sequences, which was demonstrated in several studies (section 1.2), or by non-specific amino acid condensation mediated by primordial translation machinery (section 1.4), supporting the possible prebiotic relevance of the tested peptide libraries.

The observed differences in the solubility of peptide libraries can be partially explained by the physico-chemical properties of amino acids. In general, the solubility of peptides and proteins greatly depends on the content of hydrophobic residues and charged residues, whose ionisation state defines the peptide/protein net charge (Pace *et al.* 2009; Qing *et al.* 2022). Peptide libraries composed of the reduced amino acid alphabets are characterised by the general propensity to be less soluble at acidic pH due to their enrichment in negatively charged Asp (side chain $pK_a \sim 3.9$) and Glu (side chain $pK_a \sim 4.3$) residues, whose side chains remain ionised at near-neutral and alkaline pH but become protonated and uncharged at acidic pH (Pace *et al.* 2009). The decrease of solubility of peptide libraries at high ionic strength is in line with the salting-out effect observed for globular proteins at high salt concentrations (Kherb *et al.* 2012; Duong-Ly and Gabelli 2014). Another general trend observed from the solubility analysis is that prebiotically available non-canonical amino acids (Aba, Nva, Nle, Dab) drastically increase the overall solubility of peptide libraries. The replacement of branched Val, Leu, and Ile by unbranched Aba, Nva and Nle does not affect the relative Asp and Glu content in the libraries; thus, the differences in the solubilities of 10U and 10E peptide libraries are most likely caused by the differences in the hydrophobicity of aliphatic amino acids. Unbranched

Nva and Nle are more hydrophobic than their respective branched isomers, Val, Leu and Ile, due to the higher water-exposed surface area of unbranched side chains (Reynolds *et al.* 1974; Kennedy 1990; Kovacs *et al.* 2006). However, their higher hydrophobicity is plausibly largely compensated by the ultimately lower hydrophobicity of short-chain Aba, resulting in the higher solubility of 10U peptide library (Reynolds *et al.* 1974). The addition of positively charged Dab widens the distribution of isoelectric points of individual peptide variants, moderately decreases the relative Asp and Glu content and influences their ionisation due to charge-charge interactions, resulting in the extremely high solubility of 11D peptide library (Pace *et al.* 2009). Laurents *et al.* (2003) demonstrated that the replacement of negatively charged Asp and Glu residues to positively charged Lys in ribonuclease Sa decreases the side chain pK_a of the remaining Glu residues that remain ionised at lower pH values. In contrast to the side chain amino group of Dab, the positive charge of the guanidinium moiety of Arg (side chain pK_a ~ 13.8) is strongly delocalised, leading to the poor solvation of Arg residues and the decreased solubility of 11R peptide library in comparison with 11D peptide library (Mason *et al.* 2003; Fitch *et al.* 2015). Late canonical amino acids generally decrease the solubility of peptide libraries, as indicated by the lower solubilities of 11R, 11Y, and 19F peptide libraries. The lower solubility of 11Y peptide library most likely results from the moderate decrease in Asp and Glu content caused by the addition of Tyr as well as the higher hydrophobicity of aromatic amino acids compared to aliphatic amino acids (Kennedy 1990). Several conflicting rankings have been implemented to assess the relative hydrophobicity of canonical amino acids, which hinders the direct comparison of the overall hydrophobicity of the early (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr, Asp, Glu) and late (Asn, Gln, Lys, Arg, His, Phe, Tyr, Trp, Met, Cys) subsets of canonical amino acids (Cornette *et al.* 1987; Kovacs *et al.* 2006). The full amino subset (19F) confers drastically low solubility in comparison with reduced amino acid subsets. This observation aligns with previous experimental studies that assessed the relative solubility of *in vivo* or *in vitro* expressed combinatorial protein libraries composed of reduced and full canonical amino acid alphabets (Tanaka *et al.* 2010; Newton *et al.* 2019; Tretyachenko *et al.* 2020).

Increasing pH decreases the elution volumes of monomer peaks of peptide libraries observed during size-exclusion chromatography. In theory, the elution volume of a biomolecule depends strictly on its hydrodynamic radius, which, in turn, is influenced by the size (molecular weight) and shape (conformation) of biomolecules (La Verde *et al.* 2017). The shifts in elution volumes are plausibly caused by pH-dependent conformational changes of individual peptide variants, which can undergo compaction at lower pH due to protonation of Asp and Glu side

chains and minimisation of intramolecular electrostatic repulsion. However, the changes in elution volumes can also be attributed to non-specific electrostatic interactions with the column matrix. The Superdex column matrix contains a residual amount of negatively charged groups that can affect elution volumes of peptides and proteins by electrostatic repulsion or attraction (Drevin *et al.* 1990). At low ionic strength, the peptide libraries enriched in acidic canonical amino acids (Asp, Glu) become negatively charged with increasing pH, which leads to the electrostatic repulsion with the negatively charged matrix and a decrease in the elution volume. High ionic strength reduces the electrostatic repulsion between the peptide molecules and the column matrix as well as intra- and intermolecular electrostatic repulsions of peptide molecules, simultaneously decreasing the differences in elution volumes and narrowing monomer peaks.

Size-exclusion chromatography indicates the formation of higher-order soluble aggregates in peptide libraries composed solely of canonical amino acids, specifically 10E, 11R and 11Y libraries. This finding correlates with the computational analysis, according to which non-covalent interactions involving amino acid residues with delocalised π-electrons (Asn, Gln, Arg, His, Phe, Tyr, Trp) contribute the most to protein aggregation (Hou *et al.* 2018). In fact, intermolecular association of peptides into insoluble aggregates may largely contribute (along with precipitation of monomeric peptides) to the lower solubility of peptide libraries enriched in late proteinogenic amino acids, ultimately resulting in the extremely low solubility of 19F peptide library. At higher pH, electrostatic repulsion between negatively charged peptide variants in 11R and 11Y peptide libraries may reduce aggregation propensity, increasing the overall solubility and resulting in formation of soluble rather than insoluble peptide aggregates. The high aggregation propensity of 10E peptide library is rather intriguing and will be discussed in section 5.2. In contrast, no higher-order soluble aggregates are observed after the inclusion of non-proteinogenic amino acids in 10U and 11D peptide libraries.

Far-UV CD spectroscopy demonstrates the general propensity of all peptide libraries to be largely unstructured in aqueous solutions. Water was shown to be a stronger hydrogen bond donor to an amidic carbonyl group than secondary amides, resulting in the preferential formation of intermolecular amide-water hydrogen bonds over intramolecular amide-amide hydrogen bonds and thus the extended conformation of most peptide variants in the tested peptide libraries (Harrington and Schellman 1956; Eberhardt and Raines 1994). The secondary structure analysis reveals the drastically lower secondary structure content in peptide libraries containing non-canonical amino acids across the entire pH range. This observation is consistent with several mutagenesis studies that demonstrate the loss of secondary structure after the

replacement of branched aliphatic amino acids with their unbranched analogues in particular peptide sequences (Brack 1987; Škibola *et al.* 2023), albeit the opposite effect was also observed (Lyu *et al.* 1991; Padmanabhan and Baldwin 1991). The lower secondary structure-forming propensity of unbranched non-proteinogenic amino acids most likely stems from the higher conformational freedom of their side chains leading to the significant unfavourable loss of conformational entropy upon peptide/protein folding (Weber and Miller 1981; Chellgren and Creamer 2006). In contrast, the lower secondary structure-forming propensity of Dab possibly arises from salt bridge interactions between positively charged short side chain of Dab and negatively charged short side chains of Asp and Glu that would be formed more frequently in unfolded peptides stabilising their extended conformation. The stabilising effect of electrostatic interactions between ionisable side chains on the secondary structure was shown to be strictly dependent on the relative position of amino acid residues, and given the high degeneracy of peptide libraries, the stabilising electrostatic interactions are greatly outnumbered by destabilising electrostatic interactions (Scholtz *et al.* 1993; Musafia *et al.* 1995). The lower secondary structure-forming propensities of non-canonical amino acids experimentally observed in this study are additionally confirmed by secondary structure predictions performed for each library using 1,000 random peptide sequences by Alma Carolina Sanchez Rocha, PhD.

Inclusion of late proteinogenic amino acids does not have a consistent effect on the secondary structure content of peptide libraries across the tested pH range, which in average remains relatively the same to that of 10E peptide library in aqueous solutions. In contrast to Dab, the presence of positively charged Arg is not deleterious to peptide folding possibly due to the delocalisation of positive charge of the guanidinium moiety as well as the longer side chain, which ensures more favourable geometry of Asp/Arg and Glu/Arg salt bridges (Mason *et al.* 2003; Meuzelaar *et al.* 2016). The higher secondary structure-forming propensity of late proteinogenic amino acids was distinctly observed upon TFE induction. The exact mechanism of secondary structure stabilisation by TFE is unknown but has been proposed to occur via preferential solvation of peptide molecules by TFE, which leads to the exclusion of water molecules from the local environment and facilitates formation of intramolecular hydrogen bonds in peptides (Roccatano *et al.* 2002; Fioroni *et al.* 2002; Vincenzi *et al.* 2019). At the same time, TFE was shown to promote hydrophobic interactions between amino acid side chains, ultimately leading to their hydrophobic packing (Reiersen and Rees 2000). Even though TFE is not prebiotically relevant, several researchers hypothesise that TFE mimics the hydrophobic

environment formed inside cell membranes and, possibly, even in the interior of globular proteins (Vincenzi *et al.* 2019).

Overall, the findings of our study conclude that foldability, i.e. the ability to fold into a soluble well-defined conformation, was a critical factor in the selection of the canonical amino acid alphabet. Despite their prebiotic availability, unbranched (Aba, Nva, Nle) and short-chain basic (Dab) amino acids tested in this study are not present in the modern version of the genetic code because of their tendency to produce oversolubilised unstructured polypeptides. However, it remains unknown whether non-canonical amino acids (i) were never recruited to the genetic code; or (ii) were initially present in the early version of the genetic code and later excluded or replaced by late canonical amino acids. If present in the genetic code, prebiotically available non-canonical amino acids could have provided functionalities that were not available for the early canonical amino acids. For example, incorporation of Dap and Dab into early oligopeptides could have provided positive charges that would be important for interaction with RNA molecules in the abiotic absence of Arg, Lys and His (Vázquez-Salazar and Lazcano 2018). All the tested non-canonical amino acids are still present in living organisms and fulfil important biological functions, albeit not universally distributed across three domains of life and mainly recruited for non-ribosomal peptide synthesis in microorganisms (Perkins and Cummins 1964; Kisumi *et al.* 1976; Fichtner *et al.* 2017; Schomakers *et al.* 2024). However, some of them, namely Nva and Nle, can be incorporated into polypeptide chains during translation due to the substrate promiscuity of leucyl- and methionyl-tRNA synthetases, resulting in formation of norvalyl-tRNA^{Leu} and norleucyl-tRNA^{Met} (Bogosian *et al.* 1989; Apostol *et al.* 1997). The anaerobic cultivation of *Escherichia coli* leads to the rapid accumulation of pyruvate that is subsequently converted into Nva and Nle, which led to the speculation that they may have been abundant protein components before the advent of molecular oxygen (Soini *et al.* 2008; Alvarez-Carreño *et al.* 2013).

Despite the prebiotic availability of non-canonical amino acids, primordial oligopeptides were most likely selectively enriched in early canonical amino acids due to their higher chemical reactivity in condensation reactions. Brack (1987) demonstrated that Leu undergoes easier aqueous oligomerization in the presence of activating reagent (N,N'-carbonyldiimidazole) compared to shorter non-proteinogenic 2-aminoisobutyric and 4-aminobutyric acids. In the same study, oligomerization of a mixture of proteinogenic (Gly, Ala, Val, Glu) and non-proteinogenic (Aba, β-alanine, 2-aminoisobutyric acid, 4-aminobutyric acid, isovaline) aliphatic amino acids in ratios similar to that found in the Murchinson meteorite led to the selective enrichment of the resulting peptides in proteinogenic amino acids. Similarly,

peptides containing prebiotically available positively charged amino acids (Dab, Orn) would not be efficiently produced in abiotic or proto-ribosome-mediated condensation reactions. For Fmoc-based solid-phase synthesis (employed for the synthesis of 11D peptide library), the side chain amino groups of Dab and Orn are modified with tert-butyloxycarbonyl (Boc) protecting group, enabling their efficient incorporation into a peptide chain (Conda-Sheridan and Krishnaiah 2020). However, peptide synthesis with unprotected Dab and Orn results in the premature termination of the synthesis due to their intramolecular cyclization into thermodynamically stable lactams (Poduška *et al.* 1965; Weber and Miller 1981). Moreover, Dap and Dab residues in peptides were shown to undergo spontaneous acyl migration, which disrupts the regular arrangement of peptide bonds (Poduška *et al.* 1965). Thoma and Powner (2023) demonstrated the effective coupling of Dap and Lys thioacids with glycine nitrile into peptides, which was not observed for Dab and Orn thioacids due to their spontaneous lactamization. In another study, Frenkel-Pinter *et al.* (2019) showed that basic canonical amino acids (Lys, Arg, His) mixed with simple hydroxy acids (glycolic acid, lactic acid) are more easily incorporated into linear depsipeptides during wet-dry cycling than basic non-canonical Dap, Dab, and Orn. All these factors could have largely contributed to the exclusion of non-canonical amino acids from the earliest oligopeptides and the early version of the genetic code.

5.2. Random negatively charged peptides undergo spontaneous self-assembly

The co-evolution theory implies the existence of the intermediate stage of life evolution, at which structural and functional polypeptides were composed of the limited set of early canonical amino acids (Wong 2005; Trifonov 2008; Higgs and Pudritz 2009; Longo and Blaber 2012). The early canonical amino acid alphabet has been deemed controversial by several origin-of-life researchers due to the limited functionality of early canonical amino acids and its predominantly acidic character (Griffith 2009; Raggi *et al.* 2016; Blanco *et al.* 2018; Gospodinov and Kunnev 2020). The acidic character of early canonical subset is thought to hinder interactions of primordial oligopeptides with negatively charged RNA molecules that were presumably necessary for their symbiotic co-evolution (Griffith 2009; Raggi *et al.* 2016; Blanco *et al.* 2018; Gospodinov and Kunnev 2020). To address the evolutionary plausibility of the early canonical amino acid subset, we assessed the biophysical and structural properties conferred by uncharged, negatively charged, positively charged and mixed-charge amino acid subsets that included both short-chain (Dap, Dab, Asp, Glu) and long-chain (Lys, Arg, Aad, Api) charged amino acids. All amino acid subsets share the common core represented by eight uncharged early canonical amino acids (Gly, Ala, Val, Leu, Ile, Pro, Ser, Thr). Long-chain

negatively charged amino acids (Aad, Api) were identified in smaller quantities than their respective canonical alternatives (Asp, Glu) in carbonaceous chondrites and the Ryugu asteroid but are absent in abiotic synthesis simulations; thus, they are likely to be scarce in the prebiotic environment (Koga and Naraoka 2017; Potiszil *et al.* 2023). Most of the tested combinatorial peptide libraries are not relevant to early evolutionary history of life; instead, their properties are compared to elucidate whether the negative net charge could provide certain evolutionary advantages to early oligopeptides. All experiments performed in this study follow the same methodology established in the previous study.

The solubility analysis reveals the general propensity of negatively charged peptide libraries (8E+Asp,Glu (10E) and 8E+Aad,Api) to have lower solubility with decreasing pH and increasing ionic strength, which is consistent with what has been described in section 5.1. In contrast, the positively charged and mixed-charge peptide libraries are fully solubilised across the entire range of pH and ionic strengths. The high solubility of positively charged peptide libraries is rather surprising since the opposite trend was expected due to the predominantly basic character of individual peptide variants and deprotonation of their side chains at higher pH. The extremely high solubility of mixed-charge peptide libraries is plausibly caused by the broad distribution of isoelectric points of individual peptide variants and the influence of charge-charge interactions on the side chain pK_a of ionisable residues (Laurents *et al.* 2003; Baran *et al.* 2008). The remarkably low solubility of 8E peptide library almost invariant on pH and ionic strength is in line with the absence of charged residues and the predominantly hydrophobic character of uncharged early canonical amino acids.

The aggregation propensity analysis demonstrates that increasing pH decreases the elution volumes of monomer peaks of negatively charged peptide libraries. This behaviour is consistent with the plausible compaction of negatively charged peptide variants at lower pH as well as electrostatic repulsion with the negatively charged Superdex column matrix at higher pH (Drevin *et al.* 1990). In contrast, the elution volumes of monomer peaks of positively charged libraries are almost invariant on pH, although the monomer peaks become significantly broader with increasing pH possibly due to electrostatic attraction between positively charged peptide variants and the negatively charged column matrix. Both effects are less pronounced for the mixed-charge peptide libraries due to the simultaneous presence of positively and negatively charged residues. The high ionic strength decreases the differences in elution volumes and narrows monomer peaks of all peptide libraries by supressing electrostatic interactions. The monomer peaks of 8E peptide library are uniform across the entire range of pH and ionic strengths due to the absence of charged residues.

The formation of higher-order soluble aggregates was observed only for negatively charged peptide libraries at near-neutral and alkaline pH. The biophysical and structural analyses of the individual fractions of 8E+Asp,Glu (10E) peptide library indicate that whereas the majority of peptide variants in 10E peptide library are mostly unstructured and monomeric in aqueous solutions, a certain subpopulation of peptides spontaneously assembles into highly heterogenous soluble aggregates consisting primarily of β-sheets. The compositional analysis reveals the enrichment of the soluble aggregates in branched aliphatic (Leu, Val, Ile) and depletion in Pro and charged (Asp, Glu) amino acids, which is consistent with several studies elucidating the propensities of proteinogenic amino acids to form β-sheets (Nowick and Insaf 1997; Street and Mayo 1999). The detailed analysis of individual fractions was not performed for 8E+Aad,Api peptide library due to its low yield from Fmoc-based solid-phase synthesis.

Far-UV CD spectroscopy demonstrates that positively charged and mixed-charge peptide libraries are characterised by significantly lower secondary structure content than the uncharged and negatively charged peptide libraries across the entire pH range. The hydrophobic interactions between aliphatic side chains tend to favour the formation of α-helical or β-sheet conformations in short peptides, resulting in the high secondary structure content of 8E peptide library (Blout *et al.* 1960; Kubota and Fasman 1974; Dias *et al.* 2011). The mixed-charge peptide libraries are largely unstructured plausibly due to the prevalence of destabilising salt bridge interactions over stabilising salt bridge interactions between positively and negatively charged residues in highly randomised peptide libraries (Scholtz *et al.* 1993; Musafia *et al.* 1995). The low secondary structure content of positively charged peptide libraries most likely stems from the destabilising effect of electrostatic repulsions between amino acid residues, as evidenced by structural studies of charged amino acid homopolymers (Davidson and Fasman 1967; Mirtič and Grdadolnik 2013; Morga *et al.* 2022). The secondary structure content of 8E+Lys,Arg peptide library slightly increases with increasing pH due to deprotonation of Lys and Arg residues and remains relatively higher than that of 8E+Dap,Dab peptide library, which is consistent with the observation that α-helical propensity of basic amino acids increases with the length of the side chain (Padmanabhan *et al.* 1996). The observed differences in α-helical propensity were speculated to result from hydrogen bonding between the short side-chain amino group and the peptide backbone, which interferes with α-helix formation (Padmanabhan *et al.* 1996).

At acidic pH, the higher secondary structure content of negatively charged peptide libraries can be attributed to protonation of Asp and Glu residues, which reduces electrostatic repulsion between them (Blout and Idelson 1956; Inoue *et al.* 2005). At near-neutral and

alkaline pH, the formation of soluble β-sheet aggregates largely but not solely contributes to the superior secondary structure content of 8E+Asp,Glu (10E) peptide library since its monomer fraction remains more structured than positively charged and mixed-charged peptide libraries at physiological pH. 8E+Aad,Api peptide library has lower secondary structure content than 8E+Asp,Glu peptide library in aqueous solutions but demonstrates remarkable secondary-structure forming propensity upon TFE titration. In general, the secondary structure-forming propensity of peptide libraries upon TFE addition correlates with the length of amino acid side chains since the peptide libraries that contain long-chain charged amino acids (Lys, Arg, Aad, Api) are more structured than the respective peptide libraries composed of their short-chain analogues (Dap, Dab, Asp, Glu). This tendency most likely arises from the higher hydrophobicity of the longer side chains.

Overall, the outcomes of this study demonstrate that the early canonical amino acid subset confers a unique balance between biophysical and structural properties to polypeptides and could have provided a "stepping stone" towards the emergence of soluble and folded proteins. The uncharged peptide library is characterised by high secondary structure-forming propensity but low solubility, whereas the positively charged and mixed-charged peptide libraries are remarkably soluble but largely unstructured in aqueous solutions. The short peptides composed of the negatively charged amino acid subsets demonstrate remarkable solubility but at the same time retain the ability to fold, mostly through the assembly into higher-order β-sheet structures. Predominantly considered pathological in the context of modern cellular life, aggregation could have provided a path towards first folded and functional oligopeptides as well as ensured their survival and accumulation in the prebiotic environment at the earliest stages of life evolution. It is well-known that unfolded proteins are more susceptible to proteolytic degradation than folded proteins, mostly due to the higher accessibility of conformationally flexible polypeptide chains to the active sites of proteases (Fontana *et al.* 2004). Brack (1987) showed that peptides composed of alternating hydrophilic (Lys) and hydrophobic (Val/Leu) proteinogenic amino acids adopt β-sheet conformation and are 15 times more resistant to mild acidic hydrolysis (0.1 M HCl, 110 °C) than the analogous unstructured peptides containing non-proteinogenic hydrophobic (Aba/Nva) amino acids. Similarly, Zhang *et al.* (1994) characterised 16-mer peptide composed of alternating Ala and charged (Glu/Lys) residues with unusually stable β-sheet structure that pertains over the wide range of pH (1.5–11), temperatures (25–90 °C) and denaturing conditions (SDS, urea). The self-assembly of short peptides into insoluble amyloid fibrils characterised by the cross β-sheet structure has been shown to provide high stability towards harsh environmental conditions as

well as diverse functionalities described in section 1.3. Even though currently characterised amyloidogenic peptides are enriched in late proteinogenic amino acids, Greenwald *et al.* (2016) demonstrated that COS-mediated condensation of the mixture of early canonical amino acids (Gly, Ala, Val, Asp) leads to formation of amyloid fibrils enriched in Ala and Val and nearly devoid of Asp. The soluble β-sheet aggregates formed from random peptide sequences composed solely of early canonical amino acids and observed for the first time in our study may represent an evolutionary link between primordial insoluble amyloid fibrils and extant soluble globular proteins. The catalytic and nucleotide-binding propensities of individual fractions of 8E+Asp,Glu (10E) peptide library are being currently characterised in collaboration with Prof. Florian Hollfelder's research group (University of Cambridge, UK) and Prof. Dieter Braun's research group (Ludwig Maximilian University of Munich, Germany).

The gradual evolution of the genetic code heavily implies that proto-/early cellular life could be supported by folded and functional oligopeptides composed of reduced amino acid subsets (Wong 2005; Trifonov 2008; Higgs and Pudritz 2009; Longo and Blaber 2012). Numerous theoretical and experimental studies confirm that diverse protein folds can be optimally encoded by a minimal set of 7–13 amino acids (Riddle *et al.* 1997; Murphy *et al.* 2000; Longo and Blaber 2012; Longo *et al.* 2013; Solis 2019). The sufficient functionality of reduced amino acid subsets has been confirmed by experimental studies that demonstrate the partial or complete preservation of enzymatic and RNA-binding functions in protein mutants constructed from 9- to 17-amino acid alphabets (Akanuma *et al.* 2002; Walter *et al.* 2005; Shibue *et al.* 2018; Makarov *et al.* 2021; Giacobelli *et al.* 2022; Zhao and Akanuma 2023). In particular, the author of the dissertation contributed to the reconstruction of plausible early variants of the C-terminal domain of ribosomal protein uL11 from *Bacillus stearothermophilus* and dephosphocoenzyme A kinase from *Aquifex aeolicus* that are composed of 10 and 17 canonical amino acids, respectively (Makarov *et al.* 2021; Giacobelli *et al.* 2022). The comprehensive bioinformatic analysis performed by McDonald and Storrie-Lombardi (2010) indicates the presence of 61 proteins lacking positively charged Lys and Arg residues in the extant bacterial and archaeal proteomes, including LysW protein involved in Lys biosynthesis. Possibly, the strongest support towards the gradual expansion of the genetic code comes from the partial reconstruction of Cys-free variants of enzymes involved in Cys biosynthesis, demonstrated by Fujishima *et al.* (2018), as well as the complete establishment of Trp biosynthetic pathway from Trp-less mutants in *Escherichia coli* performed in our laboratory (manuscript in preparation). These findings imply the possibility of biosynthesis of late

proteinogenic amino acids before their recruitment to the genetic code by a system of ancestral metabolic catalysts that lacked the respective amino acids.

Despite the apparent optimality of the early canonical amino acid alphabet, the genetic code was expanded with late proteinogenic amino acids to improve the structural stability and diversify the functionalities of growing polypeptides. The most notable is the contribution of aromatic amino acids to the hydrophobic core packing required for folding of longer polypeptide chains into a globular conformation (Shortle *et al.* 1990; Eriksson *et al.* 1992; Longo *et al.* 2015). The meta-ranking of amino acid residues promoting the disorder or order in extant proteins indicates that aromatic amino acids contribute the most to the emergence of ordered protein structures, exceeding aliphatic amino acids (Campen *et al.* 2008). The significance of aromatic amino acids to the tertiary (rather than secondary) structure formation is evidenced by the drastic loss of conformational stability in aromatics-less variants of dephosphocoenzyme A kinase despite the high secondary structure preservation (Makarov *et al.* 2021). The same observation can be inferred from the analysis of combinatorial peptide libraries, which demonstrates that inclusion of Tyr does not increase the secondary structure content of peptide libraries in aqueous solutions but remarkably improves the folding of peptides in the hydrophobic environment presumably mimicking the interior of globular proteins (TFE solutions). In addition to the remarkable structure-forming propensities, aromatic (as well as sulphur-containing) amino acids were plausibly recruited to the genetic code due to their unique antioxidant properties that would enable protection of polypeptides against oxidative stress caused by the introduction of molecular oxygen into the atmosphere at later stages of life evolution (Granold *et al.* 2018).

In contrast to aromatic amino acids, positively charged amino acids (Lys, Arg) are overrepresented in intrinsically disordered regions of proteins (Uversky *et al.* 2000; Campen *et al.* 2008). The inclusion of positively charged amino acids does not significantly improve (11R peptide library) and even diminishes (8E+Lys,Arg,Asp,Glu peptide library) the secondary structure content of peptide libraries both in aqueous and TFE solutions. Besides the central role of His in modern enzyme catalysis, positively charged amino acids (Lys, Arg) are mostly exposed to the globular protein surface and play a key role in molecular recognition of protein-protein and protein-nucleic acid complexes (Yokota *et al.* 2006; Campen *et al.* 2008). Many origin-of-life researchers speculate that positively charged amino acids were recruited to the genetic code to facilitate the interaction of ancestral RNA molecules with early oligopeptides that served as cofactors of RNA-based replication and catalysis in the hypothetical stage of life evolution known as the peptide-RNA world (Alva *et al.* 2015; van der Gulik and Speijer 2015;

Fried *et al.* 2022). The RNA-binding propensity conferred by positively charged (as well as other late) proteinogenic amino acids and its possible involvement in the early ribosome evolution will be discussed in the next section.

5.3. Late proteinogenic amino acids were crucial for the early ribosome evolution

The consensus between different evolutionary hypotheses implies that the ribosome originated prior to the emergence of the LUCA from an ancestral RNA fragment possessing tRNA-like structural and functional properties (section 1.4). The ancestral tRNA-like fragment underwent dimerization and fusion producing ancestral PTC, which later evolved into large ribosomal subunit by accretion of rRNA and ribosomal proteins, presenting what may be the best-preserved molecular fossil of RNA-peptide co-evolution. In this study, we utilised two proto-ribosome models that presumably represent different stages of the early ribosome evolution. The proto-ribosome inferred by Bose *et al.* (2022) follows the dimerization model and consists of the minimal part of the two-fold semi-symmetrical PTC region of 23S rRNA from *Thermus thermophilus*. The inferred WT sPTC rRNA is composed of three 23S rRNA fragments that are connected by three stem loops (5'-CUUCGG-3'). In contrast, the proto-ribosome derived by Hsiao *et al.* (2013) is based on the accretion model and includes fragments of 23S rRNA and five ribosomal proteins from *Thermus thermophilus* that lie within a hypothetical sphere that is centred at the peptidyl transfer site and has a radius of 30 Å. The inferred WT bPTC rRNA consists of 13 23S rRNA fragments that are connected by 11 stem loops (5'-GCCGUAAGGC-3') to achieve the best preservation of the PTC three-dimensional structure. The ancestral ribosomal peptides (rPeptides) are represented by largely unstructured tails of ribosomal proteins uL2, uL3, uL4, uL15, and uL22 that extend from their globular domains towards the PTC. The extended tails are characterised by high structural conservation across all domains of life and are considered the oldest components of the respective ribosomal proteins predating the emergence of their globular domains (Klein *et al.* 2004a; Hsiao *et al.* 2009). The extended tails are not found closer than 18 Å to the PTC, indicating that ribosomal proteins are not directly involved in the peptide bond formation (Nissen *et al.* 2000). Moreover, two research groups have independently developed ~70–140 nucleotides long rRNA constructs that are derived from the semi-symmetrical region of the PTC and are able to catalyse the peptidyl transfer reaction in the absence of ribosomal proteins, supporting the sovereignty of early rRNA evolution (Bose *et al.* 2022; Kawabata *et al.* 2022). However, the peptide ancestors of modern ribosomal proteins could have still provided certain evolutionary advantages necessary for the proto-ribosome survival and accumulation in the prebiotic environment. The

aim of this project was to assess the evolutionary relevance of ancestral rPeptide-rRNA interactions at two different stages of the early ribosome evolution.

The sequences of the tested rPeptides are well-conserved across bacterial species and are highly enriched in (i) Gly residues that ensure the high conformational flexibility and the tight packing of polypeptide chains and rRNA; and (ii) positively charged (Lys, Arg, His) residues that are involved in electrostatic interactions with the phosphate groups of rRNA (Klein *et al.* 2004a; Codispoti *et al.* 2024). The rPeptides were chemically synthesised using Fmoc-based solid-phase methodology, and their interactions with *in vitro* transcribed and fluorescein-labelled PTC rRNA constructs in the absence and presence of metal ions (1 mM K^+, Mg^{2+}, Ca^{2+}) were measured by microscale thermophoresis. The metal ions were chosen based on their prevalence in the primordial ocean as well as their abundance in extant ribosomal structures (Klein *et al.* 2004b; Ueda and Shibuya 2021). Microscale thermophoresis demonstrates that most rPeptides interact with both PTC rRNA constructs with moderate micromolar affinities. The similar micromolar binding affinities to WT bPTC rRNA were estimated in the absence of metal ions by Hsiao *et al.* (2013) for pL3, pL15 and pL22 rPeptides that were prepared as fusions with the maltose-binding protein. The absence of interaction with WT bPTC rRNA was reported for pL2 peptide fused with the maltose-binding protein (Hsiao *et al.* 2013). In our study, the interaction could not be detected only for pL4 rPeptide with WT bPTC rRNA in the presence of metal ions and pL2 rPeptide with WT sPTC rRNA with and without metal ions. The presence of metal ions reduces non-specific binding of positively charged rPeptides to negatively charged rRNA, as evidenced by the significant decrease of the binding stoichiometry of rPeptides with WT bPTC rRNA and the moderate (two- to threefold) decrease in the dissociation constants for both PTC rRNA constructs. The strongest interactions are observed for pL22 rPeptide that is also characterised by the highest positive charge per residue, confirming the large contribution of electrostatic interactions in the formation of PTC rRNA-rPeptide complexes. The observed interactions underlie the coacervation of both PTC rRNA constructs with the equimolar mixture of rPeptides (as well as pL22 rPeptide individually) demonstrated by MSc. Simone Codispoti and Tomoko Yamaguchi, PhD. The coacervation of rPeptides with PTC rRNA constructs is sequence- and structure-specific since the shuffled variants of both WT bPTC rRNA and WT sPTC rRNA demonstrated lower coacervation propensity leading to the frequent formation of insoluble aggregates or precipitation. Moreover, the resulting coacervates were shown to withstand extreme temperatures (up to 90 °C) and ensure protection of ancestral PTC rRNAs against nucleolytic degradation.

Overall, the findings of our study indicate that interactions between primordial ribosomal peptides enriched in late proteinogenic amino acids and ancestral PTC rRNA were essential for the early ribosome evolution. In particular, positively charged residues aided the early rRNA evolution by prompting liquid-liquid phase separation that provided the plausible means of rRNA stabilisation and accumulation in the prebiotic environment necessary for its subsequent accretion and formation of the large ribosomal subunit. The coacervation of the oppositely charged polyelectrolytes is currently a well-characterised phenomenon that has been observed for numerous polypeptide-polynucleotide systems (Hyman *et al.* 2014; Poudyal *et al.* 2018). The liquid-liquid phase separation was also recently observed between two positively charged polyelectrolytes, which was strongly mediated by cation–π interactions, but is yet to be demonstrated between two negatively charged biopolymers (Kim *et al.* 2016). The coacervation is considered one of the main driving forces of early molecular evolution enabling the membrane-free compartmentalisation, concentration and stabilisation of early biomolecules and their building blocks, which has been demonstrated in the context of the early ribosome evolution for the first time in our study (Keating 2012; Poudyal *et al.* 2018; Codispoti *et al.* 2024). Moreover, coacervation induced by short cationic peptides (poly-L-lysine and poly-L-arginine) was shown to enhance the activity of several ribozymes and even promote the ligation of short RNA oligonucleotides over their cleavage (Drobot *et al.* 2018; Poudyal *et al.* 2019; Le Vay *et al.* 2021).

The structural analyses of both natural and artificially evolved protein-RNA interactions agree that proteins build these interactions mainly via positively charged and aromatic amino acids (Ellis *et al.* 2007; Blanco *et al.* 2018; Srivastava *et al.* 2018). The prevalence of positively charged residues in protein-RNA interactions has led to the speculation that basic canonical amino acids were one of the earliest additions to the genetic code despite the numerous reports on their abiotic scarcity (Griffith 2009; Raggi *et al.* 2016; Blanco *et al.* 2018; Gospodinov and Kunnev 2020). In the abiotic absence of positively charged and aromatic residues, primordial protein-RNA interactions could have been mediated by negatively charged early canonical amino acids (Asp, Glu) with the assistance of metal ions (Vázquez-Salazar and Lazcano 2018). Few Mg^{2+}-mediated protein-RNA interactions appear to be preserved in the large ribosomal subunit even after the complete establishment of the standard genetic code (Petrov *et al.* 2012). Several Asp-enriched binding motifs are present in the active sites of extant enzymes and mediate metal ion-dependent interactions with the phosphate groups of respective substrates: DxDGD motif present in DNA- and RNA-dependent RNA polymerases; DGDxD motif employed by several mutases; and DAKVGDGD motif of dihydroxyacetone kinase (van

der Gulik *et al.* 2009). All of the catalytic motifs share the common DGD submotif that is speculated to be a common ancestral peptide that was capable of coordination with divalent metal ions and was conserved in the active sites of enzymes over the course of evolution. Another line of evidence comes from *in vitro* selection studies that successfully constructed RNA-binding oligopeptides from reduced amino acid subsets. Kumachi *et al.* (2016) utilised cDNA display to develop a 30-mer peptide that is composed solely of four canonical amino acids (Gly, Ala, Val, Asp) and is able to bind to the total tRNA from *Escherichia coli* with moderate micromolar affinity (66±3 μM). Similarly, our laboratory employed mRNA display to reconstruct the early variant of the C-terminal domain of ribosomal protein uL11 from the early canonical amino acid subset (Giacobelli *et al.* 2022). The resulting early mutant retains (2.2±0.3 nM) and even exceeds the original (25±6 nM) nanomolar binding affinity to the cognate RNA despite the drastic loss of the secondary and tertiary structure. The molecular dynamic simulations demonstrate that, in the absence of aromatic and basic residues, the early mutant binds the cognate RNA via K^+/Mg^{2+}-mediated ion bridging interactions between Glu residues and the sugar-phosphate backbone of RNA. The early mutant-RNA complex was shown to be more liable towards extreme pH and high temperatures than the respective complex formed by the wild-type protein.

The alternative hypothesis assumes that early protein-RNA interactions were mediated by positively charged non-canonical amino acids (Dap, Dab, Orn) that were abundant in the prebiotic environment (Vázquez-Salazar and Lazcano 2018). Similarly to their canonical analogues, basic non-canonical amino acids could also have potentially aided the early nucleic acid evolution. Longo *et al.* (2020) demonstrated that 29-mer peptide derived by the ancestral sequence reconstruction of four helix-hairpin-helix protein families retains (to a lower extent) its propensity to coacervate with polyuridine after replacement of Arg residues to Orn. Frenkel-Pinter *et al.* (2020) showed that cationic depsipeptides containing non-proteinogenic Dap, Dab, and Orn increase the thermal stability of a short RNA duplex, albeit to a lesser magnitude that the respective depsipeptides containing proteinogenic Lys, Arg, His. However, given the inefficient incorporation of Dap, Dab and Orn into growing polypeptide chains discussed in section 5.1, short peptides bearing positively charged non-canonical amino acids are less likely to be formed in the prebiotic environment, thus necessitating the recruitment of basic canonical amino acids (Lys, Arg, His) to the genetic code. The numerous studies (including ours) indicate that cooperative peptide-RNA interactions mediated by positively charged (Lys, Arg, His) residues attenuate the rates of their prebiotic synthesis and hydrolytic degradation and mutually improve their structural stability and catalytic functions, thus underlying the symbiotic co-

evolution of proto-peptides and ancestral RNAs at the early stages of life evolution (Barbier *et al.* 1993; Herschlag *et al.* 1994; Kashiwagi *et al.* 2009; Carny and Gazit 2011; Tagami *et al.* 2017; Frenkel-Pinter *et al.* 2020; Seal *et al.* 2022).

6. Conclusions

The main objective of this dissertation was to determine the selective factors that led to (i) the inclusion of early proteinogenic amino acids over their prebiotically available non-proteinogenic alternatives; and (ii) the inclusion of late proteinogenic amino acids into the genetic code in the context of early ribosome evolution. The conclusions of this dissertation are the following:

1. Despite the high prebiotic availability, aliphatic (2-aminobutyric acid, norvaline, norleucine) and positively charged (2,3-diaminopropionic acid, 2,4-diaminobutyric acid) non-canonical amino acids were not included in the genetic code likely due to their propensity to form oversolubilised oligopeptides with low secondary structure content.
2. The early canonical amino acid alphabet confers an optimal balance of solubility, aggregation propensity, and secondary structure formation that would be required for the emergence of earliest folded and functional oligopeptides.
3. Despite their abiotic scarcity, late canonical amino acids, in particular positively charged Lys, Arg, and His, brought multiple functional advantages upon their recruitment to the genetic code. Besides their catalytic potential, they have substantial ability to electrostatically interact with RNA/DNA. Such properties could be adaptive during early stages of ribosomal evolution as such interaction was shown to induce coacervation of ancestral ribosomal RNA that ensured its compartmentalisation, stabilisation and accumulation in the prebiotic environment.

7. References

Adamala K., Szostak J.W. Competition between model protocells driven by an encapsulated catalyst. *Nat. Chem*. **5**:6, 495–501 (2013).

Adler-Abramovich L., Vaks L., Carny O., Trudler D., Magno A., Caflisch A., Frenkel D., Gazit E. Phenylalanine assembly into toxic fibrils suggests amyloid etiology in phenylketonuria. *Nat. Chem. Biol*. **8**:8, 701–6 (2012).

Agmon I., Bashan A., Zarivach R., Yonath A. Symmetry at the active site of the ribosome: Structural and functional implications. *Biol. Chem.* **386**, 833–44 (2005).

Akanuma S., Kigawa T., Yokoyama S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc. Natl. Acad. Sci. USA*. **99**:21, 13549−53 (2002).

Akashi H., Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. USA*. **99**, 3695–700 (2002).

Al-Garawi Z.S., McIntosh B.A., Neill-Hall D., Hatimy A.A., Sweet S.M., Bagley M.C., Serpell L.C. The amyloid architecture provides a scaffold for enzyme-like catalysts. *Nanoscale*. **9**:30, 10773–83 (2017).

Altay M., Altay Y., Otto S. Parasitic behavior of self-replicating molecules. *Angew. Chem. Int. Ed. Engl*. **57**:33, 10564–8 (2018).

Alva V., Söding J., Lupas A.N. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*. **4**, e09410 (2015).

Alvarez-Carreño C., Becerra A., Lazcano A. Norvaline and norleucine may have been more abundant protein components during early stages of cell evolution. *Orig. Life Evol. Biosph*. **43**:4–5, 363–75 (2013).

Ambrogelly A, Palioura S, Söll D. Natural expansion of the genetic code. *Nat. Chem. Biol*. **3**:1, 29–35 (2007).

Amend J.P., Shock E.L. Energetics of amino acid synthesis in hydrothermal ecosystems. *Science*. **281**:5383, 1659–62 (1998).

Andersson G.E., Kurland C.G. An extreme codon preference strategy: Codon reassignment. *Mol. Biol. Evol*. **8**, 530–44 (1991).

Apostol I., Levine J., Lippincott J., Leach J., Hess E., Glascock C.B., Weickert M.J., Blackmore R. Incorporation of norvaline at leucine positions in recombinant human hemoglobin expressed in Escherichia coli. *J. Biol. Chem*. **272**:46, 28980–8 (1997).

Archetti M. Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* **59**:3, 400–15 (2004).

Aubrey A.D., Cleaves H.J., Bada J.L. The role of submarine hydrothermal systems in the synthesis of amino acids. *Orig. Life Evol. Biosph*. **39**:2, 91–108 (2009).

Baker E. G., Bartlett G. J., Porter Goff K. L., Woolfson D. N. Miniprotein design: Past, present, and prospects. *Acc. Chem. Res*. **50**:9, 2085−92 (2017).

Ban N., Beckmann R., Cate J.H., Dinman J.D., Dragon F., Ellis S.R., Lafontaine D.L., Lindahl L., Liljas A., Lipton J.M., McAlear M.A., Moore P.B., Noller H.F., Ortega J., Panse V.G., Ramakrishnan V., Spahn C.M., Steitz T.A., Tchorzewski M., Tollervey D., Warren A.J., Williamson J.R., Wilson D., Yonath A., Yusupov M. A new system for naming ribosomal proteins. *Curr. Opin. Struct. Biol*. **24**, 165–9 (2014).

Baran K.L., Chimenti M.S., Schlessman J.L., Fitch C.A., Herbst K.J., Garcia-Moreno B.E. Electrostatic effects in a network of polar and ionizable groups in staphylococcal nuclease. *J. Mol. Biol*. **379**:5, 1045–62 (2008).

Barbier B., Visscher J., Schwartz A.W. Polypeptide-assisted oligomerization of analogs in dilute aqueous solution. *J. Mol. Evol*. **37**, 554–8 (1993).

Belousoff M.J., Davidovich C., Zimmerman E., Caspi Y., Wekselman I., Rozenszajn L., Shapira T., Sade-Falk O., Taha L., Bashan A., Weiss M.S., Yonath A. Ancient machinery embedded in the contemporary ribosome. *Biochem. Soc. Trans*. **38**:2, 422–7 (2010).

Bernier C.R., Petrov A.S., Kovacs N.A., Penev P.I., Williams L.D. Translation: the universal structural core of life. *Mol. Biol. Evol*. **35**:8, 2065–76 (2018).

Bieling P., Beringer M., Adio S., Rodnina M.V. Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nat. Struct. Mol. Biol*. **13**:5, 423–8 (2006).

Biron J.P., Parkes A.L., Pascal R., Sutherland J.D. Expeditious, potentially primordial, aminoacylation of nucleotides. *Angew. Chem. Int. Ed. Engl*. **44**:41, 6731–4 (2005).

Blagojevic V., Petrie S., Bohme D.K. Gas-phase syntheses for interstellar carboxylic and amino acids. *Mon. Not. R. Astron. Soc*. **339**, 7–11 (2003).

Blanco C., Bayas M., Yan F., Chen I.A. Analysis of evolutionarily independent protein-RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Curr. Biol*. **28**:4, 526–37 (2018).

Blout E.R., Idelson M. Polypeptides. VI. Poly-α-L-glutamic acid: preparation and helix-coil conversions. *J. Am. Chem. Soc*. **78**:2, 497–8 (1956).

Blout E.R., de Lozé C., Bloom S.M., Fasman G.D. The dependence of the conformations of synthetic polypeptides on amino acid composition. *J. Am. Chem. Soc*. **82**:14, 3787–9 (1960).

Bogosian G., Violand B.N., Dorward-King E.J., Workman W.E., Jung P.E., Kane J.F. Biosynthesis and incorporation into protein of norleucine by Escherichia coli. *J. Biol. Chem*. **264**:1, 531–9 (1989).

Bokov K., Steinberg S.V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature*. **457**:7232, 977−80 (2009).

Bonner W.A., Blair N.E., Lemmon R.M. The radioracemization of amino acids by ionizing radiation: geochemical and cosmochemical implications. *Orig. Life*. **9**:4, 279–90 (1979).

Bose T., Fridkin G., Davidovich C., Krupkin M., Dinger N., Falkovich A.H., Peleg Y., Agmon I., Bashan A., Yonath A. Origin of life: protoribosome forms peptide bonds and links RNA and protein dominated worlds. *Nucleic Acids Res*. **50**:4, 1815–28 (2022).

Brack A., Orgel L.E. β structures of alternating polypeptides and their possible prebiotic significance. *Nature*. **256**:5516, 383–7 (1975).

Brack A. β-structures of alternating polypeptides and their possible role in chemical evolution. *Biosystems*. **9**:2–3, 99–103 (1977).

Brack A. Selective emergence and survival of early polypeptides in water. *Orig. Life Evol. Biosph*. **17**:3–4, 367–79 (1987).

Brearley A.J. The action of water. In: *Meteorites and the early Solar System II*. Lauretta D.S., McSween Jr. H.Y. (Eds.). Tucson: University of Arizona Press. pp. 587–624. (2006).

Brooks D.J., Fresco J.R. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Mol. Cell Proteomics*. **1**:2, 125–31 (2002).

Buck M. Trifluoroethanol and colleagues: cosolvents come of age. recent studies with peptides and proteins *Q. Rev. Biophys*. **31**:3, 297–355 (1998).

Budin I., Debnath A., Szostak J.W. Concentration-driven growth of model protocell membranes. *J. Am. Chem. Soc*. **134**:51, 20812–9 (2012).

Burton A.S., Stern J.C., Elsila J.E., Glavin D.P., Dworkin J.P. Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev*. **41**, 5459–72 (2012).

Burton Z.F. The 3-minihelix tRNA evolution theorem. *J. Mol. Evol*. **88**:3, 234–42 (2020).

Callahan M.P., Smith K.E., Cleaves H.J., Ruzicka J., Stern J.C., Glavin D.P., House C.H., Dworkin J.P. Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc. Natl. Acad. Sci. USA*. **108**:34, 13995–8 (2011).

Campbell T.D., Febrian R., McCarthy J.T., Kleinschmidt H.E., Forsythe J.G., Bracher P.J. Prebiotic condensation through wet-dry cycling regulated by deliquescence. *Nat. Commun*. **10**:1, 4508 (2019).

Campen A., Williams R.M., Brown C.J., Meng J., Uversky V.N, Dunker A.K. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett*. **15**:9, 956−63 (2008).

Carny O., Gazit E. Creating prebiotic sanctuary: self-assembling supramolecular peptide structures bind and stabilize RNA. *Orig. Life Evol. Biosph*. **41**:2, 121–32 (2011).

Cech T.R. Structural biology. The ribosome is a ribozyme. *Science*. **289**:5481, 878–9 (2000).

Chang S., Flores J., Ponnamperuma C. Peptide formation mediated by hydrogen cyanide tetramer: a possible prebiotic process. *Proc. Natl. Acad. Sci. USA*. **64**:3, 1011–5 (1969).

Chapman M.R., Robinson L.S., Pinkner J.S., Roth R., Heuser J., Hammar M., Normark S., Hultgren SJ. Role of Escherichia coli curli operons in directing amyloid fiber formation. *Science*. **295**:5556, 851–5 (2002).

Chellgren B.W., Creamer T.P. Side-chain entropy effects on protein secondary structure formation. *Proteins*. **62**:2, 411–20 (2006).

Chen F., Yang D. Condensation of amino acids to form peptides in aqueous solution induced by the oxidation of sulphur (IV): an oxidative model for prebiotic peptide formation. *Orig. Life Evol. Biosph*. **37**:1, 47–54 (2007).

Chiti F., Webster P., Taddei N., Clark A., Stefani M., Ramponi G., Dobson C.M. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc. Natl. Acad. Sci. USA*. **96**:7, 3590–4 (1999).

Chung N.M., Lohrmann R., Orgel L.E., Rabinowitz J. The mechanism of the trimetaphosphate-induced peptide synthesis. *Tetrahedron*. **27**, 1205–10 (1971).

Chyba C.F. Atmospheric science. Rethinking Earth's early atmosphere. *Science*. **308**:5724, 962–3 (2005).

Cleaves H.J., Chalmers J.H., Lazcano A., Miller S.L., Bada J.L. A reassessment of prebiotic organic synthesis in neutral planetary atmospheres. *Orig. Life Evol. Biosph*. **38**:2, 105–15 (2008).

Cleaves H.J. Prebiotic chemistry: what we know, what we don't. *Evo. Edu. Outreach.* **5**, 342–60 (2012).

Cobb A.K., Pudritz R.E. Nature's starships. I. Observed abundances and relative frequencies of amino acids in meteorites. *Astrophys. J*. **783**, 140 (2014).

Codispoti S., Yamaguchi T., Makarov M., Giacobeli V.G., Mašek M., Kolář M.H., Sanchez Rocha A.C., Fujishima K., Zanchetta G., Hlouchová K. The interplay between peptides and RNA is critical for protoribosome compartmentalization and stability. *Nucleic Acids Res*. **28**, gkae823 (2024).

Cohen Z.R., Todd Z.R., Maibaum L., Catling D.C., Black R.A. Stabilization of prebiotic vesicles by peptides depends on sequence and chirality: A mechanism for selection of protocell-associated peptides. *Langmuir*. **40**:17, 8971–80 (2024).

Conda-Sheridan M., Krishnaiah M. Protecting groups in peptide synthesis. *Methods Mol. Biol.* **2103**, 111–28 (2020).

Cornell C.E., Black R.A., Xue M., Litz H.E, Ramsay A., Gordon M., Mileant A., Cohen Z.R., Williams J.A., Lee K.K., Drobny G.P., Keller S.L. Prebiotic amino acids bind to and stabilize prebiotic fatty acid membranes. *Proc. Natl. Acad. Sci. USA*. **116**:35, 17239–44 (2019).

Cornette J.L., Cease K.B., Margalit H., Spouge J.L., Berzofsky J.A., DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol*. **195**:3, 659−85 (1987).

Crick F.H. Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol*. **19**:2, 548–55 (1966).

Crick F.H. The origin of the genetic code. *J. Mol. Biol*. **38**, 367–79 (1968).

Dalai P., Pleyer H.L., Strasdeit H., Fox S. The influence of mineral matrices on the thermal behavior of glycine. *Orig. Life Evol. Biosph*. **47**:4, 427–52 (2017).

Dalrymple G.B. The age of the Earth in the twentieth century: A problem (mostly) solved. *Geol. Soc. Spec. Pub*. **190**, 205−21 (2001).

Danger G., Boiteau L., Cottet H., Pascal R. The peptide formation mediated by cyanate revisited. N-carboxyanhydrides as accessible intermediates in the decomposition of N-carbamoylamino acids. *J. Am. Chem. Soc*. **128**:23, 7412–3 (2006).

Das R., Kiley P.J., Segal M., Norville J., Yu A.A., Wang L., Trammell S.A., Reddick L.E., Kumar R., Stellacci F., Lebedev N., Schnur J., Bruce B.D., Zhang S., Baldo M. Integration of photosynthetic protein molecular complexes in solid-state electronic devices. *Nano Lett*. **4**:6, 1079–83 (2004).

Davidovich C., Belousoff M., Bashan A., Yonath A. The evolving ribosome: from non-coded peptide bond formation to sophisticated translation machinery. *Res. Microbiol*. **160**:7, 487-92 (2009).

Davidson B., Fasman G.D. The conformational transitions of uncharged poly-L-lysine. Alpha helix-random coil-beta structure. *Biochemistry*. **6**:6, 1616–29 (1967).

Demongeot J., Seligmann H. Evolution of small and large ribosomal RNAs from accretion of tRNA subelements. *Biosystems*. **222**, 104796 (2022).

Di Giulio M. On the origin of the transfer RNA molecule. *J. Theor. Biol*. **159**:2, 199–214 (1992).

Dias C.L., Karttunen M., Chan H.S. Hydrophobic interactions in the formation of secondary structures in small peptides. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys*. **84**, 041931 (2011).

Dick T.P., Schamel W.A. Molecular evolution of transfer RNA from two precursor hairpins: implications for the origin of protein synthesis. *J. Mol. Evol*. **41**:1, 1–9 (1995).

Dong J., Shokes J.E., Scott R.A., Lynn D.G. Modulating amyloid self-assembly and fibril morphology with Zn(II). *J. Am. Chem. Soc*. **128**:11, 3540–2 (2006).

Doran D., Abul-Haija Y.M., Cronin L. Emergence of function and selection from recursively programmed polymerisation reactions in mineral environments. *Angew. Chem. Int. Ed. Engl.* **58**:33, 11253–6 (2019).

Dose K. Hypotheses on the appearance of life on Earth (review). *Adv. Space Res*. **6**:12, 181–6 (1986).

Drevin I., Larsson L., Eriksson I., Johansson B.L. Chemical properties of and solute-support interactions with the filtration medium Superdex 75 prep grade. *J. Chromatogr. A*. **514**, 137–46 (1990).

Drobot B., Iglesias-Artola J.M., Le Vay K., Mayr V., Kar M., Kreysing M., Mutschler H., Tang T.D. Compartmentalised RNA catalysis in membrane-free coacervate protocells. *Nat. Commun*. **9**:1, 3643 (2018).

Duong-Ly K.C., Gabelli S.B. Salting out of proteins using ammonium sulfate precipitation. *Methods Enzymol*. **541**, 85−94 (2014).

Eberhardt E.S., Raines R.T. Amide-amide and amide-water hydrogen bonds: implications for protein folding and stability. *J. Am. Chem. Soc*. **116**:5, 2149–50 (1994).

Ellis J.J., Broom M., Jones S. Protein-RNA interactions: structural analysis and functional classes. *Proteins*. **66**:4, 903–11 (2007).

Elsila J.E., Glavin D.P., Dworkin J.P. Cometary glycine detected in samples returned by Stardust. *Meteorit. Planet. Sci*. **44**:9, 1323–30 (2009).

Engel M.H., Nagy B. Distribution and enantiomeric composition of amino-acids in the Murchison meteorite. *Nature*. **296**, 837–40 (1982).

Engel M.H., Macko S.A. Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature.* **389**, 265–8 (1997).

Epstein C.J. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature*. **210**:5031, 25–8 (1966).

Epstein S., Krishnamurthy R.V., Cronin J.R., Pizzarello S., Yuen G.U. Unusual stable isotope ratios in amino acid and carboxylic acid extracts from the Murchison meteorite. *Nature*. **326**:6112, 477–9 (1987).

Eriksson A.E., Baase W.A., Zhang X.J., Heinz D.W., Blaber M., Baldwin E.P., Matthews B.W. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*. **255**:5041, 178−83 (1992).

Fändrich M., Meinhardt J., Grigorieff N. Structural polymorphism of Alzheimer Abeta and other amyloid fibrils. *Prion*. **3**:2, 89–93 (2009).

Ferus M., Pietrucci F., Saitta A.M., Knížek A., Kubelík P., Ivanek O., Shestivska V., Civiš S. Formation of nucleobases in a Miller-Urey reducing atmosphere. *Proc. Natl. Acad. Sci. USA*. **114**:17, 4306–11 (2017).

Fichtner M., Voigt K., Schuster S. The tip and hidden part of the iceberg: proteinogenic and non-proteinogenic aliphatic amino acids. *Biochim. Biophys. Acta Gen. Subj*. **1861**:1, 3258–69 (2017).

Finking R., Marahiel M.A. Biosynthesis of nonribosomal peptides. *Annu. Rev. Microbiol*. **58**, 453–88 (2004).

Fioroni M., Diaz M.D., Burger K., Berger S. Solvation phenomena of a tetrapeptide in water/trifluoroethanol and water/ethanol mixtures: a diffusion NMR, intermolecular NOE, and molecular dynamics study. *J. Am. Chem. Soc*. **124**:26, 7737–44 (2002).

Fitch C.A., Platzer G., Okon M., Garcia-Moreno B.E., McIntosh LP. Arginine: its pKa value revisited. *Protein Sci*. **24**:5, 752−61 (2015).

Fontana A., Polverino de Laureto P., Spolaore B., Frare E., Picotti P., Zambonin M. Probing protein structure by limited proteolysis. *Acta Biochim. Pol*. **51**:2, 299−321 (2004).

Forsythe J.G., Yu S.S., Mamajanov I., Grover M.A., Krishnamurthy R., Fernández F.M., Hud N.V. Ester-mediated amide bond formation driven by wet-dry cycles: A possible path to polypeptides on the prebiotic Earth. *Angew. Chem. Int. Ed. Engl*. **54**:34, 9871–5 (2015).

Fournier G.P., Gogarten J.P. Signature of a primitive genetic code in ancient protein lineages. *J. Mol. Evol*. **65**:4, 425–36 (2007).

Fournier G.P., Andam C.P., Alm E.J., Gogarten J.P. Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. *Orig. Life Evol. Biosph*. **41**:6, 621–32 (2011).

Fournier G.P., Alm E.J. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *J. Mol. Evol*. **80**:3–4, 171–85 (2015).

Fowler D.M., Koulov A.V., Alory-Jost C., Marks M.S., Balch W.E., Kelly J.W. Functional amyloid formation within mammalian tissue. *PLoS Biol*. **4**:1, e6 (2006).

Fowler D.M., Koulov A.V., Balch W.E., Kelly J.W. Functional amyloid - from bacteria to humans. *Trends Biochem. Sci*. **32**:5, 217–24 (2007).

Fox G.E. Origin and evolution of the ribosome. *Cold Spring Harbor. Perspect. Biol*. **2**:9, a003483 (2010).

Fox S.W., Windsor C.R. Synthesis of amino acids by the heating of formaldehyde and ammonia. *Science*. **170**:3961, 984–6 (1970).

Francklyn C., Schimmel P. Aminoacylation of RNA minihelices with alanine. *Nature*. **337**:6206, 478–81 (1989).

Francklyn C. tRNA synthetase paralogs: evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to de novo synthesis. *Proc. Natl. Acad. Sci. USA*. **100**:17, 9650–2 (2003).

Frenkel-Pinter M., Haynes J.W., C M., Petrov A.S., Burcar B.T., Krishnamurthy R., Hud N.V., Leman L.J., Williams L.D. Selective incorporation of proteinaceous over nonproteinaceous cationic amino acids in model prebiotic oligomerization reactions. *Proc. Natl. Acad. Sci. USA*. **116**:33, 16338–46 (2019).

Frenkel-Pinter M., Haynes J.W., Mohyeldin A.M., C M., Sargon A.B., Petrov A.S., Krishnamurthy R., Hud N.V., Williams L.D., Leman L.J. Mutually stabilizing interactions between proto-peptides and RNA. *Nat. Commun*. **11**:1, 3137 (2020).

Frenkel-Pinter M., Bouza M., Fernández F.M., Leman L.J., Williams L.D., Hud N.V., Guzman-Martinez A. Thioesters provide a plausible prebiotic path to proto-peptides. *Nat. Commun*. **13**:1, 2569 (2022).

Fried S.D., Fujishima K., Makarov M., Cherepashuk I., Hlouchova K. Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc. Interface*. **19**:187, 20210641 (2022).

Friedmann M.P., Torbeev V., Zelenay V., Sobol A., Greenwald J., Riek R. Towards prebiotic catalytic amyloids using high throughput screening. *PLoS One*. **10**:12, e0143948 (2015).

Friedmann N., Miller S.L. Phenylalanine and tyrosine synthesis under primitive earth conditions. *Science*. **166**:3906, 766–7 (1969).

Fuchida S., Masuda H., Shinoda K. Peptide formation mechanism on montmorillonite under thermal conditions. *Orig. Life Evol. Biosph*. **44**:1, 13–28 (2014).

Fujishima K., Sugahara J., Kikuta K., Hirano R., Sato A., Tomita M., Kanai A. Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proc. Natl. Acad. Sci. USA*. **106**:8, 2683–7 (2009).

Fujishima K., Wang K.M., Palmer J.A., Abe N., Nakahigashi K., Endy D., Rothschild L.J. Reconstruction of cysteine biosynthesis using engineered cysteine-free enzymes. *Sci. Rep*. **8**:1, 1776 (2018).

Gan D., Guo Y., Lei X., Zhang M., Fu S., Ying J., Zhao Y. Urea-mediated warm ponds: Prebiotic formation of carbamoyl amino acids on the primordial Earth. *Earth Planet. Sci. Lett*. **607**:118072 (2023).

Geiger T., Clarke S. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *J. Biol. Chem*. **262**:2, 785–94 (1987).

Giacobelli V.G., Fujishima K., Lepšík M., Tretyachenko V., Kadavá T., Makarov M., Bednárová L., Novák P., Hlouchová K. In vitro evolution reveals noncationic protein-RNA interaction mediated by metal ions. *Mol. Biol. Evol*. **39**:3, msac032 (2022).

Gibard C., Bhowmik S., Karki M., Kim E.K., Krishnamurthy R. Phosphorylation, oligomerization and self-assembly in water under potential prebiotic conditions. *Nat. Chem*. **10**:2, 212–7 (2018).

Gilbert W. Origin of life: The RNA world. *Nature.* **319**, 618 (1986).

Gilis D., Massar S., Cerf N.J., Rooman M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol*. **2**:11, research0049 (2001).

Glavin D.P., Alexander C.M.O'D., Aponte J.C., Dworkin J.P., Elsila J.E., Yabuta H. The origin and evolution of organic matter in carbonaceous chondrites and links to their parent bodies. In: *Primitive meteorites and asteroids*. Abreu N. (Ed.). Amsterdam: Elsevier. pp. 205–271 (2018).

Goodarzi H., Nejad H.A., Torabi N. On the optimality of the genetic code, with the consideration of termination codons. *Biosystems*. **77**:1–3, 163–73 (2004).

Gorlero M., Wieczorek R., Adamala K., Giorgi A., Schininà M.E., Stano P., Luisi P.L. Ser-His catalyses the formation of peptides and PNAs. *FEBS Lett*. **583**:1, 153–6 (2009).

Gospodinov A., Kunnev D. Universal codons with enrichment from GC to AU nucleotide composition reveal a chronological assignment from early to late along with LUCA formation. *Life (Basel)*. **10**:6, 81 (2020).

Granold M., Hajieva P., Toşa M.I., Irimie F.D., Moosmann B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl. Acad. Sci. USA*. **115**:1, 41–6 (2018).

Greenwald J., Friedmann M.P., Riek R. Amyloid aggregates arise from amino acid condensations under prebiotic conditions. *Angew. Chem. Int. Ed. Engl*. **55**:38, 11609−13 (2016).

Griffith R.W. A specific scenario for the origin of life and the genetic code based on peptide/oligonucleotide interdependence. *Orig. Life Evol. Biosph*. **39**:6, 517–31 (2009).

Guelin M., Cernicharo J. Organic molecules in the interstellar space: latest advances. *Front. Astron. Space Sci*. **9**, 787567 (2022).

Halevy I., Bachan A. The geologic history of seawater pH. *Science*. **355**:6329, 1069–71 (2017).

Harada K., Fox S.W. The thermal condensation of glutamic acid and glycine to linear peptides. *J. Am. Chem. Soc*. **80**:11, 2694–7 (1958).

Harada K. Thermal homopolymerization of lysine and copolymerization with neutral and acidic amino acids. *Bull. Chem. Soc. Jpn*. **32**:9, 1007–8 (1959).

Harada K. Formation of amino-acids by thermal decomposition of formamide – oligomerization of hydrogen cyanide. *Nature*. **214**, 479–80 (1967).

Harish A., Caetano-Anollés G. Ribosomal history reveals origins of modern protein synthesis. *PLoS One*. **7**:3, e32776 (2012).

Harrington W.F., Schellman J.A. Evidence for the instability of hydrogen-bonded peptide structures in water, based on studies of ribonuclease and oxidized ribonuclease. *C. R. Trav. Lab. Carlsberg Chi*m. 30:3, 21–43 (1956).

Hennet R.J., Holm N.G., Engel M.H. Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: a perpetual phenomenon? *Naturwissenschaften*. **79**:8, 361–5 (1992).

Herschlag D., Khosla M., Tsuchihashi Z., Karpel R. L. An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. *EMBO J*. **13**:12, 2913–24 (1994).

Higgs P.G., Pudritz R.E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*. **9**:5, 483–90 (2009).

Hirose Y., Ohmuro K., Saigoh M., Nakayama T., Yamagata Y. Synthesis of biomolecules from N_2, CO, and H_2O by electric discharge. *Orig. Life Evol. Biosph.* **20**, 471–81 (1990).

Holm N.G., Neubeck A. Reduction of nitrogen compounds in oceanic basement and its implications for HCN formation and abiotic organic synthesis. *Geochem. Trans.* **10**, 9 (2009).

Hou Q., Bourgeas R., Pucci F., Rooman M. Computational analysis of the amino acid interactions that promote or decrease protein solubility. *Sci. Rep*. **8**:1, 14661 (2018).

Hsiao C., Mohan S., Kalahar B.K., Williams L.D. Peeling the onion: ribosomes are ancient molecular fossils. *Mol. Biol. Evol*. **26**:11, 2415–25 (2009).

Hsiao C., Lenz T.K., Peters J.K., Fang P.Y., Schneider D.M., Anderson E.J., Preeprem T., Bowman J.C., O'Neill E.B., Lie L., Athavale S.S., Gossett J.J., Trippe C., Murray J., Petrov A.S., Wartell R.M., Harvey S.C., Hud N.V., Williams L.D. Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucleic Acids Res*. **41**:5, 3373–85 (2013).

Hudson R.L. Interstellar tryptophan revisited. *Mon. Not. R. Astron. Soc*. **526**:3, 4051–53 (2023).

Hulshof J., Ponnamperuma C. Prebiotic condensation reactions in an aqueous medium: a review of condensing agents. *Orig. Life*. **7**:3, 197–24 (1976).

Hyman A.A., Weber C.A., Jülicher F. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol*. **30**, 39–58 (2014).

Iglesias-Groth S. A search for tryptophan in the gas of the IC 348 star cluster of the Perseus molecular cloud. *Mon. Not. R. Astron. Soc*. **523**:2, 2876–86 (2023).

Illangasekare M., Yarus M. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA*. **5**:11, 1482–9 (1999).

Imai E., Honda H., Hatori K., Brack A., Matsuno K. Elongation of oligopeptides in a simulated submarine hydrothermal system. *Science*. **283**:5403, 831–3 (1999).

Inoue K., Baden N., Terazima M. Diffusion coefficient and the secondary structure of poly-L-glutamic acid in aqueous solution. *J. Phys. Chem. B*. **109**:47, 22623–8 (2005).

Ito M., Gupta L.P., Masuda H., Kawahata H. Thermal stability of amino acids in seafloor sediment in aqueous solution at high temperature. *Org. Geochem*. **37**:2, 177–88 (2006).

Jash B., Tremmel P., Jovanovic D., Richert C. Single nucleotide translation without ribosomes. *Nat. Chem*. **13**:8, 751–7 (2021).

Johnson D.B., Wang L. Imprints of the genetic code in the ribosome. *Proc. Natl. Acad. Sci. USA*. **107**:18, 8298–303 (2010).

Jukes T.H., Osawa S. The genetic code in mitochondria and chloroplasts. *Experientia*. **46**:11–12, 1117–26 (1990).

Kashiwagi N., Furuta H., Ikawa Y. Primitive templated catalysis of a peptide ligation by self-folding RNAs. *Nucleic Acids Res*. **37**:8, 2574–83 (2009).

Kasting J.F., Zahnle K.J., Pinto J.P., Young A.T. Sulphur, ultraviolet radiation, and the early evolution of life. *Orig. Life Evol. Biosph*. **19**, 95–108 (1989).

Kasting J.F. Earth's early atmosphere. *Science*. **259**:5097, 920−6 (1993).

Kawabata M., Kawashima K., Mutsuro-Aoki H., Ando T., Umehara T., Tamura K. Peptide bond formation between aminoacyl-minihelices by a scaffold derived from the peptidyl transferase center. *Life (Basel)*. **12**:4, 573 (2022).

Keating C.D. Aqueous phase separation as a possible route to compartmentalization of biological molecules. *Acc. Chem. Res*. **45**:12, 2114–24 (2012).

Kelly S.M., Jess T.J., Price N.C. How to study proteins by circular dichroism. *Biochim. Biophys. Acta*. **1751**:2, 119–39 (2005).

Kempe S., Kazmierczak J. Biogenesis and early life on Earth and Europa: favored by an alkaline ocean? *Astrobiology*. **2**:1, 123–30 (2002).

Kennedy R.M. Hydrophobic chromatography. *Methods Enzymol*. **182**, 339−43 (1990).

Kerridge JF. Carbon, hydrogen and nitrogen in carbonaceous chondrites: abundances and isotopic compositions in bulk samples. *Geochim. Cosmochim. Acta*. **49**, 1707–14 (1985).

Kherb J., Flores S.C., Cremer PS. Role of carboxylate side chains in the cation Hofmeister series. *J. Phys. Chem. B*. **116**:25, 7389−97 (2012).

Kim S., Huang J., Lee Y., Dutta S., Yoo H.Y., Jung Y.M., Jho Y., Zeng H., Hwang D.S. Complexation and coacervation of like-charged polyelectrolytes inspired by mussels. *Proc. Natl. Acad. Sci. USA*. **113**:7, E847–53 (2016).

Kirschning A. On the evolutionary history of the twenty encoded amino acids. *Chemistry*. **28**:55, e202201419 (2022).

Kisumi M., Sugiura M., Chibata I. Biosynthesis of norvaline, norleucine, and homoisoleucine in Serratia marcescens. *J. Biochem*. **80**:2, 333–9 (1976).

Kitadai N., Maruyama S. Origins of building blocks of life: A review. *Geosci. Front*. **9**:4, 1117–53 (2018).

Klein D.J., Moore P.B., Steitz T.A. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol*. **340**:1, 141–77 (2004a).

Klein D.J., Moore P.B., Steitz T.A. The contribution of metal ions to the structural stability of the large ribosomal subunit. *RNA*. **10**:9, 1366–79 (2004b).

Knight R.D., Freeland S.J., Landweber L.F. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet*. **2**:1, 49–58 (2001).

Kobylarz M.J., Grigg J.C., Takayama S.J., Rai D.K., Heinrichs D.E., Murphy M.E. Synthesis of L-2,3-diaminopropionic acid, a siderophore and antibiotic precursor. *Chem. Biol*. 21:3, 379–88 (2014).

Koga T., Naraoka H. A new family of extraterrestrial amino acids in the Murchison meteorite. *Sci. Rep*. **7**:1, 636 (2017).

Kolodny R., Nepomnyachiy S., Tawfik D.S., Ben-Tal N. Bridging themes: short protein segments found in different architectures. *Mol. Biol. Evol*. **38**:6, 2191–208 (2021).

Konn C., Charlou J.L., Holm N.G., Mousis O. The production of methane, hydrogen, and organic compounds in ultramafic-hosted hydrothermal vents of the Mid-Atlantic Ridge. *Astrobiology*. **15**:5, 381–99 (2015).

Koonin E.V., Novozhilov A.S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*. **61**:2, 99–111 (2009).

Kovacs J.M., Mant C.T., Hodges R.S. Determination of intrinsic hydrophilicity/hydrophobicity of amino acid side chains in peptides in the absence of nearest-neighbor or conformational effects. *Biopolymers*. **84**:3, 283−97 (2006).

Krasnokutski S.A., Chuang K.J., Jäger C., Ueberschaar N., Henning T. A pathway to peptides in space through the condensation of atomic carbon. *Nat. Astron*. **6**, 381–6 (2022).

Krzyzaniak A., Barciszewski J., Sałański P., Jurczak J. The non-enzymatic specific amino-acylation of transfer RNA at high pressure. *Int. J. Biol. Macromol*. **16**:3, 153–8 (1994).

Kubota S., Fasman G. Conformation and optical properties of poly(L-valine) in aqueous solution. "A single extended beta chain". *J Am Chem Soc*. **96**:14, 4684–6 (1974).

Kumachi S., Husimi Y., Nemoto N. An RNA binding peptide consisting of four types of amino acid by in vitro selection using cDNA display. *ACS Omega*. **1**:1, 52–7 (2016).

Kwiatkowski W, Bomba R, Afanasyev P, Boehringer D, Riek R, Greenwald J. Prebiotic peptide synthesis and spontaneous amyloid formation inside a proto-cellular compartment. *Angew. Chem. Int. Ed. Engl*. **60**:10, 5561–8 (2021).

la Cour R., Jørgensen H., Schjoerring J.K. Improvement of tryptophan analysis by liquid chromatography-single quadrupole mass spectrometry through the evaluation of multiple parameters. *Front. Chem*. **19**:7, 797 (2019).

La Verde V., Dominici P., Astegno A. Determination of hydrodynamic radius of proteins by size exclusion chromatography. *Bio Protoc*. **7**:8, e2230 (2017).

Lacey J.C., Weber A.L., White W.E. Jr. A model for the coevolution of the genetic code and the process of protein synthesis: Review and assessment. *Orig. Life*. **6**:1–2, 273–83 (1975).

Lahav N., White D., Chang S. Peptide formation in the prebiotic era: thermal condensation of glycine in fluctuating clay environments. *Science*. **201**:4350, 67–9 (1978).

Lal A.K. Origin of Life. *Astrophys. Space Sci.* **317**, 267–78 (2008).

Lang S.Q., Früh-Green G.L., Bernasconi S.M., Butterfield D.A. Sources of organic nitrogen at the serpentinite-hosted Lost City hydrothermal field. *Geobiology*. **11**:2, 154–69 (2013).

Laurents D.V., Huyghues-Despointes B.M., Bruix M., Thurlkill R.L., Schell D., Newsom S., Grimsley G.R., Shaw K.L., Treviño S., Rico M., Briggs J.M., Antosiewicz J.M., Scholtz J.M., Pace C.N. Charge-charge interactions are key determinants of the pK values of ionizable groups in ribonuclease Sa (pI=3.5) and a basic variant (pI=10.2). *J. Mol. Biol*. **325**:5, 1077−92 (2003).

Lawless G.D., Boynton C.D. Thermal synthesis of amino acids from a simulated primitive atmosphere. *Nature*. **243**:5407, 404−5 (1973).

Lawless J.G., Peterson E. Amino acids in carbonaceous chondrites. *Orig. Life*. **6**:1–2, 3–8 (1975).

Le Vay K., Song E.Y., Ghosh B., Tang T.D., Mutschler H. Enhanced ribozyme-catalyzed recombination and oligonucleotide assembly in peptide-RNA condensates. *Angew. Chem. Int. Ed. Engl*. **60**:50, 26096–104 (2021).

Lecompte O., Ripp R., Thierry J.C., Moras D., Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res*. **30**:24, 5382–90 (2002).

Lee D.H., Granja J.R., Martinez J.A., Severin K., Ghadiri M.R. A self-replicating peptide. *Nature*. **382**:6591, 525–8 (1996).

Lee M.R., Daly L., Floyd C., Martin P.E. CM carbonaceous chondrite falls and their terrestrial alteration. *Meteorit. Planet. Sci*. **56**:1, 34–48 (2021).

Lee N., Suga H. A minihelix-loop RNA acts as a trans-aminoacylation catalyst. *RNA*. **7**:7, 1043–51 (2001).

Leman L., Orgel L., Ghadiri M.R. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science*. **306**:5694, 283–6 (2004).

Leman L.J., Orgel L.E., Ghadiri M.R. Amino acid dependent formation of phosphate anhydrides in water mediated by carbonyl sulfide. *J. Am. Chem. Soc*. **128**:1, 20–1 (2006).

Leman L.J., Huang Z.Z, Ghadiri MR. Peptide bond formation in water mediated by carbon disulfide. *Astrobiology*. **15**:9,709–16 (2015).

Levine J.S., Augustsson T.R., Natarajan M. The prebiological paleoatmosphere: stability and composition. *Orig. Life*. **12**:3, 245–59 (1982).

Li P., Holliger P., Tagami S. Hydrophobic-cationic peptides modulate RNA polymerase ribozyme activity by accretion. *Nat. Commun.*, **13**, 3050 (2022).

Li Y., Zhao Y., Hatfield S., Wan R., Zhu Q., Li X., McMills M., Ma Y., Li J., Brown K.L., He C., Liu F., Chen X. Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. *Bioorg. Med. Chem*. **8**:12, 2675–80 (2000).

Lohrmann R., Orgel L.E. Urea-inorganic phosphate mixtures as prebiotic phosphorylating agents. *Science*. **171**:3970, 490–4 (1971).

Longo L.M., Blaber M. Protein design at the interface of the pre-biotic and biotic worlds. *Arch. Biochem. Biophys*. **526**:1, 16–21 (2012).

Longo L.M., Lee J., Blaber M. Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl. Acad. Sci. USA*. **110**:6, 2135−9 (2013).

Longo L.M., Tenorio C.A., Kumru O.S., Middaugh C.R., Blaber M. A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding. *Protein Sci*. **24**:1, 27−37 (2015).

Longo L.M., Despotović D., Weil-Ktorza O., Walker M.J., Jabłońska J., Fridmann-Sirkis Y., Varani G., Metanis N., Tawfik D.S. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. USA*. **117**:27, 15731–9 (2020).

Lupas A.N., Alva V. Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol*. **198**:2, 74–81 (2017).

Lyu P.C., Sherman J.C., Chen A., Kallenbach N.R. Alpha-helix stabilization by natural and unnatural amino acids with alkyl side chains. *Proc. Natl. Acad. Sci. USA*. **88**:12, 5317–20 (1991).

Maguire O.R., Smokers I.B.A., Huck W.T.S. A physicochemical orthophosphate cycle via a kinetically stable thermodynamically activated intermediate enables mild prebiotic phosphorylations. *Nat. Commun*. **12**, 5517 (2021).

Maguire O.R., Smokers I.B.A., Oosterom B.G., Zheliezniak A., Huck W.T.S. A prebiotic precursor to life's phosphate transfer system with an ATP analog and histidyl peptide organocatalysts. *J. Am. Chem. Soc*. **146**:11, 7839–49 (2024).

Makam P., Yamijala S.S.R.K.C., Tao K., Shimon L.J.W., Eisenberg D.S., Sawaya M.R., Wong B.M., Gazit E. Non-proteinaceous hydrolase comprised of a phenylalanine metallo-supramolecular amyloid-like structure. *Nat. Catal*. **2**, 977–85 (2019).

Makarov M., Meng J., Tretyachenko V., Srb P., Březinová A., Giacobeli V.G., Bednárová L., Vondrášek J., Dunker A.K., Hlouchová K. Enzyme catalysis prior to aromatic residues: reverse engineering of a dephospho-CoA kinase. *Protein Sci*. **30**:5, 1022−34 (2021).

Makarov M., Sanchez Rocha A.C., Krystufek R., Cherepashuk I., Dzmitruk V., Charnavets T., Faustino A.M., Lebl M., Fujishima K., Fried S.D., Hlouchova K. Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc*. **145**:9**,** 5320−9 (2023).

Makhlynets O.V., Gosavi P.M., Korendovych I.V. Short self-assembling peptides are able to bind to copper and activate oxygen. *Angew. Chem. Int. Ed. Engl*. **55**:31, 9017–20 (2016).

Marquez R., Smit S., Knight R. Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biol.* **6**:11, R91 (2005).

Marshall W.L. Hydrothermal synthesis of amino acids. *Geochim. Cosmochim. Acta*. **58**:9, 2099–106 (1994).

Martin R.B. Free energies and equilibria of peptide bond hydrolysis and formation. *Biopolymers*. **45**, 351–53 (1998).

Martinez-Rodriguez L, Erdogan O, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, Weinreb V., Collier M., Chandrasekaran S.N., Ambroggio X., Kuhlman B., Carter C.W.Jr. Functional class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. *J. Biol. Chem*. **290**:32, 19710–25 (2015).

Maruyama S., Ikoma M., Genda H., Hirose K., Yokoyama T., Santosh M. The naked planet Earth: most essential pre-requisite for the origin and evolution of life. *Geosci. Front*. **4**, 141–65 (2013).

Mason P.E., Neilson G.W., Dempsey C.E., Barnes A.C., Cruickshank J.M. The hydration structure of guanidinium and thiocyanate ions: implications for protein stability in aqueous solution. *Proc. Natl. Acad. Sci. USA*. **100**:8, 4557−61 (2003).

Maury C.P.J. Amyloid and the origin of life: self-replicating catalytic amyloids as prebiotic informational and protometabolic entities. *Cell. Mol. Life. Sci*. **75**:9, 1499–507 (2018).

McCollom T.M., Seewald J.S. Abiotic synthesis of organic compounds in deep-sea hydrothermal environments. *Chem. Rev*. **107**, 382–401 (2007).

McDonald G.D., Storrie-Lombardi M.C. Biochemical constraints in a protobiotic earth devoid of basic amino acids: the "BAA(-) world". *Astrobiology*. **10**:10, 989–1000 (2010).

Mears J.A., Cannone J.J., Stagg S.M., Gutell R.R., Agrawal R.K., Harvey S.C. Modeling a minimal ribosome based on comparative sequence analysis. *J. Mol. Biol*. **321**:2, 215–34 (2002).

Meersman F., Dobson C.M. Probing the pressure-temperature stability of amyloid fibrils provides new insights into their molecular properties. *Biochim. Biophys. Acta*. **1764**:3, 452–60 (2006).

Mehta A.K., Lu K., Childers W.S., Liang Y., Dublin S.N., Dong J., Snyder J.P., Pingali S.V., Thiyagarajan P., Lynn D.G. Facial symmetry in protein self-assembly. *J. Am. Chem. Soc*. **130**:30, 9829–35 (2008).

Meierhenrich U.J., Muñoz Caro G.M., Bredehöft J.H., Jessberger E.K, Thiemann W.H. Identification of diamino acids in the Murchison meteorite. *Proc Natl. Acad. Sci. USA*. **101**:25, 9182–6 (2004).

Meisl G., Xu C.K., Taylor J.D., Michaels T.C.T., Levin A., Otzen D., Klenerman D., Matthews S., Linse S., Andreasen M., Knowles T.P.J. Uncovering the universality of self-replication in protein aggregation and its link to disease. *Sci. Adv*. **8**:32, eabn6831 (2022).

Meng K., Chung C.Z., Söll D., Krahn N. Unconventional genetic code systems in archaea. *Front. Microbiol*. **13**, 1007832 (2022).

Menor Salván C., Bouza M., Fialho D.M., Burcar B.T., Fernández F.M., Hud N.V. Prebiotic origin of pre-RNA building blocks in a urea "warm little pond" scenario. *Chembiochem*. **21**:24, 3504–10 (2020).

Meuzelaar H., Vreede J., Woutersen S. Influence of Glu/Arg, Asp/Arg, and Glu/Lys salt bridges on α-helical stability and folding kinetics. *Biophys J*. **110**:11, 2328–41 (2016).

Mével C. Serpentinization of abyssal peridotites at mid-ocean ridges. *C. R. Geosci*. **335**:10–11, 825–52 (2003).

Miller S.L. A production of amino acids under possible primitive earth conditions. *Science*. **117**:3046, 528−9 (1953).

Miller S.L. Production of some organic compounds under possible primitive Earth conditions. *J. Am. Chem. Soc*. **77**, 2351–61 (1955).

Miller S.L. The mechanism of synthesis of amino acids by electric discharges. *Biochim. Biophys. Acta.* **23**:3, 480–89 (1957).

Miller S.L., Urey H.C. Organic compound synthesis on the primitive earth. *Science*. **130**:3370, 245−51 (1959).

Mirtič A., Grdadolnik J. The structure of poly-L-lysine in different solvents. *Biophys. Chem*. **175–176**, 47–53 (2013).

Mita H., Nomoto S., Terasaki M., Shimoyama A., Yamamoto Y. Prebiotic formation of polyamino acids in molten urea. *Int. J. Astrobiol*. **4**:2, 145–54 (2005).

Miyakawa S., Yamanashi H., Kobayashi K., Cleaves H.J., Miller S.L. Prebiotic synthesis from CO atmospheres: implications for the origins of life. *Proc. Natl. Acad. Sci. USA*. **99**:23, 14628–31 (2002).

Modica P., Martins Z., Meinert C., Zanda B., d'Hendecourt L.L.S. The amino acid distribution in laboratory analogs of extraterrestrial organic matter: a comparison to CM chondrites. *Astrophys. J*. **865**, 41 (2018).

Monasterio O., Nova E., Diaz-Espinoza R. Development of a novel catalytic amyloid displaying a metal-dependent ATPase-like activity. *Biochem. Biophys. Res. Commun*. **482**:4, 1194–200 (2017).

Morga M., Batys P., Kosior D., Bonarek P., Adamczyk Z. Poly-L-arginine molecule properties in simple electrolytes: molecular dynamic modeling and experiments. *Int. J. Environ. Res. Public Health*. **19**:6, 3588 (2022)

Müller F., Escobar L., Xu F., Węgrzyn E., Nainytė M., Amatov T., Chan C.Y., Pichler A., Carell T. A prebiotically plausible scenario of an RNA-peptide world. *Nature*. **605**:7909, 279–84 (2022).

Muñoz Caro G.M., Meierhenrich U.J., Schutte W.A., Barbier B., Arcones Segovia A., Rosenbauer H., Thiemann W.H., Brack A., Greenberg J.M. Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature*. **416**:6879, 403–6 (2002).

Murakami H., Saito H., Suga H. A versatile tRNA aminoacylation catalyst based on RNA. *Chem. Biol.* **10**, 655–62 (2003).

Murphy L.R., Wallqvist A., Levy R.M. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng*. **13**:3, 149−52 (2000).

Musafia B., Buchner V., Arad D. Complex salt bridges in proteins: statistical analysis of structure and function. *J. Mol. Biol*. **254**:4, 761–70 (1995).

Nagayama M., Takaoka O., Inomata K., Yamagata Y. Diketopiperazine-mediated peptide formation in aqueous solution. *Orig. Life Evol. Biosph*. **20**:3–4, 249–57 (1990).

Nakamura E., Kobayashi K., Tanaka R., Kunihiro T., Kitagawa H., Potiszil C., Ota T., Sakaguchi C., Yamanaka M., Ratnayake D.M., Tripathi H., Kumar R., Avramescu M.L., Tsuchida H., Yachi Y., Miura H., Abe M., Fukai R., Furuya S., Hatakeda K., Hayashi T., Hitomi Y., Kumagai K., Miyazaki A., Nakato A., Nishimura M., Okada T., Soejima H., Sugita S., Suzuki A., Usui T., Yada T., Yamamoto D., Yogata K., Yoshitake M., Arakawa M., Fujii A., Hayakawa M., Hirata N., Hirata N., Honda R., Honda C., Hosoda S., Iijima Y.I., Ikeda H., Ishiguro M., Ishihara Y., Iwata T., Kawahara K., Kikuchi S., Kitazato K., Matsumoto K., Matsuoka M., Michikami T., Mimasu Y., Miura A., Morota T., Nakazawa S., Namiki N., Noda H., Noguchi R., Ogawa N., Ogawa K., Okamoto C., Ono G., Ozaki M., Saiki T., Sakatani N., Sawada H., Senshu H., Shimaki Y., Shirai K., Takei Y., Takeuchi H., Tanaka S., Tatsumi E., Terui F., Tsukizaki R., Wada K., Yamada M., Yamada T., Yamamoto Y., Yano H., Yokota Y., Yoshihara K., Yoshikawa M., Yoshikawa K., Fujimoto M., Watanabe S.I., Tsuda Y. On the origin and evolution of the asteroid Ryugu: A comprehensive geochemical perspective. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci*. **98**:6, 227–82 (2022).

Namani T., Snyder S., Eagan J.M., Bevilacqua P.C., Wesdemiotis C., Sahai N. Amino acid specific nonenzymatic montmorillonite-promoted RNA polymerization. *ChemSystemsChem*. **3**, e2000060 (2021).

Naraoka H., Hashiguchi M., Okazaki R. Soluble sulfur-bearing organic compounds in carbonaceous meteorites: implications for chemical evolution in primitive asteroids. *ACS Earth Space Chem*. **7**:1, 41–8 (2023).

Nelson R., Sawaya M.R., Balbirnie M., Madsen A.Ø., Riekel C., Grothe R., Eisenberg D. Structure of the cross-beta spine of amyloid-like fibrils. *Nature*. **435**:7043, 773–8 (2005).

Newton M.S., Morrone D.J., Lee K.H., Seelig B. Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries. *Chembiochem*. **20**:6, 846−56 (2019).

Nigam S.N., Ressler C. Biosynthesis of 2,4-diaminobutyric acid from L-[3H]homoserine and DL-[1-14C]aspartic acid in Lathyrus sylvestris W. *Biochemistry*. **5**:11, 3426–31 (1966).

Nissen P., Hansen J., Ban N., Moore P.B., Steitz T.A. The structural basis of ribosome activity in peptide bond synthesis. *Science*. **289**:5481, 920−30 (2000).

Nowick J.S., Insaf S. The propensities of amino acids to form parallel β-sheets. *J. Am. Chem. Soc*. **119**: 45, 10903–8 (1997).

Oba Y., Takano Y., Watanabe N., Kouchi A. Deuterium fractionation during amino acid formation by photolysis of interstellar ice analogs containing deuterated methanol. *Astrophys. J. Lett.* **827**, 18–25 (2016).

Oba Y., Takano Y., Naraoka H., Watanabe N., Kouchi A. Nucleobase synthesis in interstellar ices. *Nat. Commun.* **10**:1, 4413 (2019).

Oba Y., Takano Y., Furukawa Y., Koga T., Glavin D.P., Dworkin J.P., Naraoka H. Identifying the wide diversity of extraterrestrial purine and pyrimidine nucleobases in carbonaceous meteorites. *Nat. Commun.* **13**:1, 2008 (2022).

Oba Y., Koga T., Takano Y., Ogawa N.O., Ohkouchi N., Sasaki K., Sato H., Glavin D.P., Dworkin J.P., Naraoka H., Tachibana S., Yurimoto H., Nakamura T., Noguchi T., Okazaki R., Yabuta H., Sakamoto K., Yada T., Nishimura M., Nakato A., Miyazaki A., Yogata K., Abe M., Okada T., Usui T., Yoshikawa M., Saiki T., Tanaka S., Terui F., Nakazawa S., Watanabe S.I., Tsuda Y. Uracil in the carbonaceous asteroid (162173) Ryugu. *Nat. Commun*. **14**:1, 1292 (2023).

O'Donoghue P, Sheppard K, Nureki O, Söll D. Rational design of an evolutionary precursor of glutaminyl-tRNA synthetase. *Proc. Natl. Acad. Sci. USA*. **108**:51, 20485–90 (2011).

Ogle J.M., Brodersen D.E., Clemons W.M.Jr, Tarry M.J., Carter A.P., Ramakrishnan V.. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*. **292**:5518, 897−902 (2001).

Osawa S., Jukes T.H. Codon reassignment (codon capture) in evolution. *J. Mol. Evol*. **28**:4, 271–8 (1989).

Pace C.N., Grimsley G.R., Scholtz J.M. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J. Biol. Chem*. **284**:20, 13285−9 (2009).

Padmanabhan S., Baldwin R.L. Straight-chain non-polar amino acids are good helix-formers in water. *J. Mol. Biol*. **219**:2, 135–7 (1991).

Padmanabhan S., York E.J., Stewart J.M., Baldwin R.L. Helix propensities of basic amino acids increase with the length of the side-chain. *J. Mol. Biol*. **257**:3, 726–34 (1996).

Paecht-Horowitz M. The possible role of clays in prebiotic peptide synthesis. *Orig. Life.* **5**:1, 173–87 (1974).

Pagano J.M., Clingman C.C., Ryder S.P. Quantitative approaches to monitor protein-nucleic acid interactions using fluorescent probes. *RNA*. **17**:1, 14–20 (2011).

Parker E.T., Cleaves H.J., Dworkin J.P., Glavin D.P., Callahan M., Aubrey A., Lazcano A., Bada J.L. Primordial synthesis of amines and amino acids in a 1958 Miller H_2S-rich spark discharge experiment. *Proc. Natl. Acad. Sci. USA*. **108**:14, 5526–31 (2011).

Parker E.T., Cleaves H.J., Burton A.S., Glavin D.P., Dworkin J.P., Zhou M., Bada J.L., Fernández F.M. Conducting Miller-Urey experiments. *J. Vis. Exp*. **83**, e51039 (2014).

Parker E.T., Chan Q.H., Glavin D.P., Dworkin J.P. Non-protein amino acids identified in carbon-rich Hayabusa particles. *Meteorit. Planet. Sci*., **57**:4, 776–93 (2022).

Parker E.T., McLain H.L., Glavin D.P., Dworkin J.P., Elsila J.E., Aponte J.C. Naraoka H., Takano Y., Tachibana S., Yabuta H., Yurimoto H., Sakamoto K., Yada T., Nishimura M., Nakato

A., Miyazaki A., Yogata K., Abe M., Okada T., Usui T., Yoshikawa M., Saiki T., Tanaka S., Nakazawa S., Tsuda Y., Terui F., Noguchi T., Okazaki R., Watanabe S., Nakamura T. Extraterrestrial amino acids and amines identified in asteroid Ryugu samples returned by the Hayabusa2 mission. *Geochim. Cosmochim. Acta* **347**, 42–57 (2023).

Pasek M.A., Kee T.P., Bryant D.E., Pavlov A.A., Lunine J.I. Production of potentially prebiotic condensed phosphates by phosphorus redox chemistry. *Angew. Chem. Int. Ed. Engl*. **47**:41, 7918–20 (2008).

Patel B.H., Percivalle C., Ritson D.J., Duffy C.D., Sutherland J.D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem*. **7**:4, 301–7 (2015).

Pereira J., Lupas A.N. The ancestral KH peptide at the root of a domain family with three different folds. *Bioinformatics*. **34**:23, 3961–5 (2018).

Perkins H.R., Cummins C.S. Chemical structure of bacterial cell walls. Ornithine and 2,4-diaminobutyric acid as components of the cell walls of plant pathogenic corynebacteria. *Nature*. **201**, 1105–7 (1964).

Petrov A.S., Bernier C.R., Hsiao C., Okafor C.D., Tannenbaum E., Stern J., Gaucher E., Schneider D., Hud N.V., Harvey S.C., Williams L.D. RNA-magnesium-protein interactions in large ribosomal subunit. *J. Phys. Chem. B*, **116**:28, 8113–20 (2012).

Petrov A.S., Gulen B., Norris A.M., Kovacs N.A., Bernier C.R., Lanier K.A., Fox G.E., Harvey S.C., Wartell R.M., Hud N.V., Williams L.D. History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. USA*. **112**:50, 15396−401 (2015).

Pizzarello S., Cronin J.R. Non-racemic amino acids in the Murray and Murchison meteorites. *Geochim. Cosmochim. Acta*. **64**:2, 329–38 (2000).

Poduška K., Katrukha G.S., Silaev A.B., Rudinger J. Amino acids and peptides. LII. Intramolecular aminolysis of amide bonds in derivatives of α,γ-diaminobutyric acid, α,β-diaminopropionic acid, and ornithine. *Collect. Czech. Chem. Commun*. **30**, 2410–33 (1965).

Polacek N., Mankin A.S. The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Crit. Rev. Biochem. Mol. Biol*. **40**, 285–311 (2005).

Potiszil C., Ota T., Yamanaka M., Sakaguchi C., Kobayashi K., Tanaka R., Kunihiro T., Kitagawa H., Abe M., Miyazaki A., Nakato A., Nakazawa S., Nishimura M., Okada T., Saiki T., Tanaka S., Terui F., Tsuda Y., Usui T., Watanabe S., Yada T., Yogata K., Yoshikawa M., Nakamura E. Insights into the formation and evolution of extraterrestrial amino acids from the asteroid Ryugu. *Nat. Commun*. **14**:1, 1482 (2023).

Poudyal R.R., Pir Cakmak F., Keating C.D., Bevilacqua P.C. Physical principles and extant biology reveal roles for RNA-containing membraneless compartments in origins of life chemistry. *Biochemistry*. **57**:17, 2509–19 (2018).

Poudyal R.R., Guth-Metzler R.M., Veenis A.J., Frankel E.A., Keating C.D., Bevilacqua P.C. Template-directed RNA polymerization and enhanced ribozyme catalysis inside membraneless compartments formed by coacervates. Nat. Commun. **10**:1, 490 (2019).

Prosdocimi F., Zamudio G.S., Palacios-Pérez M., Torres de Farias S., V José M. The ancient history of peptidyl transferase center formation as told by conservation and information analyses. *Life (Basel)*. **10**:8, 134 (2020).

Qing R., Hao S., Smorodina E., Jin D., Zalevsky A., Zhang S. Protein design: from the aspect of water solubility and stability. *Chem. Rev*. **122**:18, 14085−179 (2022).

Qiu F., Chen Y., Zhao X. Comparative studies on the self-assembling behaviors of cationic and catanionic surfactant-like peptides. *J. Colloid. Interface Sci*. **336**:2, 477–84 (2009).

Rabinowitz J., Flores J., Krebsbach R., Rogers G. Peptide formation in the presence of linear or cyclic polyphosphates. *Nature*. **224**, 795–6 (1969).

Raggi L., Bada J.L., and Lazcano, A. On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Phys. Chem. Chem. Phys*. **18**, 20028–32 (2016).

Randau L., Münch R., Hohn M.J., Jahn D., Söll D. Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature*. **433**:7025, 537–41 (2005).

Reches M., Gazit E. Casting metal nanowires within discrete self-assembled peptide nanotubes. *Science*. **300**:5619, 625–7 (2003).

Reiersen H, Rees AR. Trifluoroethanol may form a solvent matrix for assisted hydrophobic interactions between peptide side chains. *Protein Eng*. **13**:11, 739–43 (2000).

Remmert M., Biegert A., Linke D., Lupas A.N., Söding J. Evolution of outer membrane β-barrels from an ancestral ββ hairpin. *Mol. Biol. Evol*. **27**:6, 1348–58 (2010).

Reynolds J.A., Gilbert D.B., Tanford C. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Natl. Acad. Sci. USA*. **71**:8, 2925−7 (1974).

Ribas de Pouplana L. The evolution of aminoacyl-tRNA synthetases: From dawn to LUCA. *Enzymes*. **48**, 11–37 (2020).

Ribeiro A.J.M., Tyzack J.D., Borkakoti N., Holliday G.L., Thornton J.M. A global analysis of function and conservation of catalytic residues in enzymes. *J. Biol. Chem*. **295**:2, 314–24 (2020).

Riddle D.S., Santiago J.V., Bray-Hall S.T., Doshi N., Grantcharova V.P., Yi Q., Baker D. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**:10, 805−9 (1997).

Riek R., Eisenberg D.S. The activities of amyloids from a structural perspective. *Nature*. **539**:7628, 227–35 (2016).

Riggi V.S., Watson E.B., Steele A., Rogers K.L. Mineral-mediated oligoribonucleotide condensation: Broadening the scope of prebiotic possibilities on the early earth. *Life (Basel).* **13**:9, 1899 (2023).

Ring D., Wolman Y., Friedmann N., Miller S.L. Prebiotic synthesis of hydrophobic and protein amino acids. *Proc. Natl. Acad. Sci. USA*. **69**:3, 765–8 (1972).

Roberts S.J., Liu Z., Sutherland J.D. Potentially prebiotic synthesis of aminoacyl-RNA via a bridging phosphoramidate-ester intermediate. *J. Am. Chem. Soc*. **144**:9, 4254–9 (2022).

Robertson M.P., Miller S.L. An efficient prebiotic synthesis of cytosine and uracil. Nature. 375:6534, 772–4 (1995).

Robinson N.E. Protein deamidation. *Proc. Natl. Acad. Sci. USA*. **99**:8, 5283–8 (2002).

Roccatano D., Colombo G., Fioroni M., Mark A.E. Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: a molecular dynamics study. *Proc. Natl. Acad. Sci. USA*. **99**:19, 12179–84 (2002).

Rode B.M., Schwendinger M.G. Copper-catalyzed amino acid condensation in water – A simple possible way of prebiotic peptide formation. *Orig. Life Evol. Biosph.* **20**, 401–10 (1990).

Rodin S.N., Ohno S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. *Orig. Life Evol. Biosph*. **25**:6, 565–89 (1995).

Root-Bernstein M., Root-Bernstein R. The ribosome as a missing link in the evolution of life. *J. Theor. Biol*. **367**, 130–58 (2015).

Rosenheck K., Doty P. The far ultraviolet absorption spectra of polypeptide and protein solutions and their dependence on conformation. *Proc. Natl. Acad. Sci. USA*. **47**:11, 1775–85 (1961).

Rubinov B., Wagner N., Rapaport H., Ashkenasy G. Self-replicating amphiphilic beta-sheet peptides. *Angew. Chem. Int. Ed. Engl*. **48**:36, 6683–6 (2009).

Rufo C.M., Moroz Y.S., Moroz O.V., Stör J., Smith T.A., Hu X., DeGrado W.F., Korendovych I.V. Short peptides self-assemble to produce catalytic amyloids. *Nat. Chem*. **6**, 303–9 (2014).

Ruiz-Bermejo M., Menor-Salván C., Osuna-Esteban S., Veintemillas-Verdaguer S. The effects of ferrous and other ions on the abiotic formation of biomolecules using aqueous aerosols and spark discharges. *Orig. Life Evol. Biosph*. **37**:6, 507–21 (2007).

Ruiz-Bermejo M., Zorzano M.P., Osuna-Esteban S. Simple organics and biomonomers identified in HCN polymers: an overview. *Life (Basel)*. **3**:3, 421–48 (2013).

Sagan C., Khare B.N. Long-wavelength ultraviolet photoproduction of amino acids on the primitive Earth. *Science*. **173**:3995, 417−20 (1971).

Saghatelian A., Yokobayashi Y., Soltani K., Ghadiri M.R. A chiroselective peptide replicator. *Nature*. **409**:6822, 797–801 (2001).

Saier M.H.Jr. Understanding the genetic code. *J. Bacteriol*. **201**:15, e00091–19 (2019).

Santoso S., Hwang W., Hartman H., Zhang S. Self-assembly of surfactant-like peptides with variable glycine tails to form nanotubes and nanovesicles. *Nano Lett*. **2**:7, 687–91 (2002).

Sardesai N.Y., Green R., Schimmel P. Efficient 50S ribosome-catalyzed peptide bond synthesis with an aminoacyl minihelix. *Biochemistry*. **38**:37, 12080–8 (1999).

Schimmel P., Ribas de Pouplana L. Transfer RNA: from minihelix to genetic code. *Cell*. **81**:7, 983–6 (1995).

Schimpl A., Lemmon R.M., Calvin M. Cyanamide formation under primitive Earth conditions. *Science.* **147**:149–50 (1968).

Schlesinger G., Miller S.L. Prebiotic synthesis in atmospheres containing CH_4, CO, and CO_2. I. Amino acids. *J. Mol. Evol*. **19**:5, 376–82 (1983).

Schneider C., Becker S., Okamura H., Crisp A., Amatov T., Stadlmeier M., Carell T. Noncanonical RNA nucleosides as molecular fossils of an early earth-generation by prebiotic methylations and carbamoylations. *Angew. Chem. Int. Ed. Engl*. **57**:20, 5943–6 (2018).

Scholtz J.M., Qian H., Robbins V.H., Baldwin R.L. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry*. **32**:37, 9668–76 (1993).

Schomakers B.V., Jillings S.L., van Weeghel M., Vaz F.M., Salomons G.S., Janssens G.E., Houtkooper R.H. Ophthalmic acid is a glutathione regulating tripeptide. *FEBS J.* **291**:15, 3317–30 (2024).

Schultz D.W., Yarus M. Transfer RNA mutations and the malleability of the genetic code. *J. Mol. Evol*. **235**:5, 1377–80 (1994)**.**

Schwartz A.W. Chemical evolution: the first stages. *Naturwissenschaften*. **70**:8, 373–7 (1983).

Seal M., Weil-Ktorza O., Despotović D., Tawfik D.S., Levy Y., Metanis N., Longo L.M., Goldfarb D. Peptide-RNA coacervates as a cradle for the evolution of folded domains. *J. Am. Chem. Soc*. **144**, 14150–60 (2022).

Seidel S.A., Dijkman P.M., Lea W.A., van den Bogaart G., Jerabek-Willemsen M., Lazic A., Joseph J.S., Srinivasan P., Baaske P., Simeonov A., Katritch I., Melo F.A., Ladbury J.E., Schreiber G., Watts A., Braun D., Duhr S. Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods*. **59**:3, 301–15 (2013).

Sengupta S., Yang X., Higgs P.G. The mechanisms of codon reassignments in mitochondrial genetic codes. *J. Mol. Evol*. **64**:6, 662–88 (2007).

Sheehan J.C., Hess G.P. A new method of forming peptide bonds. *J. Am. Chem. Soc.* **77**:4, 1067–8 (1955).

Shen C., Yang L., Miller S.L., Oro J. Prebiotic synthesis of histidine. *J. Mol. Evol*. **31**:3, 167–74 (1990).

Sheppard K., Yuan J., Hohn M. J., Jester B., Devine K. M., Söll D. From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic Acids Res*. **36**:6, 1813–25 (2008).

Shibue R., Sasamoto T., Shimada M., Zhang B., Yamagishi A., Akanuma S. Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep*. **8**:1, 1227 (2018).

Shih F. F., Kalmar A. D. Determination of glutamine and asparagine by isocratic elution reverse phase liquid chromatography with fluorescent detection. *J. Liq. Chromatogr.* **7**:6, 1169–83 (1984).

Shimoyama A., Ogasawara R. Dipeptides and diketopiperazines in the Yamato-791198 and Murchison carbonaceous chondrites. *Orig. Life Evol. Biosph*. **32**:2, 165–79 (2002).

Shortle D., Stites W.E., Meeker A.K. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*. **29**:35, 8033−41 (1990).

Shu W., Yu Y., Chen S., Yan X., Liu Y., Zhao Y. Selective formation of Ser-His dipeptide via phosphorus activation. *Orig. Life Evol. Biosph*. **48**:2, 213–22 (2018).

Sibilska I., Chen B., Li L., Yin J. Effects of trimetaphosphate on abiotic formation and hydrolysis of peptides. *Life (Basel)*. **7**:4, 50 (2017).

Sibilska I., Feng Y., Li L., Yin J. Trimetaphosphate activates prebiotic peptide synthesis across a wide range of temperature and pH. *Orig. Life Evol. Biosph*. **48**:3, 277–87 (2018).

Sievers A., Beringer M., Rodnina M.V., Wolfenden R. The ribosome as an entropy trap. *Proc. Natl. Acad. Sci. USA*. **101**:21, 7897–901 (2004).

Škibola Z., Sovulj I.G., Maršavelski A. Impact of non-proteinogenic amino acid norvaline and proteinogenic valine misincorporation on a secondary structure of a model peptide. *J. Mol. Graph. Model*. **123**, 108528 (2023).

Smith T.F., Lee J.C., Gutell R.R., Hartman H. The origin and evolution of the ribosome. *Biol. Direct*. **3**:16 (2008).

Söding J., Lupas A.N. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*. **25**:9, 837–46 (2003).

Soini J., Falschlehner C., Liedert C., Bernhardt J., Vuoristo J., Neubauer P. Norvaline is accumulated after a down-shift of oxygen in Escherichia coli W3110. *Microb. Cell Fact*. **7**, 30 (2008).

Solis A.D. Reduced alphabet of prebiotic amino acids optimally encodes the conformational space of diverse extant protein folds. *BMC Evol. Biol*. **19**:1, 158 (2019).

Srivastava A., Ahmad S., Gromiha M.M. Deciphering RNA-recognition patterns of intrinsically disordered proteins. *Int. J. Mol. Sci*. **19**:6, 1595 (2018).

Steinman G., Lemmon R.M., Calvin M. Cyanamide: A possible key compound in chemical evolution. *Proc. Natl. Acad. Sci. USA*. **52**:1, 27–30 (1964).

Street A.G., Mayo S.L. Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. USA*. **96**:16, 9074–6 (1999).

Su M., Schmitt C., Liu Z., Roberts S.J., Liu K.C., Röder K., Jäschke A., Wales D.J, Sutherland J.D. Triplet-encoded prebiotic RNA aminoacylation. *J. Am. Chem. Soc*. **145**:29, 15971–80 (2023)

Sun F.J., Caetano-Anollés G. The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J. Mol. Evol.* **66**, 21–35 (2008).

Suwannachot Y, Rode B.M. Mutual amino acid catalysis in salt-induced peptide formation supports this mechanism's role in prebiotic peptide evolution. *Orig. Life Evol. Biosph*. **29**:5, 463–71 (1999).

Tadeus C., Wirström E.S., Bergman P., Charnley S.B., Chuang Y.L., Kuan Y.J. Deep search for glycine conformers in Barnard 5. *Mon. Not. R. Astron. Soc.* **524**:4, 5993–6003 (2023).

Tagami S., Attwater J., Holliger P. Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat. Chem*. **9**:4, 325–32 (2017).

Taillades J., Beuzelin I., Garrel L., Tabacik V., Bied C., Commeyras A. N-carbamoyl-alpha-amino acids rather than free alpha-amino acids formation in the primitive hydrosphere: a novel proposal for the emergence of prebiotic peptides. *Orig. Life Evol. Biosph*. **28**:1, 61-77 (1998).

Takahashi J., Hosokawa T., Masuda H., Kaneko T., Kobayashi K., Saito T., Utsumi Y. Abiotic synthesis of amino acids by X-ray irradiation of simple inorganic gases. *Appl. Phys. Lett*. **74**, 877–9 (1999).

Takahashi Y., Mihara H. Construction of a chemically and conformationally self-replicating system of amyloid-like fibrils. *Bioorg. Med. Chem*. **12**:4, 693–9 (2004).

Takano Y., Takahashi J.I., Kaneko T., Marumo K., Kobayashi K. Asymmetric synthesis of amino acid precursors in interstellar complex organics by circularly polarized light. *Earth Planet. Sci. Lett*. **254**, 106–14 (2007).

Tamura K., Schimmel P. Oligonucleotide-directed peptide synthesis in a ribosome- and ribozyme-free system. *Proc. Natl. Acad. Sci. USA*. **98**:4, 1393–7 (2001).

Tamura K., Schimmel P. Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *Proc. Natl. Acad. Sci. USA*. **100**:15, 8666–9 (2003).

Tamura K., Schimmel P. Chiral-selective aminoacylation of an RNA minihelix. *Science*. **305**:5688, 1253 (2004).

Tamura K. Ribosome evolution: emergence of peptide synthesis machinery. *J. Biosci*. **36**:5, 921–8 (2011).

Tanaka J., Doi N., Takashima H., Yanagawa H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci*. **19**:4, 786−95 (2010).

Tang G.Q., Hu H., Douglas J., Carter C.W.Jr. Primordial aminoacyl-tRNA synthetases preferred minihelices to full-length tRNA. *Nucleic Acids Res*. **52**:12, 7096–111 (2024).

Tena-Solsona M., Nanda J., Díaz-Oltra S., Chotera A., Ashkenasy G., Escuder B. Emergent catalytic behavior of self-assembled low molecular weight peptide-based aggregates and hydrogels. *Chemistry*. **22**:19, 6687–94 (2016).

Thoma B, Powner MW. Selective synthesis of lysine peptides and the prebiotically plausible synthesis of catalytically active diaminopropionic acid peptide nitriles in water. *J. Am. Chem Soc*. **145**:5, 3121–30 (2023).

Tian F., Toon O.B., Pavlov A.A., De Sterck H. A hydrogen-rich early Earth atmosphere. *Science*. **308**:5724, 1014−7 (2005).

Tiller K.E., Tessier P.M. Lifting the veil on amyloid drug design. *eLife*. **2**, e01089 (2013).

Tong A., Tang X., Zhang F., Wang B. Study on the shift of ultraviolet spectra in aqueous solution with variations of the solution concentration. *Spectrochim. Acta A. Mol. Biomol. Spectrosc*. **234**, 118259 (2020).

Torres de Farias S. Suggested phylogeny of tRNAs based on the construction of ancestral sequences. *J. Theor. Biol*. **335**, 245–8 (2013).

Torres de Farias S., Rêgo T.G., José M.V. Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio*. **4**, 175–8 (2014).

Torres de Farias S., Gaudêncio Rêgo T., José M.V. peptidyl transferase center and the emergence of the translation system. *Life (Basel)*. **7**:2, 21 (2017).

Tretyachenko V., Vymětal J., Neuwirthová T., Vondrášek J, Fujishima K, Hlouchová K. Modern and prebiotic amino acids support distinct structural profiles in proteins. *Open Biol*. **12**:6, 220040 (2022).

Trifonov E.N. Tracing Life back to elements. *Phys. Life Rev*. **5**:2, 121−32 (2008).

Tsuge M., Watanabe N. Radical reactions on interstellar icy dust grains: experimental investigations of elementary processes. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci*. **99**:4, 103–30 (2023).

Turanov A.A., Xu X.M., Carlson B.A., Yoo M.H., Gladyshev V.N., Hatfield D.L. Biosynthesis of selenocysteine, the 21st amino acid in the genetic code, and a novel pathway for cysteine biosynthesis. *Adv. Nutr*. **2**:2, 122–8 (2011).

Turk R.M., Chumachenko N.V., Yarus M. Multiple translational products from a five-nucleotide ribozyme. *Proc. Natl. Acad. Sci. USA*. **107**:10, 4585–9 (2010).

Udenfriend S., Stein S., Böhlen P., Dairman W., Leimgruber W., Weigele M. Fluorescamine: a reagent for assay of amino acids, peptides, proteins, and primary amines in the picomole range. *Science*. **178**:4063, 871–2 (1972).

Ueda H., Shibuya T. Composition of the primordial ocean just after its formation: constraints from the reactions between the primitive crust and a strongly acidic, CO_2-rich fluid at elevated temperatures and pressures. *Minerals*. **11**:4, 389 (2021).

Uversky V.N., Gillespie J.R., Fink A.L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins*. **41**:3, 415–27 (2000).

van der Gulik P., Massar S., Gilis D., Buhrman H., Rooman M. The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* **261**:4, 531–9 (2009).

van der Gulik P.T., Speijer D. How amino acids and peptides shaped the RNA world. *Life (Basel)*. **5**:1, 230–46 (2015).

Vauthey S., Santoso S., Gong H., Watson N., Zhang S. Molecular self-assembly of surfactant-like peptides to form nanotubes and nanovesicles. *Proc. Natl. Acad. Sci. USA*. **99**:8, 5355–60 (2002).

Vázquez-Salazar A., Tan G., Stockton A., Fani R., Becerra A., Lazcano A. Can an imidazole be formed from an alanyl-seryl-glycine tripeptide under possible prebiotic conditions? *Orig. Life Evol. Biosph*. **47**:3, 345–54 (2017).

Vázquez-Salazar A., Lazcano A. Early life: embracing the RNA world. *Curr. Biol*. **28**:5, R220–2 (2018).

Vegotsky A., Harada K., Fox S.W. The characterization of polyaspartic acid and some related compounds. *J. Am. Chem. Soc*. **80**:13, 3361–6 (1958).

Vincenzi M., Mercurio F.A., Leone M. About TFE: old and new findings. *Curr. Protein Pept. Sci*. **20**:5, 425–51 (2019).

von Maltzahn G., Vauthey S., Santoso S., Zhang S. Positively charged surfactant-like peptides self-assemble into nanostructures. *Langmuir*. **19**:10, 4332–7 (2003).

Wächtershäuser G. Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev*. **52**, 452–84 (1988).

Walter K.U., Vamvaca K., Hilvert D. An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem*. **280**:45, 37742−6 (2005).

Weber A.L., Orgel L.E. Amino acid activation with adenosine 5'-phosphorimidazolide. *J. Mol. Evol*. **11**:1, 9–16 (1978).

Weber A.L., Miller S.L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol*. **17**:5, 273–84 (1981)

Weber A.L., Pizzarello S. The peptide-catalyzed stereospecific synthesis of tetroses: a possible model for prebiotic molecular evolution. *Proc. Natl. Acad. Sci. USA*. **103**:34, 12713–7 (2006).

Widmann J., Di Giulio M., Yarus M., Knight R. tRNA creation by hairpin duplication. *J. Mol. Evol*. **61**:4, 524–30 (2005).

Wieczorek R., Dörr M., Chotera A., Luisi P.L., Monnard P.A. Formation of RNA phosphodiester bond by histidine-containing dipeptides. *Chembiochem*. **14**:2, 217–23 (2013).

Wilde S.A., Valley J.W., Peck W.H., Graham C.M. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature*. **409**:6817, 175−8 (2001).

Woese C.R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA*. **54**:6, 1546–52 (1965).

Woese C.R. The molecular basis for gene expression. Chapters 6–7. In: *The Genetic Code*. New York: Harper & Row. pp. 156–60 (1967).

Wolman Y., Haverland W.J., Miller S.L. Nonprotein amino acids from spark discharges and their comparison with the Murchison meteorite amino acids. *Proc. Natl. Acad. Sci. USA*. **69**:4, 809–11 (1972).

Wong J.T. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA*. **72**:5, 1909–12 (1975).

Wong J.T. Coevolution theory of the genetic code at age thirty. *Bioessays*. **27**:4, 416–25 (2005).

Wong T.F. Triphasic development of the genetic code. *Chem. Rev*. **124**:17, 9866–72 (2024).

Wu L.F., Su M., Liu Z., Bjork S.J., Sutherland J.D. Interstrand aminoacyl transfer in a tRNA acceptor stem-overhang mimic. *J. Am. Chem. Soc*. **143**:30, 11836–42 (2021).

Yamagata Y., Watanabe H., Saitoh M., Namba T. Volcanic production of polyphosphates and its relevance to prebiotic evolution. *Nature*. **352**:6335, 516–9 (1991).

Yamagata Y., Inomata K. Condensation of glycylglycine to oligoglycines with trimetaphosphate in aqueous solution. II: catalytic effect of magnesium ion. *Orig. Life Evol. Biosph*. **27**:4, 339–44 (1997).

Yanagawa H., Kobayashi K. An experimental approach to chemical evolution in submarine hydrothermal systems. *Orig. Life Evol. Biosph.* **22**, 147–59 (1992).

Yao S., Ghosh I., Zutshi R., Chmielewski J. A pH-modulated, self-replicating peptide. *J. Am. Chem. Soc*. **119**:43, 10559–60 (1997).

Yarus M., Caporaso J.G., Knight R. Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem*. **74**, 179–98 (2005).

Yarus M., Widmann J.J., Knight R. RNA–amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.* **69**, 406–29 (2009).

Yeh J.I., Du S., Tortajada A., Paulo J., Zhang S. Peptergents: peptide detergents that improve stability and functionality of a membrane protein, glycerol-3-phosphate dehydrogenase. *Biochemistry*. **44**:51, 16912–9 (2005).

Ying J., Fu S., Li X., Feng L., Xu P., Liu Y., Xiang G., Zhao Y. A plausible model correlates prebiotic peptide synthesis with the primordial genetic code. *Chem. Commun (Camb.)*. **54**:62, 8598–601 (2018).

Yokota K., Satou K., Ohki S.Y. Comparative analysis of protein thermostability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Sci. Technol. Adv. Mater*. **7**, 255–62 (2006).

Young G., Hundt N., Cole D., Fineberg A., Andrecka J., Tyler A., Olerinyova A., Ansari A., Marklund E.G., Collier M.P., Chandler S.A., Tkachenko O., Allen J., Crispin M., Billington N., Takagi Y., Sellers J.R., Eichmann C., Selenko P., Frey L., Riek R., Galpin M.R., Struwe W.B., Benesch J.L.P, Kukura P. Quantitative mass imaging of single biological macromolecules. *Science*. **360**:6387, 423–7 (2018).

Yu S., Solano M.D., Blanchard M.K., Soper-Hopper M.T., Krishnamurthy R., Fernández FM, Hud N.V., Schork F.J., Grover MA. Elongation of model prebiotic proto-peptides by continuous monomer feeding. *Macromolecules*. **50**: 9286–94 (2017).

Zaia D.A., Zaia C.T., De Santana H. Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph*. **38**:6, 469–88 (2008).

Zearfoss N.R., Ryder S.P. End-labeling oligonucleotides with chemical tags after synthesis. *Methods Mol. Biol*. **941**, 181–93 (2012).

Zeichner S.S., Aponte J.C., Bhattacharjee S., Dong G., Hofmann A.E., Dworkin J.P., Glavin D.P., Elsila J.E., Graham H.V., Naraoka H., Takano Y., Tachibana S., Karp A.T., Grice K., Holman A.I., Freeman K.H., Yurimoto H., Nakamura T., Noguchi T., Okazaki R., Yabuta H., Sakamoto K., Yada T., Nishimura M., Nakato A., Miyazaki A., Yogata K., Abe M., Okada T., Usui T., Yoshikawa M., Saiki T., Tanaka S., Terui F., Nakazawa S., Watanabe S.I., Tsuda Y., Hamase K., Fukushima K., Aoki D., Hashiguchi M., Mita H., Chikaraishi Y., Ohkouchi N., Ogawa N.O., Sakai S., Parker E.T., McLain H.L., Orthous-Daunay F.R., Vuitton V., Wolters C., Schmitt-Kopplin P., Hertkorn N., Thissen R., Ruf A., Isa J., Oba Y., Koga T., Yoshimura T., Araoka D., Sugahara H., Furusho A., Furukawa Y., Aoki J., Kano K., Nomura S.M., Sasaki K., Sato H., Yoshikawa T., Tanaka S., Morita M., Onose M., Kabashima F., Fujishima K., Yamazaki T., Kimura Y., Eiler J.M. Polycyclic aromatic hydrocarbons in samples of Ryugu formed in the interstellar medium. *Science*. **382**:6677, 1411–16 (2023).

Zhang S., Lockshin C., Cook R., Rich A. Unusually stable beta-sheet formation in an ionic self-complementary oligopeptide. *Biopolymers*. **34**:5, 663−72 (1994).

Zhao F., Akanuma S. Ancestral sequence reconstruction of the ribosomal protein uS8 and reduction of amino acid usage to a smaller alphabet. *J. Mol. Evol*. **91**:1, 10−23 (2023).

Zhao X., Nagai Y., Reeves P.J., Kiley P., Khorana H.G., Zhang S. Designer short peptide surfactants stabilize G protein-coupled receptor bovine rhodopsin. *Proc. Natl. Acad. Sci. USA*. **103**:47, 17707–12 (2006).

Zhu C.T., Zeng X.B., Huang W.D. Codon usage decreases the error minimization within the genetic code. *J. Mol. Evol*. **57**:5, 533–7 (2003).

Zozulia O., Dolan M.A., Korendovych I.V. Catalytic peptide assemblies. *Chem. Soc. Rev*. **47**:10, 3621–39 (2018).

Zuo Z., Peng D., Yin X., Zhou X., Cheng H., Zhou R. Genome-wide analysis reveals origin of transfer RNA genes from tRNA halves. *Mol. Biol. Evol*. **30**:9, 2087–98 (2013).

8. Publications and submitted manuscripts

The list of articles and submitted manuscripts included in this thesis:

1. Makarov M., Sanchez Rocha A.C., Krystufek R., Cherepashuk I., Dzmitruk V., Charnavets T., Faustino A.M., Lebl M., Fujishima K., Fried S.D., Hlouchova K. Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc*. **145**:9**,** 5320−9 (2023).
2. Makarov M., Kryštůfek R., Manriquez-Sandoval E., Dutta S., Kalvoda T., Kormaník J.M., Chernavets T., Lébl M., Brown S.M., Konvalinka J., Rulíšek L., Singharoy A., Fried S.D., Hlouchová K. Self-assembly of random primordial acidic peptides as a stepping stone to foldable polymers. *Under review in Chem*. (2024).
3. Codispoti S., Yamaguchi T., Makarov M., Giacobeli V.G., Mašek M., Kolář M.H., Sanchez Rocha A.C., Fujishima K., Zanchetta G., Hlouchová K. The interplay between peptides and RNA is critical for protoribosome compartmentalization and stability. *Nucleic Acids Res*. **28**, gkae823 (2024).

The list of articles and submitted manuscripts not included in this thesis:

4. Makarov M., Meng J., Tretyachenko V., Srb P., Březinová A., Giacobeli V.G., Bednárová L., Vondrášek J., Dunker A.K., Hlouchová K. Enzyme catalysis prior to aromatic residues: reverse engineering of a dephospho-CoA kinase. *Protein Sci*. **30**:5, 1022−34 (2021).
5. Giacobelli V.G., Fujishima K., Lepšík M., Tretyachenko V., Kadavá T., Makarov M., Bednárová L., Novák P., Hlouchová K. In vitro evolution reveals noncationic protein-RNA interaction mediated by metal ions. *Mol. Biol. Evol*. **39**:3, msac032 (2022).
6. Fried S.D., Fujishima K., Makarov M., Cherepashuk I., Hlouchova K. Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc Interface*. **19**:187, 20210641 (2022).
7. Sanchez-Rocha A.C., Makarov M., Pravda L., Novotný M., Hlouchová K. Coenzyme-protein interactions since early life. *eLife.* **13**, RP94174 (2024).

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Article

Early Selection of the Amino Acid Alphabet Was Adaptively Shaped by Biophysical Constraints of Foldability

Mikhail Makarov, Alma C. Sanchez Rocha, Robin Krystufek, Ivan Cherepashuk, Volha Dzmitruk, Tatsiana Charnavets, Anneliese M. Faustino, Michal Lebl, Kosuke Fujishima, Stephen D. Fried,* and Klara Hlouchova*

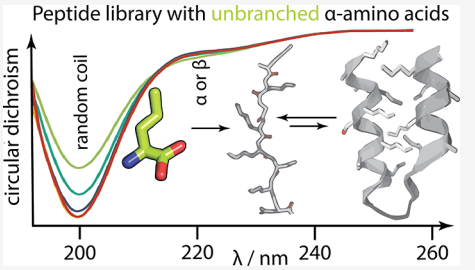
Cite This: *J. Am. Chem. Soc.* 2023, 145, 5320–5329

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via 147.231.236.32 on October 22, 2024 at 18:09:50 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: Whereas modern proteins rely on a quasi-universal repertoire of 20 canonical amino acids (AAs), numerous lines of evidence suggest that ancient proteins relied on a limited alphabet of 10 "early" AAs and that the 10 "late" AAs were products of biosynthetic pathways. However, many nonproteinogenic AAs were also prebiotically available, which begs two fundamental questions: Why do we have the current modern amino acid alphabet and would proteins be able to fold into globular structures as well if different amino acids comprised the genetic code? Here, we experimentally evaluate the solubility and secondary structure propensities of several prebiotically relevant amino acids in the context of synthetic combinatorial 25-mer peptide libraries. The most prebiotically abundant linear aliphatic and basic residues were incorporated along with or in place of other early amino acids to explore these alternative sequence spaces. The results show that foldability was likely a critical factor in the selection of the canonical alphabet. Unbranched aliphatic amino acids were purged from the proteinogenic alphabet despite their high prebiotic abundance because they generate polypeptides that are oversolubilized and have low packing efficiency. Surprisingly, we find that the inclusion of a short-chain basic amino acid also decreases polypeptides' secondary structure potential, for which we suggest a biophysical model. Our results support the view that, despite lacking basic residues, the early canonical alphabet was remarkably adaptive at supporting protein folding and explain why basic residues were only incorporated at a later stage of protein evolution.



INTRODUCTION

The role of peptides and proteins in life's emergence has historically been sidelined as they have been, at times, perceived as only relevant after the evolution of nucleotide-based polymers, the genetic code, and a translation apparatus. However, the prebiotic abundance of amino acids and their ease of condensation resulting in polymers capable of creating functional hubs along with various cofactors (such as metal ions and organic compounds) has prompted some chemists to reconsider peptides' role during life's early evolution.[1−4] While extant proteins are built from a sequence space spanned by the 20 canonical amino acids (cAAs) and rely on the specificity of the Central Dogma, early peptides (and peptide-like polymers) likely arose from a larger pool of prebiotically plausible monomers.

The modern protein alphabet was most likely selected during the first 10−15% of Earth history (~4.5−3.9 Ga), but the factors guiding its chemical and biological evolution remain unclear.[5] Some of the cAAs are thought to be "late" additions to the genetic code, enabled only (or enriched significantly) by the evolution of their biosynthetic pathways.[6] On the other hand, the "early" amino acids (Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr, and Val) have been found abundant in several different prebiotic sources, including atmospheric spark discharge (Miller−Urey type) experiments, hydrothermal vent syntheses, and even meteorites.[7−10] Although some studies have detected possible prebiotic synthesis pathways for other cAAs under very specific conditions,[11,12] a consensus "early" canonical alphabet has been defined by two independent meta-analyses based on reports of the prebiotic abundance of amino acids and the order with which they entered the genetic coding.[6,13] The ranking derived by these two meta-analyses has since provided a base for multiple independent studies recapitulating the protein consequences of the evolving alphabet.[14−16] Nevertheless, besides the "early" cAAs, many other noncanonical amino acids (ncAAs) and their alternatives (such as β- and γ-amino acids, hydroxy acids or dicarboxylic acids) were apparently available through prebiotic synthesis.[2,3,7,8,17] Why certain amino acids were selected over others (and when) has been repeatedly discussed in the literature over the last few decades.[5,8,17−19] This question can

Received: December 6, 2022
Published: February 24, 2023



ACS Publications

© 2023 American Chemical Society

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

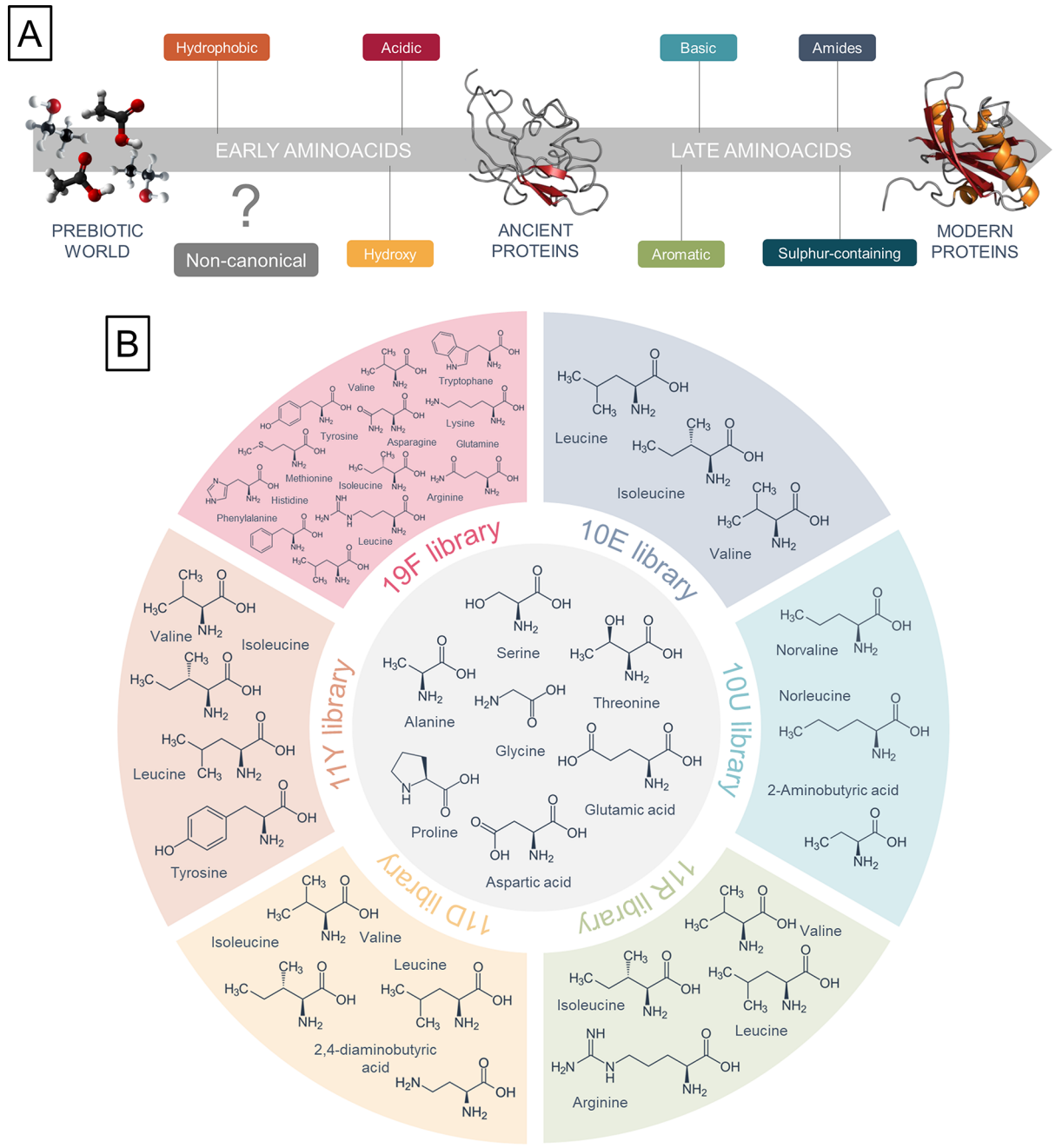


Figure 1. (A) Assumed early and late stages of amino acid alphabet incorporation during protein evolution and (B) design of peptide libraries based on this order.

be further subdivided into two separate but related questions, namely: (i) Why were the 10 early cAAs selected from the prebiotic environment and (ii) what factors drove the selection of the additional residues in the following era? Has protein evolution been successful as a consequence of the selected cAAs, or could similar structural and functional spaces be formed with alternative alphabets?[8]

While some of the choices for the early amino acids are quite expected (especially for the C2 and C3 amino acids), others have been repeatedly considered thought-provoking. Most strikingly, the early amino acid alphabet lacked positively charged residues, even though diaminopropionic and diaminobutyric acids (C3 and C4 amino acids, DAP and DAB, respectively) have been shown to be prebiotically plausible.[20,21] Neither lysine nor arginine are considered prebiotically plausible, while ornithine is known to cyclize, promoting peptide chain scission.[8] Moreover, the canonical alphabet does not include some of the most abundant linear aliphatic amino acids such as α-amino-n-butyric acid (ABA), norvaline (Nva), and norleucine (Nle) (C4, C5, and C6, respectively) while their branched alternatives (such as Val, Leu, Ile) were selected[8] despite being available at much reduced abundances. It has been hypothesized that perhaps the earlier version of the amino acid alphabet included some ncAAs that were later replaced by late cAAs.[2,22] This is an intriguing hypothesis especially in light of the fact that some of these ncAAs have been reported as promiscuous substrates of some aminoacyl-tRNA synthetases.[23]

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

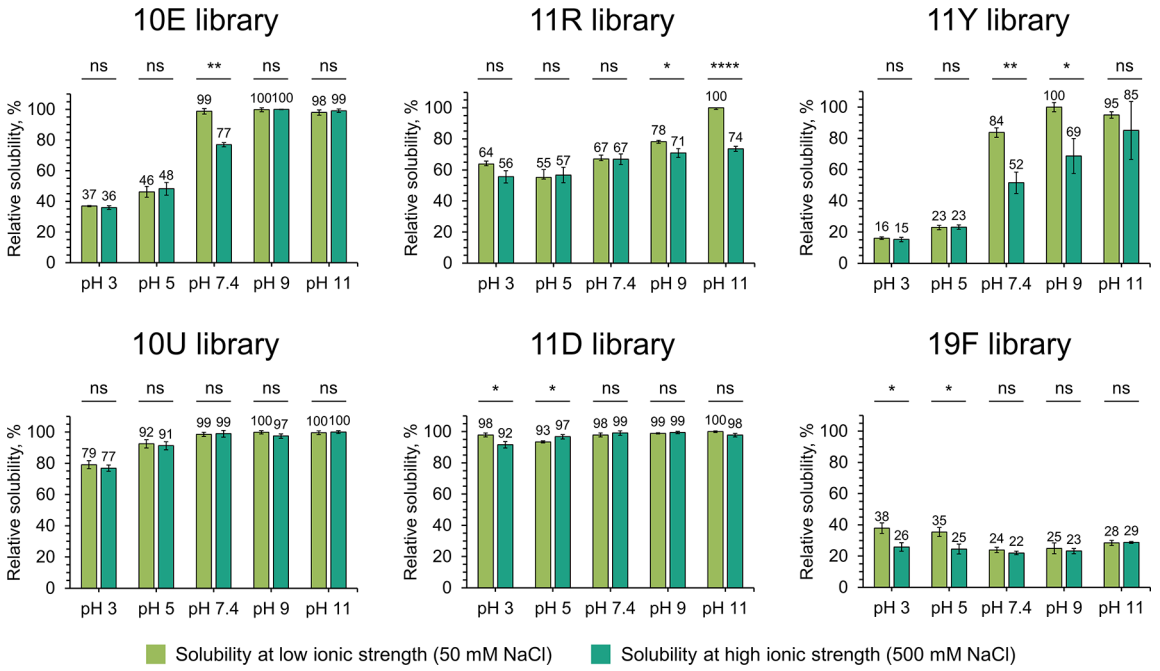


Figure 2. Solubility of 25-mer combinatorial peptide libraries at different pHs and ionic strengths. Solubility was measured for 0.5 mg/mL nominal peptide library solutions in 20 mM ABP buffers (pH 3−11) at 50 and 500 mM NaCl. Solubility of 19F peptide library was measured by fluorescence spectroscopy, and solubility of the other peptide libraries was measured by UV−vis absorption spectroscopy. Solubility measurements were conducted in technical triplicate ($n = 3$) on separate preparations of each library, and error bars correspond to standard deviations. Student's t tests were performed to assess significance of salt on solubility (ns, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$). For detailed statistical analysis, see Table S2.

This study experimentally determines the properties that would have accompanied some of the most feasible ncAA candidates from the prebiotic pool. We incorporated the selected ncAAs into combinatorial peptide libraries along with (or replacing) other early amino acids to evaluate their effect on fundamental physicochemical properties such as solubility and secondary structure formation. The outcomes of our study show that an inclusion of some of the most feasible alternatives would *not* have supported the twin requirements of rich secondary structure potential and solubility. Hence, our study demonstrates that the biophysical requirement of foldability (by which we mean "ability to fold into a soluble defined conformation") provided a selective pressure shaping the building block selection in the early alphabet and therefore explains why certain early AAs (and not others) were selected to construct proteins.

RESULTS

Design and Synthesis of Peptide Libraries from Alternative Alphabets. Twenty-five-mer combinatorial peptide libraries were synthesized using solid-support chemical synthesis with isokinetic mixtures of amino acids (see Table S1).

To model the sequence space available to different subsets of the amino acid alphabets, the libraries included the entire canonical alphabet without Cys (19F; F = full), its prebiotically available subset of 10 (10E; E = early), an alternative of 10E where the branched aliphatic amino acids were substituted with their unbranched prebiotically abundant alternatives (10U; U = unbranched), the 10E library supplemented with Arg as a representative of a modern cationic cAA (11R; R = Arg); or DAB as a representative of a potentially early cationic AA (11D; D = 2,4-diaminobutyric acid); or Tyr as a representative of an aromatic AA (11Y; Y = Tyr) (Figure 1).

MALDI spectra confirmed the expected molecular weight range and distribution of the combinatorial libraries, reflecting their respective compositions (Figure S1). Apart from the DAB in 11D library, the amino acid analyses of all libraries confirmed the expected composition (Figure S2). The analysis protocol could not detect DAB whose presence in the 11D library (contrary to the 10E library) was instead determined using fluorescamine assays (Figure S3). All the libraries passed quality control and were lyophilized to establish Cl^- as the counterion for all downstream analyses.

Solubility and Aggregation Profiles. The ability to remain soluble under different conditions could represent an important selection factor during the formation of the early protein alphabet, as prebiotically relevant environments spanned from alkaline hydrothermal vents to acidic lakes.[24,25] We measured the solubility profiles of these peptide libraries in the pH range 3−11 and also at low vs high ionic strength, spectrophotometrically by absorption of peptide bonds at 215 nm and fluorometrically in case of library 19F due to its very poor solubility (Figure 2). In general, ionic strength did not significantly modulate solubility, with the exception of the 11Y library (see below).

At a fixed nominal peptide concentration (0.5 mg/mL), the 19F library is significantly less soluble than the libraries with narrower amino acid repertoires. This result is in line with previous studies reporting that subsets of the canonical library that are enriched with the early amino acids are significantly more soluble than the full alphabet version.[14,16,26] This phenomenon could be partly explained by the physicochemical nature of the full alphabet (such as the presence of aromatic amino acids) leading to increased aggregation propensity[27] or

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

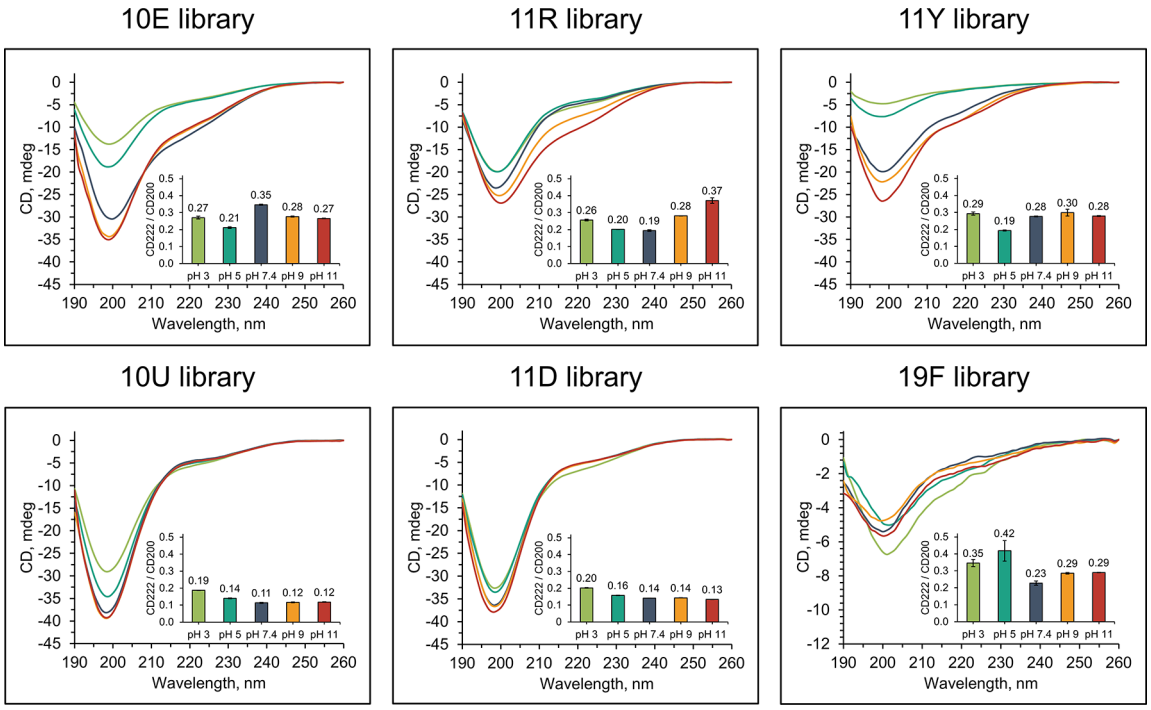


Figure 3. Effect of pH on the secondary structure of 25-mer combinatorial peptide libraries. CD spectra were recorded at a nominal concentration of 0.2 mg/mL peptide library (however is less for aggregation prone libraries, e.g., 19F) in a series of 10 mM ABP buffers (pH 3−11) at 25 mM NaCl. The inset graphs show the ratios of CD signals at 222 to 200 nm. CD spectra were acquired in duplicate and averaged and the 222/200 ellipticity ratio calculated. Measurements were then conducted in technical duplicate ($n = 2$) on separate preparations of each library, and error bars correspond to standard deviations. For detailed statistical analysis, refer to Table S3.

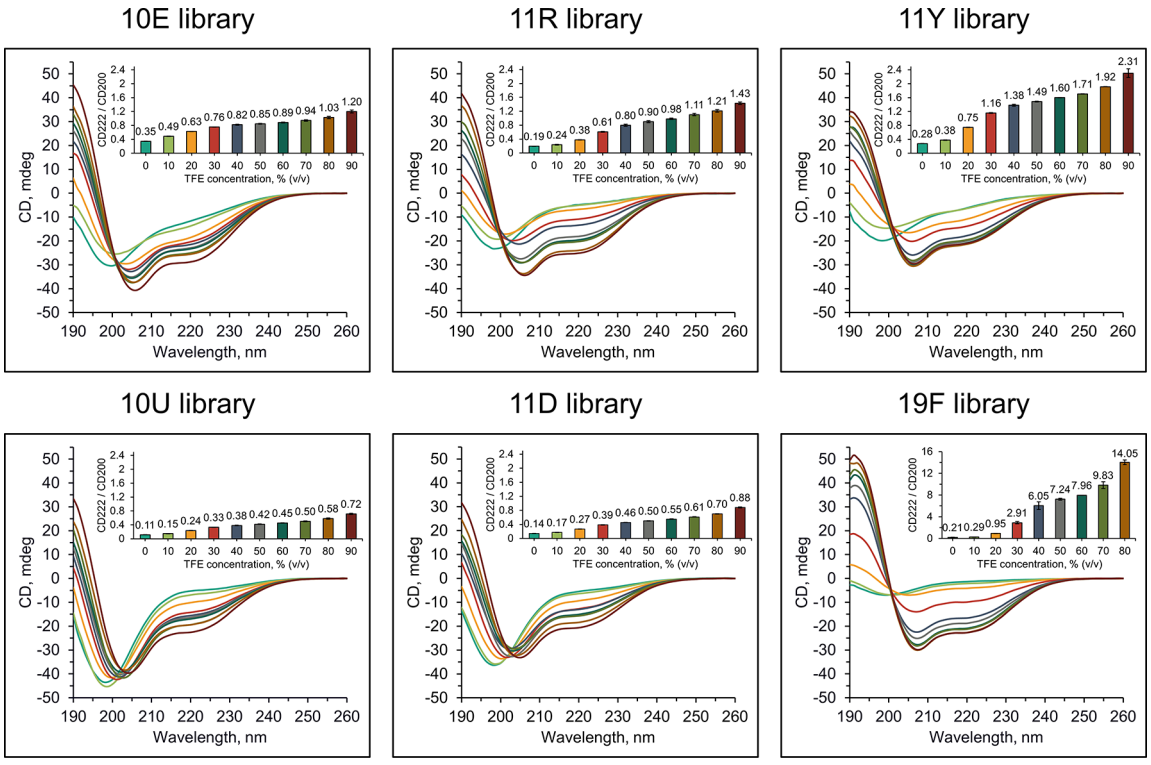


Figure 4. Effect of 2,2,2-trifluoroethanol to induce secondary structure of 25-mer combinatorial peptide libraries. CD spectra were recorded at 0.2 mg/mL nominal concentration in 10 mM ABP buffer (pH 7.4) with 0−90% (v/v) 2,2,2-trifluoroethanol. The inset graphs show the ratio of CD signal at 222−200 nm. The *y*-axis has the same scale for all except for 19F library. CD spectra were acquired in duplicate, averaged, and the 222/200 ellipticity ratio calculated. Measurements were then conducted in technical duplicate ($n = 2$) on separate preparations of each library, and error bars correspond to standard deviations. For detailed statistical analysis, refer to Table S4.

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

the greater "search complexity" of a larger alphabet (leading to a more rugged landscape). Regardless of the mechanism, the results support the view that peptides composed of the full alphabet are more reliant on chaperones and/or translation to fold.[2,27]

The solubilities of the canonical library subsets (10E, 11R, and 11Y) trend upward with increasing pH (going from pH 3 to pH 11) and 11Y (which includes Tyr as a representative aromatic amino acid) is the least soluble of these three as expected. This result can be potentially explained by pointing out that all of these alphabets (as do modern proteomes)[28] tilt toward being acidic; hence, at alkaline pH, the peptides will accrue larger net negative charge and will repel one another electrostatically (inhibiting aggregation). This is broadly consistent with the practice of using mildly alkaline buffers in protein refolding experiments, which tend to facilitate refolding by preventing aggregation.[29] In general, high salt concentrations (500 mM NaCl) did not perturb solubility levels to the same extent as pH change, with the exception of the 11Y library for which high salt significantly lowered solubility at the alkaline pH's where it was soluble. We speculate this may be due to cation-π interactions with the nonphysiologically high tyrosine content (ca. 9.4% in 11Y vs 2−3% in natural proteomes).

The two libraries that include the noncanonical alternatives of aliphatic (10U) and basic (11D) amino acids are strikingly more soluble than 10E and 11R throughout the examined pH range (Figure 2). Both of these results are unexpected for distinct reasons. Based on the physical chemistry of elementary hydrocarbons, one might expect that the 10U library would be less soluble because linear hydrocarbons are more hydrophobic than branched hydrocarbons of equal carbon-count because they create cavities with greater surface area.[30] Moreover, the inclusion of a basic amino acid would raise the average isoelectric point of the library (relative to 10E) to be closer to neutral, which would decrease intermolecular repulsion and thereby increase aggregation.

The soluble fractions of the libraries were additionally screened for the occurrence of soluble aggregates using size-exclusion chromatography (Figures S4 and S5, in low and high ionic strength, respectively). No significant amount of such phenomenon was observed with the exception of 11Y and 10E libraries where a minor (up to ∼10%) fraction of soluble aggregates was detected (Figure S4). Overall, the elution profiles correlated with the spectrophotometric solubility measurements in which the 11R library was the most soluble of the canonical subset libraries, and the 10U and 11D libraries were fully soluble across the entire pH range.

Secondary Structure Propensity. Natural nonfolding sequences as well as random sequences with low secondary structure content have been previously reported to be highly soluble.[16,31] The solubility profiles of the 11D and 10U libraries therefore suggest that the incorporation of the respective ncAAs may decrease the potential for secondary structure formation. CD spectroscopy was used to estimate average secondary structure content of the libraries over the studied pH range (Figure 3) and upon its induction with 2,2,2-trifluoroethanol (TFE) in pH 7.4 (Figure 4). Because no dominant secondary structure motif is expected in these highly diverse libraries and we are rather interested in average structural propensities, the secondary structure content was evaluated as the ratio of CD signal at 222 nm to the CD signal at 200 nm (shown by inset graphs in Figures 3 and 4), because both α helices and β sheets exhibit higher ellipticities than random coils at 222 nm.[32] We did not observe any significant induction of secondary structure upon changing the metal ion concentration (Figure S6).

The CD spectra of the 10U and 11D libraries showed significantly lower relative signals at 222 nm, indicating that these peptides possessed profoundly decreased secondary structural propensity when compared with the respective canonical libraries 10E and 11R (Figure 3). At pH 7.4, the CD222/CD200 ratio was 3-fold lower for the 10U library relative to 10E (P-value = 0.0004 by Student's t test (n = 2)) and 2.4-fold lower for the 11D library relative to 10E (P-value = 0.0003 by Student's t test (n = 2)). These differences are also significant relative to 11R (P-values of 0.0056 and 0.0101, respectively). Such significant and large differences in secondary structure propensity between 10U and 11D relative to 10E carried across the entire pH range (Figure 3 and Table S3). Hence, the 10E library with branched aliphatic amino acids has a higher structural propensity than the 10U library with unbranched aliphatic alternatives in general. The addition of DAB in the 11D library also significantly decreases secondary structure propensity while the addition of canonical amino acids, Arg (in 11R) or Tyr (in 11Y), increases the secondary structure propensity mildly under some pH values. This trend becomes more pronounced upon TFE titration, which is often used to induce helical structure and increase protein stability.[32−34] Because the overall folding stability of our 25-mer peptide libraries is small, we sought to test whether the trends we observed would extrapolate to a scenario in which folding stabilities are uniformly increased. TFE addition potentiates secondary structure content robustly in the 10E library, an effect that is further amplified in alphabets that include Arg and Tyr (Figure 4). Remarkably, the inclusion of DAB (in the 11D library) significantly decreases the potentiating capacity of TFE, implying that the inclusion of DAB still reduces the secondary structure potential even when they have been artificially stabilized. While the poor solubility of the 19F library is reflected in the overall low intensity of its CD spectra, its elevated structural propensity (when compared with all the other library subsets) becomes evident in the TFE titration (Figure 4).

DISCUSSION

The absence of ABA, Nva, and Nle in the proteinogenic alphabet has been considered striking given their high prebiotic abundance.[10] Some authors have commented that this universal feature of biochemistry is a potential example of a "frozen accident" due to the absence of an obvious advantage of the canonical aliphatic residues over these ncAAs. At the same time, Weber and Miller hypothesized that inclusion of linear aliphatic amino acids would increase the side chain mobility and hence would not promote formation of ordered tertiary structure.[8] Our experimental observations support this hypothesis. Moreover, structural prediction with PEPstrMOD (a molecular dynamics based predictor that generalizes to noncanonical amino acids) on 1,000 randomly chosen peptides from each library show that the 10U library has considerably less secondary structure content (11%) than the 10E library (26%), confirming the same trend (Figure 5).[35]

To explain these findings, we start with the accepted model that the driving force for folding (and therefore secondary structure acquisition) is hydrophobic packing interactions.[36,37] Isolated secondary structures are generally not stable in neutral

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

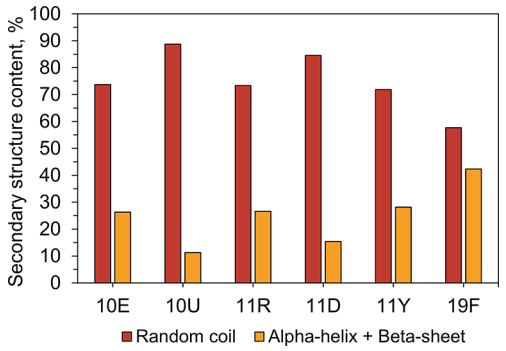


Figure 5. PEPstrMOD prediction of secondary structure content of 1,000 sequences selected randomly from the 25-mer combinatorial peptide libraries.

water[38]—this is because the interamide hydrogen bonds associated with secondary structure are equally stabilizing as amide-solvent hydrogen bonds in the unfolded state.[39−41] As a consequence, observation of secondary structure in our CD spectra can be treated as a proxy for peptides possessing hydrophobic packing interactions to stabilize those secondary structures.

Figure 6A explains why the 10U library has lower secondary structure-forming potential. Because unbranched hydrocarbons have more conformational freedom than their branched isomers, the ordering that is implicit in packing is expected to be more entropically costly. The lower packing efficiency of these side chains simultaneously explains the higher solubility of the 10U library. Hence, the requirement of foldability most likely acted to purge ABA, Nva, and Nle from the pool of amino acids that became proteinogenic, and ultimately canonical.

The more prebiotically abundant diamino acids, DAP and DAB, are also strikingly absent in the modern amino acid alphabet. One possible scenario is that DAP and DAB were utilized in an ancient alphabet and then were ultimately displaced with Arg, Lys, and His.[18] Ornithine has also been proposed as a prebiotically plausible diamino acid that could have assisted in nucleic acid binding.[42−44] However, it should be noted that ornithine is at least 1 order of magnitude less abundant in prebiotic environments and significantly more unstable compared to DAP and DAB.[20,43,44] Another school of thought holds that proteins lacked positively charged residues altogether until biosynthetic pathways for the prebiotically unavailable Arg, Lys, and His evolved.[6,13] While the second theory might seem unlikely at first glance, proteins lacking positively charged amino acids have recently been shown capable of folding and even binding with nucleic acids, assisted by metal ion cofactors.[9,27,45,46] Our results concerning the 11D library further support the second theory. Hence, this body of work is building a compelling case that primordial proteins were highly acidic.

We find that, when DAB is incorporated into the 10E alphabet, it reduces secondary structure potential. Hence, its inclusion would have been a step *backward* rather than a step *forward* toward the goal of assembling foldable polypeptides. The same trend is observed in the PEPstrMOD prediction on 1,000 randomly chosen peptides from the 11D library (Figure 5). The 11D library is also exceptionally soluble, so the combination of high solubility and low foldability implies that 11D polypeptides are mostly unfolded in aqueous solution. By way of contrast, the 19F library that has the most folding capacity (as seen in PEPStrMOD and in CD) is also the least soluble (cf. Figure 2), because its bulky aromatic amino acids can alternatively lead to folding or aggregation depending on

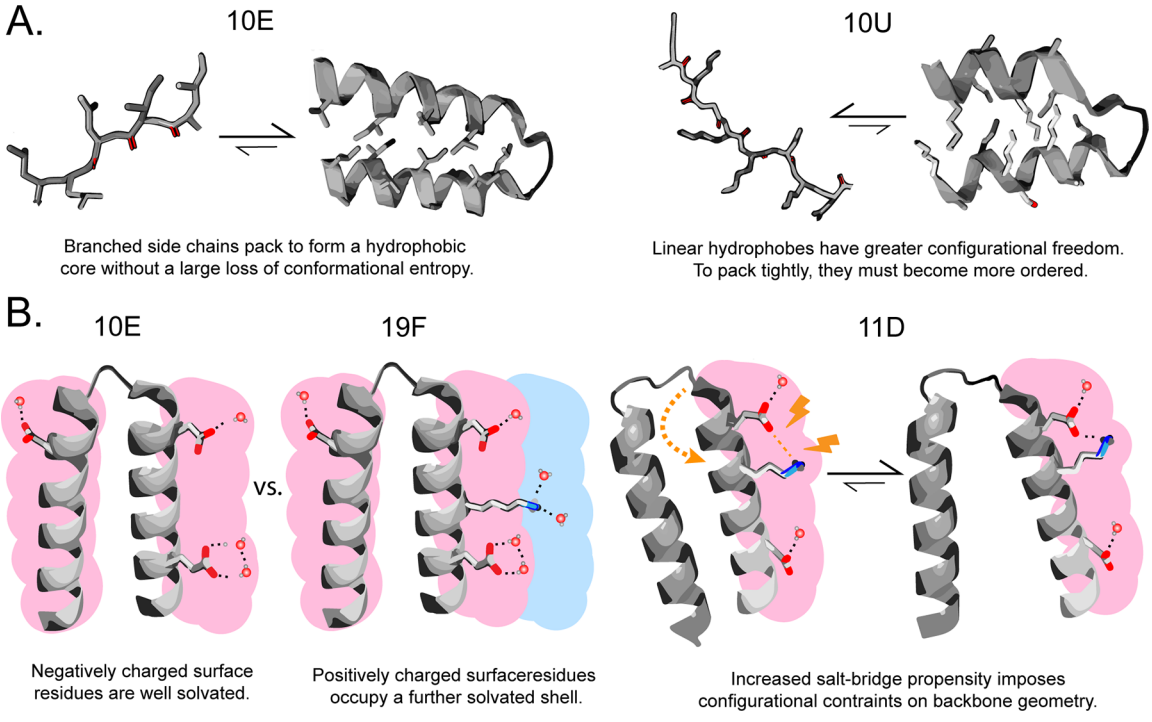


Figure 6. Biophysical models for the detrimental effect of unbranched hydrophobic side chains (A) and short-chained basic residues (B) on folding stability of small proteins. (A) Unbranched hydrophobic residues render core packing more entropically expensive. (B) Short-chained basic residues occupy the same solvation sphere as canonical acidic residues, increasing the frequency of salt bridges over solvation with water.

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

whether they are placed in appropriate positions in the peptide sequences.

Why does DAB harm protein structure propensity so significantly? It has been previously pointed out that these short-chain amines suffer from lactamization and acyl migration in peptides.[8,47] This point is important in relation to the long-term stability of peptide chains, but it cannot feasibly explain the results of our study given that lactamization is a slow reaction in relation to the time scale of our experimental work. Moreover, the MALDI spectra of the library did not show any detectable scission products. It has also been proposed that charged groups close to the backbone interfere with backbone H-bonding,[48] though this does not explain the high foldability of short-chained acidic residues in 10E. A third possibility that we propose (Figure 6B) is that DAB disrupts structural potential because of a conflict associated with its simultaneous presence with Asp and Glu, i.e., negatively charged AAs with short side chains. Especially in these short 25-mer polypeptides, charged residues are likely to be on the surface. However, by having surface-anions and surface-cations in such close proximity, a polypeptide would be inclined to fold in such a way to favor formation of ion-pair salt bridges (Figure 6B, right). The backbone angles required to form such salt bridges would create additional constraints that might be hard to simultaneously satisfy along with hydrophobic packing, hence the ion-pairs would stabilize the protein in an unpacked (and therefore unfolded) conformation. Naturally, there are two ways to avoid this folding problem: Either (1) place positive charges further away from the backbone (on a longer side chain) so that they will primarily interact by hydrogen bonding with water in a distinct solvation shell or (2) do not include positively charged residues at all (Figure 6). We hypothesize that early proteins avoided ion pair-induced unfolding using the second strategy, whereas modern proteins adopted the first, though this strategy only became available once Arg and Lys could be biosynthesized. One experiment that could further test this hypothesis is the synthesis of a counterfactual peptide library in which basic residues use short side chains (diaminobutyric acid) while acidic residues use long side chains (e.g., 2-aminoadipic acid). Our model would predict that such a combination would also support secondary structure formation. Such a finding addresses the central challenge for rational design of proteinogenic xeno alphabets: even a small library of candidate amino acid structures implies functionally infinite possible alphabets.[49,50] Limiting alphabets that contain both positive and negatively charged side chains to those with unequal side chain lengths could be a crucial principle guiding protein design in synthetic sequence space.

CONCLUSION

In summary, our study on the solubility and secondary structure propensity of several prebiotically relevant amino acid alphabets supports the assertion that foldability played a critical early role in governing which amino acids ultimately became part of the canonical alphabet.[51,52] Unbranched amino acids and short-chain basic amino acids were excluded, despite their prebiotic abundance because they "oversolubilized" polypeptides by stabilizing their unfolded conformations. Polypeptides with propensity to cooperatively fold *via* hydrophobic packing would have then been better adapted for binding cofactors, recognizing RNA, and performing enzymatic catalysis.[2] More broadly, this study supports the view that the early canonical alphabet (10E), despite its deficiencies in relation to the modern canonical alphabet, was remarkably adaptive at supporting folding for the earliest proteins.

EXPERIMENTAL PROCEDURES

All chemicals if not stated otherwise were purchased from Sigma-Aldrich.

Design and Preparation of Combinatorial Peptide Libraries. Split and mix synthesis of peptide libraries has its limitations.[53] A complete library of 25-mers composed of 10 amino acids would in theory require 10^{25} beads. One gram of solid phase support contains about 10^6 beads—therefore, it would cover only 10^{-17}% of possible structures. To increase the probability of completeness of the possible structures, it is necessary to conduct the synthesis considering individual molecules.

Average molecular weight of a 25-mer peptide composed of 10 amino acids is about 2500 Da. Therefore, 2.5 kg of these peptides would potentially contain 6×10^{23} individual peptides. In order to have a chance to synthesize a complete library, we would need ∼250 kg of peptides, where each peptide would be present only once. But, the chance of finding any given peptide in the mixture is about 70%. Therefore, studying a complete library is practically impossible.

We decided to synthesize the peptide mixture on 200 mg of the solid support and therefore create theoretically only $\sim 10^{-4}$% of possible structures. The approach using individual molecules is still capable of production of $10^{17}\times$ more possible structures than split and mix strategy. However, considering complete randomness of the synthetic process, we concluded that the sample is a representative collection of possible structures. One milligram of synthetic peptide mixture contains about 2.4×10^{17} individual peptides.

Synthesis of the peptide mixture was accomplished by coupling mixtures of amino acids in which the ratio of individual amino acids was adjusted according to their reactivities. This approach was used by several authors, and various amino acid ratios were reported.[54] We have used as the basis the ratios reported by Santi et al.[55,56] and adjusted the ratios for unnatural amino acids based on their structure and our previous experience (Table S1).

Synthesis was performed on the automatic peptide synthesizer Spyder Mark IV, using standard Fmoc peptide synthesis protocol.[57] Fmoc amino acids were individually weighted (see Table S1), mixed together, and dissolved to create 0.3 M solution in 0.3 M *N*-hydroxybenzotriazole (HOBt) in dimethylformamide (DMF). Rink resin (200 mg, 0.42 mmol/g) was distributed into 10 mL syringes of the synthesizer and swelled in DMF for 10 min. The mixtures from Table S1 were defined as extra amino acids 1−6 and placed into amino acid containers 21−26. Synthetic protocol was as follows: 2 × 1 min washing with DMF, 1 × 1 and 1 × 20 min treatment with 20% piperidine in DMF, 4 × 1 min wash with DMF, 1 × 1 min wash with 0.3 M HOBt, 2 × 60 min coupling with amino acid mixture, and 1 M diisopropylcarbodiimide in DMF. After 25 cycles of the synthesis, the resin was washed with DMF and dichloromethane and dried *in vacuo*. Dried resin was treated with 5 mL of Mixture K (trifluoroacetic acid−thioanisol−water−phenol−ethanedithiol, 82.5:5:5:5:2.5, v/v) for 2 h.[58] Resin was filtered off and peptide mixture was precipitated with diethyl ether three times and dried *in vacuo*. The pellet was then dissolved in 20% acetic acid and lyophilized. Prior to further experiments, all the peptide libraries were lyophilized three times with 1 mM HCl overnight.

Quality Control of Peptide Libraries. The molecular weight distributions of combinatorial peptide libraries were confirmed by mass spectrometry using UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) according to the standard procedure.

Prior to the amino acid analysis, the library samples were hydrolyzed in 6 M hydrochloric acid at 110 °C for 20 h and the hydrolysate was evaporated and reconstituted with 0.1 M hydrochloric acid containing the internal standard. Amino acid analysis was performed on Agilent 1260 HPLC (Agilent Technologies, Wald-

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

bronn, Germany) equipped with a fluorescence detector using automated *o*-phtalaldehyde/2-mercaptopropionic acid (OPA/MPA) or ninhydrin derivatization in the case of 10E, 10U, 11R and 11D, 11Y, and 19F libraries, respectively.

The presence of 2,4-diaminobutyric acid in the 11D library was confirmed by fluorescamine assay.[59] For the assay, peptide library solution (75 μL) in PBS buffer was incubated with 25 μL of fresh stock of 3 mg/mL fluorescamine in DMSO at room temperature for 1 h in a 96-well plate (Greiner 650209). Fluorescence intensity (λ_{Ex} = 365 nm, λ_{Em} = 470 nm) was recorded using Tecan Spark plate reader. Primary amine concentrations were determined by linear interpolation of measured intensities for peptide H-GTIQPYPFSWGY-NH_2 in concentration range 87.5−2.7 μM. Obtained amine concentrations were then divided by the concentration of the analyzed peptide determined by amino acid analysis to calculate the amount of amine equivalents per molecule. Experiments were conducted in duplicate.

Solubility and Aggregation Propensity Measurements. A series of 10 Britton−Robinson buffers at pH 3.0, 5.0, 7.4, 9.0, and 11.0 were prepared by mixing 20 mM acetic acid, 20 mM boric acid, and 20 mM phosphoric acid and adjusting pH to the desired value with 5 M NaOH. The ionic strength was adjusted to either 50 mM (low ionic strength) or 500 mM (high ionic strength) with NaCl, and all Britton−Robinson buffers were filter-sterilized using 0.22 μm PVDF membrane before use. Lyophilized peptide libraries were thoroughly resuspended in autoclaved Milli-Q water to 5 mg/mL and subsequently diluted 10 times into Britton−Robinson buffer to the final concentration of 0.5 mg/mL. The samples were gently shaken at room temperature for 30 min and then centrifuged at maximum speed (21,300*g*) for 15 min at 4 °C in order to remove insoluble fraction. The supernatant (soluble fraction) was taken and used for solubility and aggregation propensity analyses.

The relative amount of peptides in soluble fractions was estimated spectrophotometrically by absorption of peptide bonds at 215 nm for all peptide libraries except the 19F peptide library. Due to the extremely low solubility, solubility of the 19F peptide library was estimated by fluorescence spectroscopy using Pierce quantitative fluorometric peptide assay (Thermo Fisher Scientific), and 0.5 mg/mL solution in DMSO was used as a standard. Each peptide library was prepared at least three times independently. For each preparation, spectrophotometric measurements were performed five times, whereas fluorometric measurements were performed in triplicates.

The relative amount of soluble aggregates was estimated by size-exclusion chromatography. The 100 μL aliquot of supernatant was loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva) that was pre-equilibrated with two bed volumes of the corresponding 20 mM Britton−Robinson buffer. The peptides were eluted from the column by one bed volume of the buffer at 0.5 mL/min flow rate at room temperature, and the eluted peptides were detected by absorption at 215 nm.

Circular Dichroism Spectroscopy. The CD spectra were recorded using a Chirascan-plus spectrophotometer (Applied Photophysics, Leatherhead, UK) over the wavelength range 190−260 nm in steps of 1 nm with an averaging time of 1 s per step. Cleared samples at 0.2 mg/mL nominal concentration in 1 mm path-length quartz cells were placed into a cell holder, and spectra were recorded at room temperature. The CD signal was obtained as ellipticity in units of millidegrees, and the resulting spectra were averaged from two scans and buffer-spectrum subtracted. All CD measurements were performed twice, and the resulting ratio of ellipticity at 222 nm to ellipticity at 200 nm was averaged and used to estimate the secondary structure content in peptide libraries.

To estimate the effect of pH on secondary structure, 5 mg/mL suspensions of peptide libraries in Milli-Q water were diluted into 10 mM Britton−Robinson buffers at pH 3.0, 5.0, 7.4, 9.0, and 11.0 to the final nominal concentration of 0.2 mg/mL, gently mixed at room temperature for 30 min, and then centrifuged at maximum speed (21,300*g*) for 15 min at 4 °C in order to remove insoluble fraction.

To estimate the effect of 2,2,2-trifluoroethanol on secondary structure, 5 mg/mL suspensions of peptide libraries in Milli-Q water were diluted in a series of 10 mM Britton−Robinson buffer (pH 7.4) containing 0−90% (v/v) 2,2,2-trifluoroethanol.

To estimate the effect of metal ions on secondary structure, 5 mg/mL suspensions of peptide libraries in Milli-Q water were diluted in a series of 10 mM Tris buffer at pH 7.4 supplemented with 0, 10, 100, and 1,000 μM mixtures of NaCl, KCl, $MgCl_2$, $MnCl_2$, and $ZnCl_2$.

Bioinformatic Analysis of in Silico Peptide Libraries. One thousand 25-mer sequences were randomly generated *in silico* for the libraries, and their structures were predicted using PEPstrMOD.[35] PEPstrMOD uses force field libraries Force field_NCAA and Force field_PTM and integrates them in Molecular Dynamics (MD) simulations using AMBER. Additionally, it implements the force field library SwissSideChain with the MD package GROMACS. The parameters employed in the MD simulations were 50 ps of simulation time, and the selected peptide environment was set to "vacuum".

DSSP[60] was employed to determine the secondary structure elements from the peptide tertiary structures generated by PEPstrMOD. The obtained annotations of secondary structure were grouped into three larger classes: helix (H, α helix; G, 310 helix; and I, π-helix), sheet (B, isolated β-bridge and E, extended strand), and loop (T, hydrogen bonded turn; S, bend; and "_" blank spaces).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.2c12987.

Figures of molecular weight distribution, amino acid composition, quantification of free amino acid groups, aggregation propensity, and effect of metal ions on the secondary structure propensities and tables of weights and amounts of protected amino acids, statistical analysis of relative solubilities, and statistical analysis of CD222/CD200 values (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Stephen D. Fried − *Department of Chemistry and T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, United States;* orcid.org/0000-0003-2494-2193; Email: sdfried@jhu.edu

Klara Hlouchova − *Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague 12843, Czech Republic; Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague 16610, Czech Republic;* orcid.org/0000-0002-5651-4874; Email: klara.hlouchova@natur.cuni.cz

Authors

Mikhail Makarov − *Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague 12843, Czech Republic*

Alma C. Sanchez Rocha − *Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague 12843, Czech Republic*

Robin Krystufek − *Department of Physical Chemistry, Faculty of Science, Charles University, Prague 12843, Czech Republic; Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague 16610, Czech Republic;* orcid.org/0000-0003-4096-1159

Ivan Cherepashuk − *Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague 12843, Czech Republic*

Volha Dzmitruk − *Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Vestec 25250, Czech Republic*

Tatsiana Charnavets − *Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Vestec 25250, Czech Republic*
Anneliese M. Faustino − *Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States*
Michal Lebl − *Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague 16610, Czech Republic*
Kosuke Fujishima − *Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo 1528550, Japan; Graduate School of Media and Governance, Keio University, Fujisawa 2520882, Japan;* orcid.org/0000-0002-8844-812X

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.2c12987

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Dr. Vyacheslav Tretyachenko, Dr. Valerio Guido Giacobelli, Dr. Lucie Bednarova, and Prof. Stephen Freeland for helpful discussions about this project. The authors would also like to acknowledge Dr. Radko Souček and Dr. Martin Hubálek for their assistance with amino acid analysis and MALDI analyses, respectively. This work was supported by the Human Frontier Science Program grant HFSP-RGY0074/2019 to K.F., S.D.F., and K.H. M.M., A.C.S.R., and I.C. acknowledge support 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. I.C. is supported by the Visegrad Fund Scholarship (no. 52110039). V.D., T.Ch., K.H., M.M., A.C.S.R., and I.C. acknowledge CMS CF Biophysical techniques of CIISB, Instruct-CZ Centre, supported by MEYS CR Infrastructure project LM2018127 and European Regional Development Fund-Project, "UP CIISB" (No. CZ.02.1.01/0.0/0.0/18_046/0015974). S.D.F. acknowledges support from the NIH Director's New Innovator Award (DP2-GM140926).

■ REFERENCES

(1) Milner-White, E. J.; Russell, M. J. Functional Capabilities of the Earliest Peptides and the Emergence of Life. *Genes (Basel)* **2011**, *2* (4), 671−688.
(2) Fried, S. D.; Fujishima, K.; Makarov, M.; Cherepashuk, I.; Hlouchova, K. Peptides Before and During the Nucleotide World: An Origins Story Emphasizing Cooperation Between Proteins and Nucleic Acids. *J. R. Soc. Interface* **2022**, *19* (187), 20210641.
(3) Frenkel-Pinter, M.; Samanta, M.; Ashkenasy, G.; Leman, L. J. Prebiotic Peptides: Molecular Hubs in the Origin of Life. *Chem. Rev.* **2020**, *120* (11), 4707−4765.
(4) Freeland, S. Undefining Life's Biochemistry: Implications for Abiogenesis. *J. R. Soc. Interface* **2022**, *19* (187), 20210814.
(5) Cleaves, H. J. The Origin of the Biologically Coded Amino Acids. *J. Theor. Biol.* **2010**, *263* (4), 490−498.
(6) Higgs, P. G.; Pudritz, R. E. A Thermodynamic Basis for Prebiotic Amino Acid Synthesis and the Nature of the First Genetic Code. *Astrobiology* **2009**, *9* (5), 483−490.
(7) Kitadai, N.; Maruyama, S. Origins of building blocks of life: A review. *Geosci. Front.* **2018**, *9* (4), 1117−1153.
(8) Weber, A. L.; Miller, S. L. Reasons for the Occurrence of the Twenty Coded Protein Amino Acids. *J. Mol. Evol.* **1981**, *17* (5), 273−284.
(9) Longo, L. M.; Blaber, M. Protein Design at The Interface of The Pre-Biotic and Biotic Worlds. *Arch. Biochem. Biophys.* **2012**, *526* (1), 16−21.
(10) Zaia, D. A. M.; Zaia, C. T. B. V.; De Santana, H. Which Amino Acids Should Be Used in Prebiotic Chemistry Studies? *Orig. Life Evol. Biosph.* **2008**, *38* (6), 469−488.
(11) Ménez, B.; Pisapia, C.; Andreani, M.; Jamme, F.; Vanbellingen, Q. P.; Brunelle, A.; Richard, L.; Dumas, P.; Réfrégiers, M. Abiotic Synthesis of Amino Acids in the Recesses of the Oceanic Lithosphere. *Nature* **2018**, *564* (7734), 59−63.
(12) Patel, B. H.; Percivalle, C.; Ritson, D. J.; Duffy, C. D.; Sutherland, J. D. Common Origins of RNA, Protein and Lipid Precursors in a Cyanosulfidic Protometabolism. *Nat. Chem.* **2015**, *7* (4), 301−307.
(13) Trifonov, E. N. Consensus Temporal Order of Amino Acids and Evolution of The Triplet Code. *Gene.* **2000**, *261* (1), 139−151.
(14) Newton, M. S.; Morrone, D. J.; Lee, K. H.; Seelig, B. Genetic Code Evolution Investigated through the Synthesis and Characterisation of Proteins from Reduced-Alphabet Libraries. *ChemBioChem.* **2019**, *20* (6), 846−856.
(15) Solis, A. D. Reduced Alphabet of Prebiotic Amino Acids Optimally Encodes the Conformational Space of Diverse Extant Protein Folds. *BMC Evol. Biol.* **2019**, *19* (1), 158.
(16) Tretyachenko, V.; Vymětal, J.; Bednárová, L.; Kopecký, V.; Hofbauerová, K.; Jindrová, H.; Hubálek, M.; Souček, R.; Konvalinka, J.; Vondrášek, J.; Hlouchová, K. Random Protein Sequences Can Form Defined Secondary Structures and Are Well-Tolerated In Vivo. *Sci. Rep.* **2017**, *7* (1), 15449.
(17) Freeland, S. Terrestrial" Amino Acids and Their Evolution. *Amino Acids, Peptides and Proteins in Organic Chemistry: Origins and Synthesis of Amino Acids*, 1st ed.; Wiley-VCH Verlag GmbH & Co. KgaA, 2010; Vol. *1*, pp 43−75.
(18) Vázquez-Salazar, A.; Lazcano, A. Early Life: Embracing the RNA World. *Curr. Biol.* **2018**, *28* (5), R220−R222.
(19) Raggi, L.; Bada, J. L.; Lazcano, A. On the Lack of Evolutionary Continuity between Prebiotic Peptides and Extant Enzymes. *Phys. Chem. Chem. Phys.* **2016**, *18* (30), 20028−20032.
(20) Meierhenrich, U. J.; Munoz Caro, G. M.; Bredehoft, J. H.; Jessberger, E. K.; Thiemann, W. H.-P. Identification of Diamino Acids in the Murichison Meteorite. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (25), 9182−9186.
(21) Wolman, Y.; Haverland, W. J.; Miller, S. L. Nonprotein Amino Acids from Spark Discharges and Their Comparison with the Murchison Meteorite Amino Acids. *Proc. Natl. Acad. Sci. U. S. A.* **1972**, *69* (4), 809−811.
(22) Alvarez-Carreño, C.; Becerra, A.; Lazcano, A. Norvaline and Norleucine May Have Been More Abundant Protein Components during Early Stages of Cell Evolution. *Orig. Life Evol. Biosph.* **2013**, *43* (4−5), 363−375.
(23) Apostol, I.; Levine, J.; Lippincott, J.; Leach, J.; Hess, E.; Glascock, C. B.; Weickert, M. J.; Blackmore, R. Incorporation of Norvaline at Leucine Positions in Recombinant Human Hemoglobin Expressed in Escherichia Coli. *J. Biol. Chem.* **1997**, *272* (46), 28980−28988.
(24) Martin, W.; Russell, M. J. On the Origin of Biochemistry at an Alkaline Hydrothermal Vent. *Philos. Trans. R. Soc. London B. Biol. Sci.* **2007**, *362* (1486), 1887−1925.
(25) Grotzinger, J. P.; Kasting, J. F. New Constraints on Precambrian Ocean Composition. *J. Geol.* **1993**, *101* (2), 235−243.
(26) Tanaka, J.; Doi, N.; Takashima, H.; Yanagawa, H. Comparative Characterization of Random-Sequence Proteins Consisting of 5, 12, and 20 Kinds of Amino Acids. *Protein Sci.* **2010**, *19* (4), 786−795.
(27) Tretyachenko, V.; Vymětal, J.; Neuwirthová, T.; Vondrášek, J.; Fujishima, K.; Hlouchová, K. Modern and Prebiotic Amino Acids Support Distinct Structural Profiles in Proteins. *Open Biol.* **2022**, *12* (6), 220040.
(28) Kozlowski, L. P. Proteome-pI: Proteome Isoelectric Point Database. *Nucleic Acids Res.* **2017**, *45* (D1), D1112−D1116.

(29) Wang, Y.; Van Oosterwijk, N.; Ali, A. M.; Adawy, A.; Anindya, A. L.; Dömling, A. S. S.; Groves, M. R. A Systematic Protein Refolding Screen Method using the DGR Approach Reveals that Time and Secondary TSA are Essential Variables. *Sci. Rep.* **2017**, *7* (1), 9355.
(30) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area. *Proc. Natl. Acad. Sci. U. S. A.* **1974**, *71* (8), 2925−2927.
(31) Vymětal, J.; Vondrášek, J.; Hlouchová, K. Sequence Versus Composition: What Prescribes IDP Biophysical Properties? *Entropy (Basel)* **2019**, *21* (7), 654.
(32) Kelly, S. M.; Jess, T. J.; Price, N. C. How to Study Proteins by Circular Dichroism. *Biochim. Biophys. Acta* **2005**, *1751* (2), 119−139.
(33) Buck, M. Trifluoroethanol and Colleagues: Cosolvents Come of Age. Recent Studies with Peptides and Proteins. *Q. Rev. Biophys.* **1998**, *31* (3), 297−355.
(34) Starzyk, A.; Barber-Armstrong, W.; Sridharan, M.; Decatur, S. M. Spectroscopic Evidence for Backbone Desolvation of Helical Peptides by 2,2,2-Trifluoroethanol: An Isotope-Edited FTIR Study. *Biochemistry* **2005**, *44* (1), 369−376.
(35) Singh, S.; Singh, H.; Tuknait, A.; Chaudhary, K.; Singh, B.; Kumaran, S.; Raghava, G. P. PEPstrMOD: Structure Prediction of Peptides Containing Natural, Non-Natural and Modified Residues. *Biol. Direct.* **2015**, *10* (1), 73.
(36) Tanford, C. The Hydrophobic Effect and The Organization of Living Matter. *Science* **1978**, *200* (4345), 1012−1018.
(37) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29* (31), 7133−7155.
(38) Soman, K. V.; Karimi, A.; Case, D. A. Unfolding of an Alpha-helix in Water. *Biopolymers* **1991**, *31*, 1351−1361.
(39) Klotz, I. M.; Franzen, J. S. Hydrogen Bonds between Model Peptide Groups in Solution. *J. Am. Chem. Soc.* **1962**, *84* (18), 3461−3466.
(40) Koh, J. T.; Cornish, V. W.; Schultz, P. G. An Experimental Approach to Evaluating The Role of Backbone Interactions in Proteins using Unnatural Amino Acid Mutagenesis. *Biochemistry* **1997**, *36* (38), 11314−11322.
(41) Baldwin, R. L. In Search of The Energetic Role of Peptide Hydrogen Bonds. *J. Biol. Chem.* **2003**, *278* (20), 17581−17588.
(42) Longo, L. M.; Despotović, D.; Weil-Ktorza, O.; Walker, M. J.; Jabłońska, J.; Fridmann-Sirkis, Y.; Varani, G.; Metanis, N.; Tawfik, D. S. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (27), 15731−15739.
(43) Hay, R. W.; Morris, P. J. Proton Ionisation Constants and Kinetics of Base Hydrolysis of Some α-Amino-acid Esters in Aqueous Solution. Part III. Hydrolysis and Intramolecular Aminolysis of $\alpha\omega$-Diamino-acid Methyl Esters. *J. Chem. Soc., Perkin Trans. 2* **1972**, 1021−1029.
(44) McGee, W. M.; McLuckey, S. A. The Ornithine Effect in Peptide Cation Dissociation. *J. Mass. Spectrom.* **2013**, *48* (7), 856−861.
(45) Despotović, D.; Longo, L. M.; Aharon, E.; Kahana, A.; Scherf, T.; Gruic-Sovulj, I.; Tawfik, D. S. Polyamines Mediate Folding of Primordial Hyperacidic Helical Proteins. *Biochemistry* **2020**, *59* (46), 4456−4462.
(46) Giacobelli, V. G.; Fujishima, K.; Lepšík, M.; Tretyachenko, V.; Kadavá, T.; Makarov, M.; Bednárová, L.; Novák, P.; Hlouchová, K. In Vitro Evolution Reveals Noncationic Protein-RNA Interaction Mediated by Metal Ions. *Mol. Biol. Evol.* **2022**, *39* (3), msac032.
(47) Poduska, K.; Katrukha, G. S.; Silaev, A. B.; Rudinger, J. Amino Acids and Peptides. LII. Intramolecular Aminolysis of Amide Bonds in Derivatives of Alpha,Gamma-Diaminobutyric Acid, Alpha,Beta-Diaminoproprionic Acid, and Ornithine. *Collect. Czechoslov. Chem. Commun.* **1965**, *30*, 2410−2433.
(48) Padmanabhan, S.; York, E. J.; Stewart, J. M.; Baldwin, R. L. Helix Propensities of Basic Amino Acids Increase with the Length of the Side-chain. *J. Mol. Biol.* **1996**, *257*, 726−734.
(49) Mayer-Bacon, C.; Agboha, N.; Muscalli, M.; Freeland, S. Evolution as a Guide to Designing xeno Amino Acid Alphabets. *Int. J. Mol. Sci.* **2021**, *22* (6), 2787.
(50) Brown, S.; Voráček, V.; Freeland, S. What Would an Alien Amino Acid Alphabet Look Like and Why?. *UMBC GABS Symposium 2022*, 2023.
(51) Ilardo, M.; Bose, R.; Meringer, M.; Rasulev, B.; Grefenstette, N.; Stephenson, J.; Freeland, S.; Gillams, R. J.; Butch, C. J.; Cleaves, H. J., 2nd Adaptive Properties of the Genetically Encoded Amino Acid Alphabet Are Inherited from Its Subsets. *Sci. Rep.* **2019**, *9* (1), 12468.
(52) Ilardo, M.; Meringer, M.; Freeland, S.; Rasulev, B.; Cleaves, H. J., 2nd Extraordinarily Adaptive Properties of the Genetically Encoded Amino Acids. *Sci. Rep.* **2015**, *5*, 9414.
(53) Lam, K. S.; Lebl, M.; Krchňák, V. The "One-Bead-One-Compound" Combinatorial Library Method. *Chem. Rev.* **1997**, *97* (2), 411−448.
(54) Ostresh, J. M.; Winkle, J. H.; Hamashin, V. T.; Houghten, R. A. Peptide Libraries: Determination of Relative Reaction Rates of Protected Amino Acids in Competitive Couplings. *Biopolymers* **1994**, *34* (12), 1681−1689.
(55) Rutter, W.; Santi, D. Peptide Mixtures. US5420246A, 1995.
(56) Ivanetich, K. M.; Santi, D. V. Preparation of Equimolar Mixtures of Peptides by Adjustment of Activated Amino Acid Concentrations. *Methods Enzymol.* **1996**, *267*, 247−260.
(57) Lebl, M.; Flegelova, Z.; Poncar, P.; Mudra, P.; Paces, O.; Knor, M.; Buzek, M.; Smrz, J.; Pokorny, V.; Pesek, V. Device for Parallel Oligomer Synthesis, Method of Parallel Oligomer Synthesis and Use Thereof. WO-2019115514-A1, 2017.
(58) King, D. S.; Fields, C. G.; Fields, G. B. A Cleavage Method Which Minimizes Side Reactions Following Fmoc Solid Phase Peptide Synthesis. *Int. J. Pept. Protein Res.* **1990**, *36* (3), 255−266.
(59) Udenfriend, S.; Stein, S.; Böhlen, P.; Dairman, W.; Leimgruber, W.; Weigele, M. Fluorescamine: a Reagent for Assay of Amino Acids, Peptides, Proteins, and Primary Amines in The Picomole Range. *Science* **1972**, *178* (4063), 871−872.
(60) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.

https://doi.org/10.1021/jacs.2c12987
J. Am. Chem. Soc. 2023, 145, 5320−5329

Supplementary Information

Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability

Mikhail Makarov[1], Alma C. Sanchez Rocha[1], Robin Krystufek[2,3], Ivan Cherepashuk[1], Volha Dzmitruk[4], Tatsiana Charnavets[4], Anneliese M. Faustino,[5] Michal Lebl[3], Kosuke Fujishima[6,7], Stephen D. Fried*[5,8], and Klara Hlouchova*[1,3]

[1]*Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague, 12843, Czech Republic*

[2]*Department of Physical Chemistry, Faculty of Science, Charles University, Prague, 12843, Czech Republic*

[3]*Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, 16610, Czech Republic*

[4]*Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, 25250 Vestec, Czech Republic*

[5]*Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21212, United States*

[6]*Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, 1528550, Japan*

[7]*Graduate School of Media and Governance, Keio University, Fujisawa, 2520882, Japan*

[8]*Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21212, United States*

* To whom correspondence may be addressed: klara.hlouchova@natur.cuni.cz, sdfried@jhu.edu

Table of contents

Figure S1. Molecular weight distribution of 25-mer combinatorial peptide libraries 1

Figure S2. Amino acid composition of 25-mer combinatorial peptide libraries estimated by HPLC amino acid analysis. 2

Figure S3. The quantification of free amino acid groups in 10E and 11D peptide libraries using fluorescamine assay 3

Figure S4. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and low ionic strength 4

Figure S5. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and high ionic strength 5

Figure S6. The effect of metal ions on the secondary structure of 25-mer combinatorial peptide libraries.....6

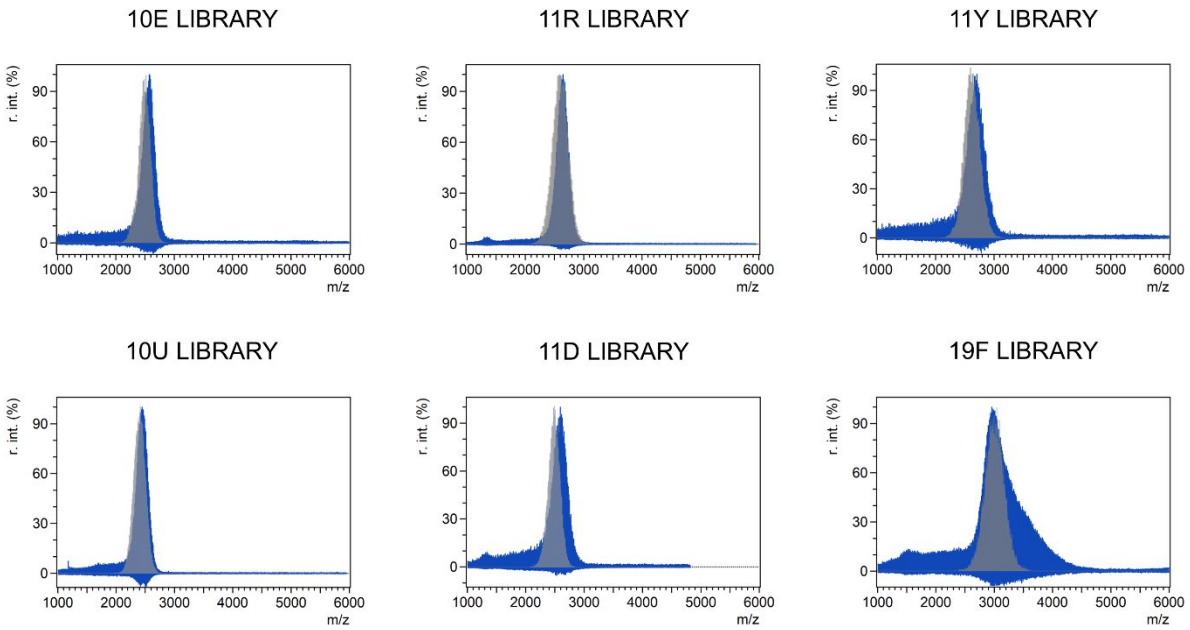
Table S1: Weights (mg) of protected amino acids in individual isokinetic mixtures used for synthesis 7

Supplementary Information

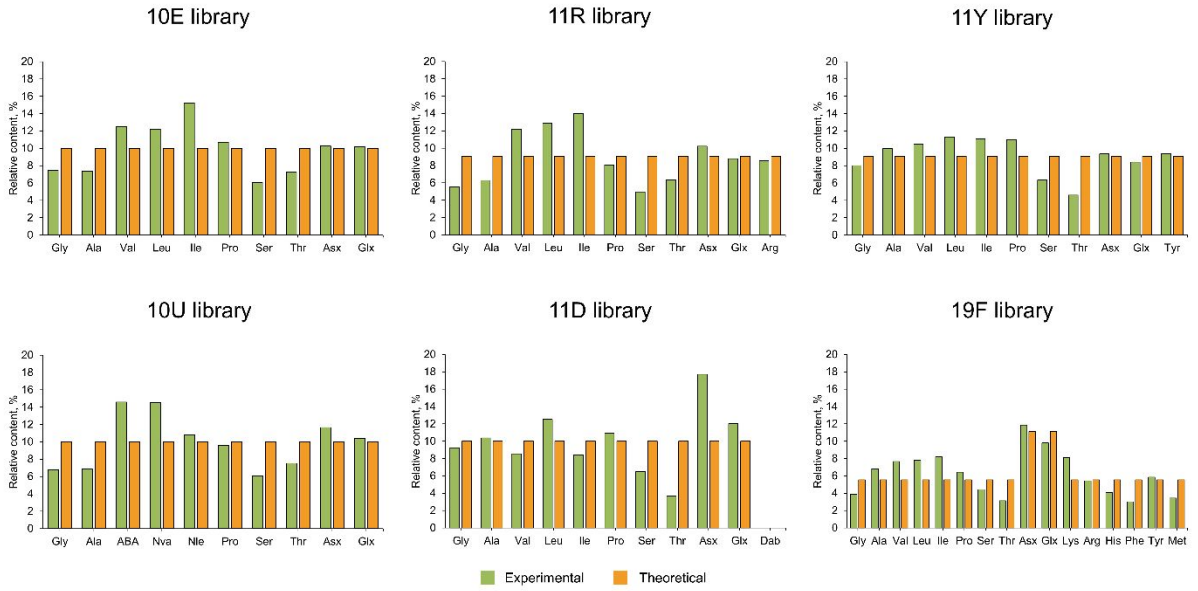
Table S2: Statistical analysis of relative solubilities at different pH and ionic strengths……8

Table S3: Statistical analysis of CD222/CD200 values at different pH values…….……12

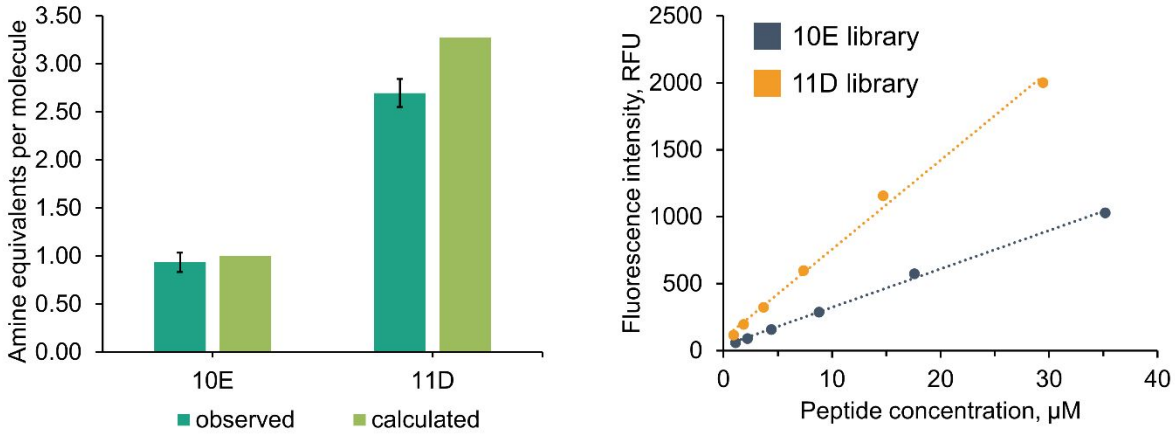
Table S4: Statistical analysis of CD222/CD200 values at different 2,2,2-trifluoroethanol concentrations……..14



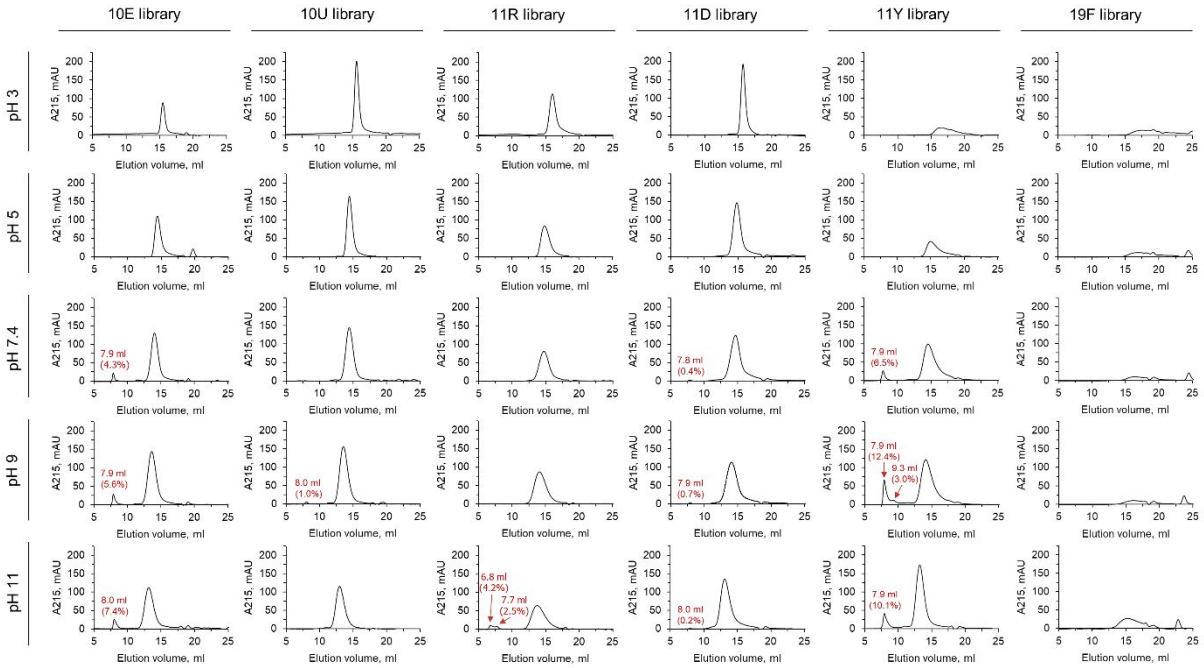
Supplementary Figure S1. Molecular weight distribution of 25-mer combinatorial peptide libraries estimated by MALDI-TOF-MS using 2,5-dihydrobenzoic acid (DHB) matrix. Observed m/z distribution is shown in blue, expected m/z distribution is shown as a gray overlay. Peptide libraries were dissolved in acetonitrile:water (1:1) mixture + 0.1 % (v/v) acetic acid. 19F peptide library was dissolved in acetonitrile:water (1:1) mixture.



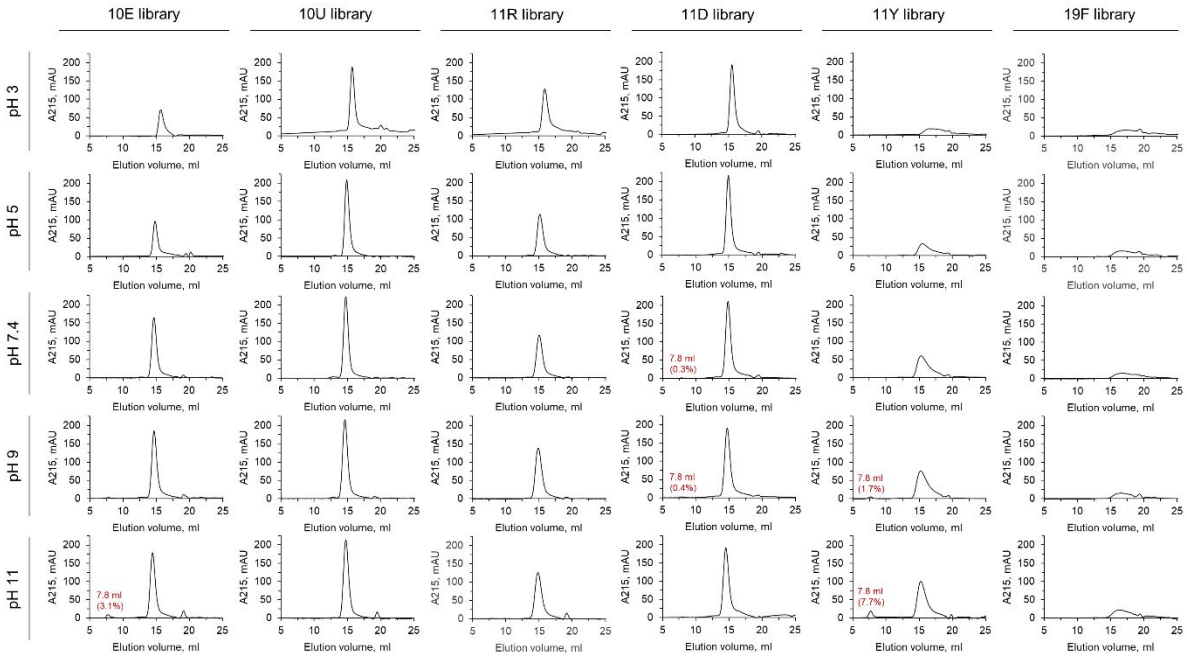
Supplementary Figure S2. Amino acid composition of 25-mer combinatorial peptide libraries determined by amino acid analysis.



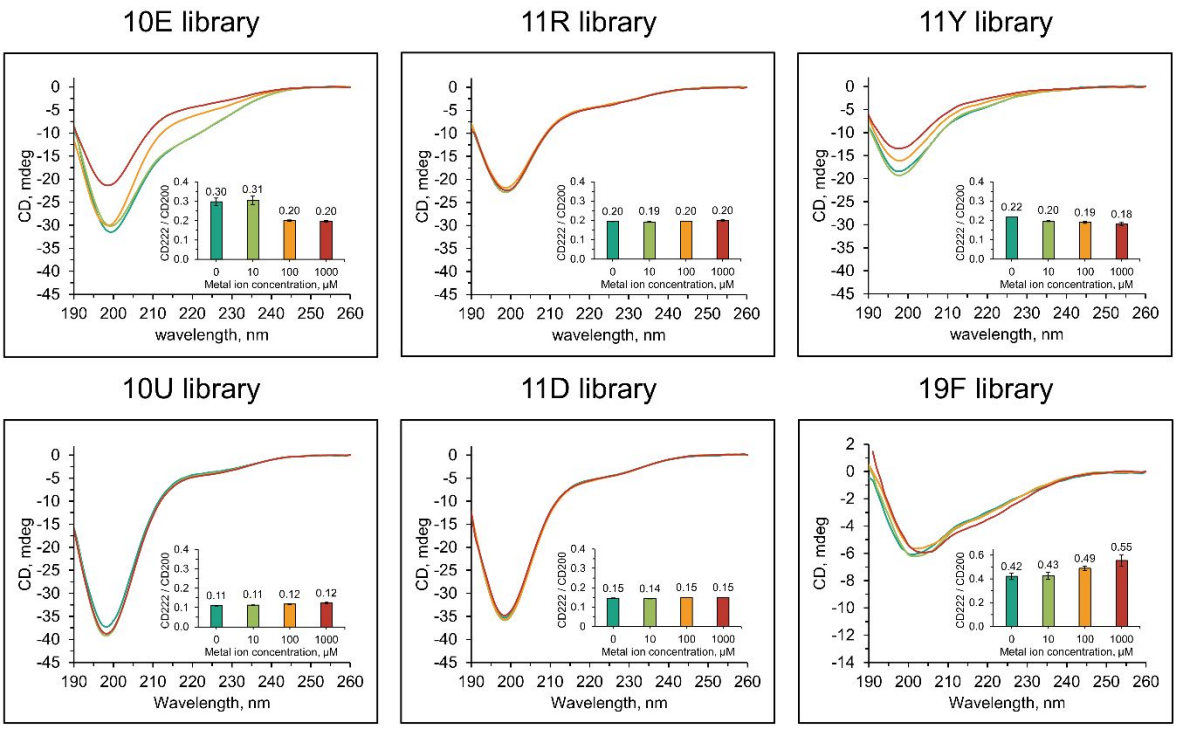
Supplementary Figure S3. The quantification of free amino acid groups in 10E and 11D peptide libraries using fluorescamine assay.



Supplementary Figure S4. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and low ionic strength (50 mM NaCl). Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in a series of 20 mM ABP buffers (pH 3–11) by size-exclusion chromatography.



Supplementary Figure S5. Aggregation propensity of 25-mer combinatorial peptide libraries at different pH and high ionic strength (500 mM NaCl). Aggregation propensity was measured for 0.5 mg/ml nominal peptide library solutions in a series of 20 mM ABP buffers (pH 3–11) by size-exclusion chromatography.



Supplementary Figure S6. The effect of metal ions on the secondary structure of 25-mer combinatorial peptide libraries. CD spectra were measured for 0.2 mg/ml nominal peptide library solutions in a series of 10 mM Tris buffers at pH 7.4 supplemented with 0 (green line); 10 (yellow line); 100 (orange line); and 1,000 (red line) μM mixture of NaCl, KCl, $MgCl_2$, $MnCl_2$, and $ZnCl_2$. The inlet graph shows the ratios of CD signal at 222 to 200 nm.

Supplementary Table S1: Weights (mg) and amounts (mmol) of protected amino acids in individual isokinetic mixtures used for synthesis. Mixtures were dissolved in a final volume listed in the last row yielding 0.3 M solution of amino acids in 0.3 M HOBt in DMF.

	19F		10E		11R		11Y		10U		11D	
	m mg	n mmol	m mg	n mmol	m mg	n mmol	m mg	n mmol	m mg	n mmol	m mg	n mmol
Fmoc-Ala	202	0.61	332	1.01	300	0.91	312	0.95	508	1.54	303	0.92
Fmoc-Asp(OtBu)	276	0.67	455	1.11	409	0.99	426	1.04	694	1.69	414	1.01
Fmoc-Glu(OtBu)	296	0.67	489	1.10	440	0.99	457	1.03	745	1.68	444	1.00
Fmoc-Phe	187	0.48										
Fmoc-Gly	164	0.55	270	0.91	243	0.82	253	0.85	412	1.39	246	0.83
Fmoc-His(Trt)	422	0.68										
Fmoc-Ile	1,171	3.31	1,934	5.47	1,740	4.92	1,809	5.12			1,759	4.98
Fmoc-Lys(Boc)	556	1.19										
Fmoc-Leu	335	0.95	553	1.56	497	1.41	517	1.46			503	1.42
Fmoc-Met	163	0.44										
Fmoc-Asn(Trt)	609	1.02										
Fmoc-Pro	278	0.78	460	1.29	414	1.16	430	1.21	702	1.98	418	1.18
Fmoc-Gln(Trt)	620	1.02										
Fmoc-Arg(Pmc)	807	1.22			1,199	1.81						
Fmoc-Ser(tBu)	203	0.53	336	0.88	302	0.79	314	0.82	512	1.34	306	0.80
Fmoc-Thr(tBu)	363	0.91	599	1.51	539	1.36	560	1.41	914	2.30	545	1.37
Fmoc-Val	731	2.15	1,208	3.56	1,086	3.20	1,130	3.33			1,099	3.24
Fmoc-Trp(Boc)	254	0.48										
Fmoc-Tyr(tBu)	363	0.79					560	1.22				
Fmoc-Nva									752	2.22		
Fmoc-Nle									775	2.19		
Fmoc-Aba									648	1.99		
Fmoc-Dab(Boc)											734	1.67
V / ml	61.5		61.3		61.2		61.4		39.7		55.8	

Supplementary table S2. Statistical analysis of relative solubilities at different pH and ionic strengths. The tables include ratios of relative solubilities and P-values. P-values are calculated using a two-tailed Student's t-test and assuming homoscedasticity.

(a) Statistical analysis of solubilities at low (50 mM NaCl) vs high (500 mM NaCl) ionic strengths

		10E	10U	11R	11D	11Y	19F
pH 3.0	Ratio	1.0264	1.0285	1.1484	1.0682	1.0560	1.4677
	P-value	0.3448	0.3946	0.0564	0.0214	0.4727	0.0172
pH 5.0	Ratio	0.9580	1.0137	0.9736	0.9656	0.9895	1.4433
	P-value	0.6268	0.6617	0.7776	0.0395	0.8594	0.0226
pH 7.4	Ratio	1.2833	0.9966	1.0013	0.9875	1.6230	1.0857
	P-value	0.0002	0.8585	0.9781	0.4112	0.0031	0.2538
pH 9.0	Ratio	0.9979	1.0239	1.1019	0.9945	1.4544	1.0753
	P-value	0.8370	0.1289	0.0263	0.4117	0.0160	0.5570
pH 11.0	Ratio	0.9881	0.9957	1.3588	1.0231	1.1160	0.9891
	P-value	0.4577	0.6999	2.1E-5	0.0546	0.4751	0.8085

(b) Statistical analysis of solubilities at low (50 mM NaCl) ionic strength at pH 3.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.4673					
	P-value	2.1E-5					
11R	Ratio	0.5790	1.2391				
	P-value	4.0E-5	0.0028				
11D	Ratio	0.3778	0.8085	0.6524			
	P-value	3.3E-7	0.0007	3.1E-5			
11Y	Ratio	2.2921	4.9050	3.9584	6.0671		
	P-value	4.5E-6	4.7E-6	5.0E-6	1.5E-7		
19F	Ratio	0.9739	2.0842	1.6820	2.5780	0.4249	
	P-value	0.6288	0.0002	0.0007	2.0E-5	0.0008	

(c) Statistical analysis of solubilities at low (50 mM NaCl) ionic strength at pH 5.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.4997					
	P-value	0.0001					
11R	Ratio	0.8370	1.6750				
	P-value	0.0958	0.0008				
11D	Ratio	0.4947	0.9901	0.5911			
	P-value	4.9E-5	0.5904	0.0004			
11Y	Ratio	2.0150	4.0324	2.4075	4.0729		
	P-value	0.0009	4.9E-6	0.0008	2.5E-7		
19F	Ratio	1.3055	2.6127	1.5598	2.6389	0.6479	
	P-value	0.0309	3.4E-5	0.0081	1.0E-5	0.0044	

(d) Statistical analysis of solubilities at low (50 mM NaCl) ionic strength at pH 7.4

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.0024					
	P-value	0.9157					
11R	Ratio	1.4744	1.4709				
	P-value	0.0001	0.0001				
11D	Ratio	1.0098	1.0074	0.6849			
	P-value	0.6847	0.7210	0.0001			
11Y	Ratio	1.1801	1.1773	0.8004	1.1687		
	P-value	0.0025	0.0021	0.0052	0.0022		
19F	Ratio	4.1326	4.1229	2.8029	4.0927	3.5020	
	P-value	1.8E-6	1.1E-6	3.8E-5	1.0E-6	1.6E-5	

(e) Statistical analysis of solubilities at low (50 mM NaCl) ionic strength at pH 9.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.9998					
	P-value	0.9490					
11R	Ratio	1.2767	1.2769				
	P-value	4.6E-5	2.3E-5				
11D	Ratio	1.0094	1.0096	0.7907			
	P-value	0.5400	0.3905	6.7E-6			
11Y	Ratio	0.9979	0.9982	0.7817	0.9886		
	P-value	0.6588	0.6219	0.0006	0.8537		
19F	Ratio	3.9955	3.9963	3.1296	3.9582	4.0037	
	P-value	9.5E-6	8.5E-6	3.1E-5	7.8E-6	2.0E-5	

(f) Statistical analysis of solubilities at low (50 mM NaCl) ionic strength at pH 11.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.9840					
	P-value	0.3263					
11R	Ratio	0.9798	0.9957				
	P-value	0.0895	0.2452				
11D	Ratio	0.9798	0.9957	1.0000			
	P-value	0.1452	0.4924	0.2810			
11Y	Ratio	1.0313	1.0480	1.0526	1.0526		
	P-value	0.0860	0.0213	0.0070	0.0099		
19F	Ratio	3.4413	3.4971	3.5124	3.5124	3.3370	
	P-value	2.0E-6	8.9E-7	3.7E-7	4.5E-7	3.2E-6	

(g) Statistical analysis of solubilities at high (500 mM NaCl) ionic strength at pH 3.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.4683					
	P-value	1.6E-5					
11R	Ratio	0.6478	1.3835				
	P-value	0.0024	0.0027				
11D	Ratio	0.3932	0.8397	0.6069			
	P-value	4.9E-6	0.0019	0.0003			
11Y	Ratio	2.3581	5.0360	3.6400	5.9976		
	P-value	7.3E-5	3.5E-6	0.0002	1.5E-6		
19F	Ratio	1.3926	2.9741	2.1497	3.5420	0.5906	
	P-value	0.0096	3.1E-5	0.0010	1.1E-5	0.0073	

(h) Statistical analysis of solubilities at high (500 mM NaCl) ionic strength at pH 5.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.5287					
	P-value	0.0002					
11R	Ratio	0.8506	1.6087				
	P-value	0.1249	0.0010				
11D	Ratio	0.4986	0.9431	0.5862			
	P-value	9.6E-5	0.0519	0.0004			
11Y	Ratio	2.0812	3.9362	2.4468	4.1737		
	P-value	0.0012	4.8E-6	0.0007	6.7E-7		
19F	Ratio	1.9669	3.7201	2.3124	3.9446	0.9451	
	P-value	0.0031	2.1E-5	0.0015	7.9E-6	0.4946	

(i) Statistical analysis of solubilities at high (500 mM NaCl) ionic strength at pH 7.4

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.7784					
	P-value	0.0003					
11R	Ratio	1.1505	1.4779				
	P-value	0.0209	0.0004				
11D	Ratio	0.7770	0.9982	0.6754			
	P-value	9.2E-5	0.8350	0.0003			
11Y	Ratio	1.4926	1.9174	1.2974	1.9208		
	P-value	0.0050	0.0006	0.0330	0.0005		
19F	Ratio	3.4963	4.4915	3.0391	4.4995	2.3425	
	P-value	1.6E-6	1.4E-6	5.7E-5	4.5E-7	0.0034	

(j) Statistical analysis of solubilities at high (500 mM NaCl) ionic strength at pH 9.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.0257					
	P-value	0.0698					
11R	Ratio	1.4097	1.3743				
	P-value	0.0001	0.0003				
11D	Ratio	1.0059	0.9807	0.7136			
	P-value	0.6154	0.1297	0.0002			
11Y	Ratio	1.4544	1.4179	1.0318	1.4459		
	P-value	0.0125	0.0168	0.6691	0.0130		
19F	Ratio	4.3053	4.1973	3.0542	4.2800	2.9602	
	P-value	3.5E-7	1.1E-6	3.4E-5	5.1E-7	0.0041	

(k) Statistical analysis of solubilities at high (500 mM NaCl) ionic strength at pH 11.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	0.9916					
	P-value	0.4210					
11R	Ratio	1.3474	1.3588				
	P-value	5.4E-5	4.1E-5				
11D	Ratio	1.0145	1.0231	0.7529			
	P-value	0.3492	0.1158	6.6E-5			
11Y	Ratio	1.1647	1.1746	0.8644	1.1481		
	P-value	0.2884	0.2646	0.4659	0.3213		
19F	Ratio	3.4451	3.4743	2.5568	3.3959	2.9578	
	P-value	1.4E-7	9.6E-8	3.3E-6	1.6E-7	0.0111	

Supplementary table S3. Statistical analysis of CD222/CD200 values at different pH values. The tables include ratios of CD222/CD200 values and P-values. P-values are calculated using a two-tailed Student's t-test and assuming homoscedasticity.

(a) Statistical analysis of CD222/CD200 values at pH 3.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.4454					
	P-value	0.0102					
11R	Ratio	1.0564	0.7309				
	P-value	0.2795	0.0054				
11D	Ratio	1.3374	0.9253	1.2660			
	P-value	0.0159	0.0213	0.0102			
11Y	Ratio	0.9235	0.6389	0.8741	0.6905		
	P-value	0.2158	0.0076	0.0730	0.7528		
19F	Ratio	0.7829	0.5416	0.7410	0.5853	0.8477	
	P-value	0.0796	0.0170	0.0531	0.9173	0.1479	

(b) Statistical analysis of CD222/CD200 values at pH 5.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.5078					
	P-value	0.0059					
11R	Ratio	1.0490	0.6957				
	P-value	0.1904	0.0012				
11D	Ratio	1.3382	0.8875	1.2757			
	P-value	0.0101	0.0249	0.0019			
11Y	Ratio	1.0940	0.7255	1.0429	0.8175		
	P-value	0.0928	0.0051	0.1186	0.5937		
19F	Ratio	0.5058	0.3354	0.4822	0.3379	0.4623	
	P-value	0.0760	0.0440	0.0695	0.6946	0.0651	

(c) Statistical analysis of CD222/CD200 values at pH 7.4

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	3.0516					
	P-value	0.0004					
11R	Ratio	1.7835	0.5844				
	P-value	0.0017	0.0056				
11D	Ratio	2.4441	0.8009	1.3704			
	P-value	0.0003	0.0118	0.0101			
11Y	Ratio	1.2538	0.4109	0.7030	0.5130		
	P-value	0.0037	0.0006	0.0052	0.9884		
19F	Ratio	1.5215	0.4986	0.8531	0.6225	1.2135	
	P-value	0.0122	0.0130	0.1373	0.8600	0.0655	

(d) Statistical analysis of CD222/CD200 values at pH 9.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.3852					
	P-value	0.0006					
11R	Ratio	0.9834	0.4123				
	P-value	0.2634	0.0002				
11D	Ratio	1.9297	0.8091	1.9623			
	P-value	0.0005	0.0096	3.0E-5			
11Y	Ratio	0.9257	0.3881	0.9413	0.4797		
	P-value	0.3894	0.0121	0.4754	0.9183		
19F	Ratio	0.9682	0.4059	0.9845	0.5017	1.0459	
	P-value	0.2465	0.0010	0.4470	0.9906	0.5910	

(e) Statistical analysis of CD222/CD200 values at pH 11.0

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.2522					
	P-value	0.0003					
11R	Ratio	0.7160	0.3179				
	P-value	0.0240	0.0043				
11D	Ratio	1.9858	0.8817	2.7734			
	P-value	0.0002	0.0204	0.0049			
11Y	Ratio	0.9523	0.4228	1.3301	0.4796		
	P-value	0.0564	0.0005	0.0317	0.9635		
19F	Ratio	0.9157	0.4066	1.2790	0.4611	0.9616	
	P-value	0.0088	0.0002	0.0397	0.9408	0.0774	

Supplementary table S4. Statistical analysis of CD222/CD200 values at different 2,2,2-trifluoroethanol concentrations. The tables include ratios of CD222/CD200 values and P-values. P-values are calculated using a two-tailed Student's t-test and assuming homoscedasticity.

(a) Statistical analysis of CD222/CD200 values at 0% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	3.0516					
	P-value	0.0004					
11R	Ratio	1.7910	0.5869				
	P-value	0.0013	0.0045				
11D	Ratio	2.4441	0.8009	1.3647			
	P-value	0.0003	0.0118	0.0075			
11Y	Ratio	1.2538	0.4109	0.7000	0.5130		
	P-value	0.0037	0.0006	0.0039	0.9884		
19F	Ratio	1.6823	0.5513	0.9393	0.6883	1.3417	
	P-value	0.0560	0.1169	0.7543	0.8254	0.1796	

(b) Statistical analysis of CD222/CD200 values at 10% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	3.3216					
	P-value	0.0005					
11R	Ratio	2.0610	0.6205				
	P-value	0.0030	0.0154				
11D	Ratio	2.8413	0.8554	1.3786			
	P-value	0.0006	0.0032	0.0292			
11Y	Ratio	1.2974	0.3906	0.6295	0.4566		
	P-value	0.0090	0.0010	0.0092	0.9237		
19F	Ratio	1.6901	0.5088	0.8201	0.5948	1.3027	
	P-value	0.0027	0.0023	0.0588	0.8806	0.0128	

(c) Statistical analysis of CD222/CD200 values at 20% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.6675					
	P-value	0.0001					
11R	Ratio	1.6638	0.6237				
	P-value	0.0006	0.0025				
11D	Ratio	2.3442	0.8788	1.4089			
	P-value	0.0002	0.0341	0.0047			
11Y	Ratio	0.8460	0.3171	0.5085	0.3609		
	P-value	0.0032	0.0002	0.0006	0.7995		
19F	Ratio	0.6663	0.2498	0.4005	0.2843	0.7876	
	P-value	0.0005	0.0001	0.0003	0.7014	0.0023	

(d) Statistical analysis of CD222/CD200 values at 30% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.2870					
	P-value	4.9E-5					
11R	Ratio	1.2366	0.5407				
	P-value	0.0074	0.0021				
11D	Ratio	1.9516	0.8534	1.5782			
	P-value	0.0003	0.0138	0.0038			
11Y	Ratio	0.6514	0.2849	0.5268	0.3338		
	P-value	0.0007	0.0002	0.0009	0.7630		
19F	Ratio	0.2595	0.1134	0.2098	0.1330	0.3983	
	P-value	0.0097	0.0068	0.0085	0.4754	0.0146	

(e) Statistical analysis of CD222/CD200 values at 40% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.1465					
	P-value	0.0015					
11R	Ratio	1.0300	0.4799				
	P-value	0.5512	0.0055				
11D	Ratio	1.8074	0.8420	1.7547			
	P-value	0.0019	0.0123	0.0076			
11Y	Ratio	0.5943	0.2769	0.5770	0.3288		
	P-value	0.0032	0.0008	0.0048	0.7472		
19F	Ratio	0.1359	0.0633	0.1320	0.0752	0.2287	
	P-value	0.0184	0.0157	0.0182	0.3865	0.0229	

(f) Statistical analysis of CD222/CD200 values at 50% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	2.0240					
	P-value	0.0010					
11R	Ratio	0.9415	0.4652				
	P-value	0.2231	0.0036				
11D	Ratio	1.6898	0.8349	1.7948			
	P-value	0.0014	0.0136	0.0052			
11Y	Ratio	0.5706	0.2819	0.6060	0.3377		
	P-value	0.0007	0.0002	0.0027	0.7691		
19F	Ratio	0.1174	0.0580	0.1246	0.0694	0.2057	
	P-value	0.0008	0.0007	0.0008	0.4596	0.0010	

(g) Statistical analysis of CD222/CD200 values at 60% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.9533					
	P-value	0.0010					
11R	Ratio	0.9008	0.4612				
	P-value	0.0515	0.0016				
11D	Ratio	1.6126	0.8256	1.7901			
	P-value	0.0019	0.0146	0.0026			
11Y	Ratio	0.5535	0.2834	0.6145	0.3433		
	P-value	0.0003	4.2E-5	0.0011	0.7827		
19F	Ratio	0.1113	0.0570	0.1236	0.0690	0.2011	
	P-value	8.0E-6	5.7E-6	1.4E-5	0.4779	6.7E-6	

(h) Statistical analysis of CD222/CD200 values at 70% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.8746					
	P-value	0.0025					
11R	Ratio	0.8534	0.4553				
	P-value	0.0475	0.0029				
11D	Ratio	1.5377	0.8203	1.8019			
	P-value	0.0037	0.0106	0.0041			
11Y	Ratio	0.5533	0.2952	0.6483	0.3598		
	P-value	0.0007	0.0001	0.0028	0.8018		
19F	Ratio	0.0961	0.0513	0.1126	0.0625	0.1736	
	P-value	0.0049	0.0044	0.0051	0.4219	0.0058	

(i) Statistical analysis of CD222/CD200 values at 80% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.7646					
	P-value	0.0070					
11R	Ratio	0.8485	0.4809				
	P-value	0.0632	0.0039				
11D	Ratio	1.4690	0.8325	1.7312			
	P-value	0.0104	0.0245	0.0049			
11Y	Ratio	0.5381	0.3049	0.6341	0.3663		
	P-value	0.0014	0.0002	0.0026	0.8132		
19F	Ratio	0.0734	0.0416	0.0864	0.0499	0.1363	
	P-value	0.0011	0.0010	0.0011	0.4384	0.0012	

(j) Statistical analysis of CD222/CD200 values at 90% (v/v) 2,2,2-trifluoroethanol

		10E	10U	11R	11D	11Y	19F
10E	Ratio						
	P-value						
10U	Ratio	1.6794					
	P-value	0.0091					
11R	Ratio	0.8393	0.4997				
	P-value	0.0550	0.0029				
11D	Ratio	1.3614	0.8106	1.6221			
	P-value	0.0213	0.0201	0.0052			
11Y	Ratio	0.5214	0.3105	0.6213	0.3830		
	P-value	0.0143	0.0063	0.0217	0.7891		
19F	Ratio	0.0184	0.0110	0.0220	0.0135	0.0353	
	P-value	0.0270	0.0266	0.0272	0.3064	0.0279	

Self-assembly of random primordial acidic peptides as a steppingstone to foldable polymers

Mikhail Makarov[1], Robin Kryštůfek[2], Edgar Manriquez-Sandoval[3], Soumya Dutta[4], Tadeáš Kalvoda[2], Ján Michael Kormaník[2], Tatsiana Charnavets[5], Michal Lebl[2], Sean M. Brown[6], Jan Konvalinka[2], Abhishek Singharoy[4], Stephen D. Fried*[3,7], Klára Hlouchová*[1,2]

[1] *Department of Cell Biology, Faculty of Science, Charles University, Prague, 12800, Czech Republic*

[2] *Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, 16610, Czech Republic*

[3] *T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA*

[4] *School of Molecular Sciences, Arizona State University, Tempe, AZ, USA*

[5] *Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Vestec, 25250, Czech Republic*

[6] *Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD 21250, USA*

[7] *Department of Chemistry, Johns Hopkins University, Baltimore, MD 21218, USA*

* Correspondence to: sdfried@jhu.edu, klara.hlouchova@natur.cuni.cz

Highlights

- A prebiotically plausible acidic amino acid alphabet stands out in its capacity to promote self-assembly and folding.
- Any inclusion of basic amino acid residues, both canonical and prebiotic, either in isolation or alongside acidic residues, abrogates folding propensity.
- Observation of soluble oligomeric β-sheet assemblies arising from random peptide sequences comprising early amino acids

Summary

Modern proteins are remarkable polymers built from a 20-amino-acid alphabet, shaped by billions of years of evolution. However, some amino acids were likely scarce during Earth's prebiotic stages. In particular, the roles of charged residues are unclear since the canonical basic amino acids (lysine, arginine, and histidine) were unavailable, despite their key role in many biochemical functions. Here, we synthesize random peptide libraries that span a broad charge spectrum and systematically interrogate their properties. The data indicate that a prebiotically plausible acidic alphabet stands out in its propensity for secondary structure and higher-order soluble assembly via formation of β-sheets. In fact, any inclusion of basic residues – both canonical and prebiotic, in isolation or alongside acidic residues – abrogates such properties. This study showcases that unevolved sequences of prebiotically-abundant amino acids can readily produce foldable self-assembling polymers, potentially providing a steppingstone toward the first proteins, prior to the onset of purifying selection.

Introduction

Biochemistry researchers and students alike have long asked, Why did life since the last universal common ancestor (LUCA) 'select' the twenty L-α-amino acids as the established language of building proteins and peptides?[1,2,3,4,5] The groundbreaking work of Miller and Urey[6,7,8] in the 1950s provided early evidence that unsupervised gas phase reactions under reducing conditions could have provided abundant access to some α-amino acids early in Earth's history. Even at this incipient stage in the field of prebiotic chemistry, an apparent asymmetry was noted: the modern acidic amino acids – L-aspartic acid and L-glutamic acid – were generated in high yields, whereas the modern basic amino acids – L-arginine, L-histidine, and L-lysine – were not detected. The absence of these essential building blocks was rightly perceived as a roadblock for an early emergence of functional proteins[2], given that extant proteins are comprised of a balanced mix of acidic and basic residues, and basic residues are almost always involved in interactions with nucleotides and nucleic acids. Analyses of the amino acids that are carried to Earth by carbonaceous meteorites[9] have only reinforced the view that the modern acidic amino acids were plentiful whereas the modern basic amino acids were unavailable in prebiotic sources. The overlap between the amino acid compositions from gas-phase atmospheric chemistry and meteorites have led to propositions of a consensus 'early' amino acid alphabet[10] that comprised of ten L-α-amino acids: alanine (Ala), aspartic acid (Asp), glutamic acid (Glu), glycine (Gly), leucine (Leu), isoleucine (Ile), proline (Pro), serine (Ser), threonine (Thr), and valine (Val) (Fig. 1, dark blue and red circles).

Whether such a reduced chemical repertoire could have served as a practical inventory for building up the earliest biopolymers is not apparent, given the essential roles aromatic, basic, and sulfur-containing amino acids play in modern protein biochemistry. Recent efforts from us and others[11,12] have at least substantiated this possibility by showing that these 'early acidic' proteins can adopt conformations with secondary structure, can form strong associations with RNA mediated by divalent cation coordination[13], and strikingly, are much less dependent on chaperones to adopt soluble stable conformations[12,14,15].

On the other hand, firm claims are hard to make in prebiotic chemistry because one must always be cognizant of the large space of potential alternative scenarios that are not reflected in extant biology, but which nevertheless might have taken place or been significant in the distant past. For instance, even though there is spare evidence for L-lysine and L-arginine in prebiotic sources, modest levels of L-2,3-diaminopropionic acid (Dap), and L-2,4-diaminobutyric acid (Dab) (Fig. 1, cyan circles) are detected in meteorite sources among many other noncanonical amino acids[16]. It is possible that these noncanonical basic amino acids played a key role in an earlier phase of biochemical history, only later to be displaced by their peers with longer side chains. The existing biological record provides no way to disambiguate between these possibilities.

Recently, we proposed that foldability could have provided a selective force to govern which among the vast array of available amino acids became promoted to the biotic pool[17]. The model is premised on the fact that now many prebiotically-plausible routes to condense α-amino acids into peptides have been documented (e.g., COS-catalysis[18], wet-dry cycling[19], thiol catalysis[20,21]). Building blocks that provided peptides with a propensity to fold/compact or oligomerize/assemble would have been adaptive by generating sequences with greater stability and function. On the other hand, monomers that impair foldability would have been selected against. This model provides a satisfying explanation for why the canonical alphabet contains branched aliphatic amino acids (Val, Leu, Ile) even though their unbranched isomers

(norvaline, norleucine) were possibly more prebiotically abundant: the former are more adept at producing secondary structures than the latter (especially β-sheets). Whether foldability could have more broadly steered the early evolution of biopolymers, and how adaptive any proposed early amino acid alphabet is in relation to alternatives has not been presently established.

Here, we extensively characterise nine random peptide libraries (mixtures of sequences), each composed of a different alphabet. Each alphabet consists of various combinations of acidic and basic amino acids (some canonical, some prebiotic, and some counterfactual), to interrogate effects on solubility, folding propensity, and oligomerisation potential. The data shows that an acidic prebiotically-plausible alphabet is remarkable in its capacity to promote folding as well as oligomer assembly mediated through β-sheets. We find unambiguously that any inclusion of basic amino acid residues – either in isolation or alongside acidic residues – diminishes efficient folding and assembly. These experimental findings are supported by a large language model for protein structure prediction (ESMFold[22]). Altogether, these findings point to acidic polypeptides as a potential intermediate in the early evolution of peptides and proteins. Though alien in some ways compared to today's charge-balanced polypeptides, the early acidic alphabet could provide key advantages that facilitated the initial emergence of biopolymers.

Results

Synthesis of Charge-Variant Libraries. Libraries of 25-mer peptides, constructed as random combinations of amino acids within the library, were synthesised by standard Fmoc-based solid phase peptide synthesis using the isokinetic mixtures as previously reported[17] (see Methods and Supplemental Data S1). Each library sampled one of nine different possible alphabets (Fig. 1A), which could comprise either 8, 10, or 12 constituent L-α-amino acids. The common core of all libraries (library "8E") only uses the eight uncharged canonical and prebiotic amino acids (dark blue circles), which we synthesised as a control. We considered two basic alphabets: 8E+Lys,Arg adds the two modern basic amino acids with long side chains (lime circles), whereas 8E+Dap,Dab adds the two most prebiotically available basic amino acids (cyan circles). We considered two acidic alphabets: 8E+Asp,Glu which adds the two canonical (and prebiotically plausible) acidic amino acids, and 8E+Aad,Api (L-2-aminoadipic acid, L-2-aminopimelic acid; orange circles) which is a 'counterfactual' alphabet. These amino acids were not prebiotically abundant but mimic the longer side chains of Lys and Arg and provide orthogonal means to address whether Asp and Glu provide unique advantages over hypothetical alternatives. Finally, we also considered four mixed alphabets in which acidic and basic amino acids coexist with varying side chain lengths.

We established the quality of these libraries with both MALDI mass spectrometry (Supplemental Fig. S1) and HPLC amino acid analysis (Supplemental Fig. S2, measurements compiled in Supplemental Data S2). MALDI showed distributions of precursor masses quite similar to theoretical predictions for each library (Supplemental Fig. S1). For instance, the predicted average molecular weight for a 25-mer peptide in the 8E+Dap,Dab library is 2330 Da; we observe that it is 2288 Da (difference 42 Da or ~0.44 amino acids), which is similar for the other libraries. Amino acid analysis showed correct incorporation of the amino acids in each library and close-to-uniform distributions in the amino acids' stoichiometry (Supplemental Fig. S2). For instance, since the 8E+Lys,Arg,Aad,Api library has 12 amino acids, each amino acid should represent 8.3% of the total amino acid content theoretically; we observe that each amino acid is present at 8 ± 3% of the content, and this is similar for the other eight libraries.

We conclude that the isokinetic mix strategy can produce reasonably representative random peptide libraries, even spanning those with highly distinct isoelectric points and noncanonical amino acids.

Solubility Profiles of Charge-Variant Libraries. Lyophilised peptide libraries were resuspended in aqueous Britton-Robinson (ABP) buffers of various pH's and supplemented with NaCl to either a medium (50 mM) or high (500 mM) ionic strength. After mixing, the solutions were centrifuged at 21,000 x g for 30 min to remove insoluble material, and the solubility of the alphabets were assessed by measuring peptide concentration in the soluble fraction by UV-Vis absorption spectroscopy (Fig. 2, measurements compiled in Supplemental Data S2).

Similarly to what we previously observed in 8E+Asp,Glu[17], the acidic 8E+Aad,Api alphabet showed poor solubility at low pH, and high solubility at neutral and high pH. These results can be rationalised since carboxylic acids are expected to be protonated and neutral at pH below 5, thus promoting aggregation. To our surprise, all other six libraries that contained basic amino acids were highly soluble at all pH's (Fig. 2), when we might have expected 8E+Lys,Arg and 8E+Dap,Dab to mirror the acidic libraries and be insoluble at higher pH's (although the pK_a of Arg is further away from physiological pH).

In general, ionic strength has minimal effect on solubility, with the only exception being the acidic alphabets (8E+Asp,Glu and 8E+Aad,Api) at physiological pH, where high salt concentration is deleterious. In other words, it is only the acidic alphabets that evidence the Hofmeister salting-out effect typically ascribed to globular proteins[23,24]. This result suggests that the acidic alphabets have a rather unique balance of hydrophobicity and hydrophilicity, which can promote aggregation under some conditions (i.e., low pH, high ionic strength) and potentially folding under other conditions (i.e., neutral pH, medium ionic strength).

Secondary Structure Formation in Charge-Variant Libraries. Both α-helices and β-sheets exhibit strong negative dichroism at 222 nm, whereas loops and coils exhibit strong negative dichroism at 200 nm and hardly any dichroism at 222 nm. Therefore, we adopt the ellipticity ratio between 222 nm and 200 nm as a general measure for secondary structure formation propensity, recognizing this ratio is an estimate that does not account for all effects that influence far-UV circular dichroism (CD) spectra[25]. Previously we had observed that the 8E+Asp,Glu alphabet expressed a higher ability to form stable secondary motifs as indicated by a clear negative feature in its CD spectrum at 222 nm, especially under physiological pH[17]. With the greater context afforded by comparing 8E+Asp,Glu to seven other alphabets, its remarkable folding potential comes into starker focus (Fig. 3, spectral data compiled in Supplemental Data S2). Even the chemically similar 8E+Aad,Api alphabet shows a diminished spectral feature at 222 nm suggesting peptides with attenuated secondary structure when the negatively charged carboxylic acids are placed further away from the backbone.

Moreover, we find that all six alphabets containing basic residues, either alone or in combination with acidic residues, exhibit strikingly diminished capacity to form secondary structures (Fig. 3). We were surprised by these results particularly in relation to the four mixed-charge libraries, as they have isoelectric point distributions more representative of naturally evolved proteins. These findings are invariant to pH, and indeed it is only the acidic alphabets whose secondary structure capacity show any pronounced dependence on pH. As a control to ensure that the folding properties were not influenced by the ABP buffer (which enabled the interrogation of a broad pH range), we recorded CD spectra for 8E+Asp,Glu, 8E+Lys,Arg, and

8E+Lys,Arg,Asp,Glu in Tris-HCl (pH 7.4) and found this change resulted in no significant differences with respect to ABP buffer (Supplemental Fig. S3).

As a further control, we acquired CD spectra of these random peptide libraries with varying amounts of 2,2,2-trifluoroethanol (TFE) added (between 10-90% (v/v)) as a cosolvent in order to induce α-helical structure (Supplemental Fig. S4). All the alphabets responded to TFE induction, with pronounced negative features at 222 nm appearing in all cases with TFE concentrations of 50% and higher. These experiments serve as useful controls that show all libraries have, in principle, the capacity to fold, and none of the noncanonical amino acids impose steric constraints that explicitly prevent it. Indeed, at 50% TFE, all but two of the other libraries showed *greater* helical content than 8E+Asp,Glu, with the 'long side chained alphabet' (8E+Lys,Arg,Aad,Api) showing the most pronounced structural induction by the organic cosolvent (Supplemental Fig. S4). Nevertheless, in aqueous conditions, loop and coil motifs are favoured over α-helices and β-sheets for all alphabets surveyed except for 8E+Asp,Glu.

Soluble Aggregate Formation in Charge-Variant Libraries. Size exclusion chromatography (SEC) can provide an orthogonal means to assess for aggregation, and unlike centrifugal pelleting assays, which measure precipitating aggregation, SEC is sensitive to the formation of soluble aggregates. We examined all of the peptide libraries in the pH range between 3–11, finding that in all cases, the majority of the peptide mass eluted as a single peak with an elution volume between 13-16 ml. These data indicate that the majority of the soluble peptides are monomeric and not forming oligomers or soluble aggregates (Fig. 4).

The two acidic libraries (8E+Asp,Glu and 8E+Aad,Api) showed the greatest pH-dependence on their elution profile in a predictable fashion. At pH 11, the peptides eluted at 13 ml, implying larger hydrodynamic radii when the molecules will be the most polyanionic (and most likely unfolded). At pH 3, where these peptides are more neutralised and higher levels of precipitation are observed (cf. Fig. 2), the soluble peptides eluted at 16 ml, implying more compact structures with lower hydrodynamic radii. Curiously, the two basic libraries (8E+Lys,Arg and 8E+Dap,Dab) did not show the expected reciprocal behaviour of becoming more compact at high pH, where the peptides would become more charge neutralised. Instead, the position of the primary peak remained at ca. 16 ml, though the chromatographic feature does broaden and narrow in a pH-dependent manner (Fig. 4). The SEC profiles of the four libraries with both acidic and basic residues responded to pH in a similar, if slightly attenuated, way as the acidic libraries.

We detected soluble aggregates forming as a separate feature at low elution volume (8 ml), but *only* in the two acidic libraries and *only* under pH's where these libraries did *not* produce precipitating aggregates (Figs. 2 and 4). This is a key observation as self-assembly into larger structures has been posited as an important property for prebiotic peptides[26,27].

These experiments were repeated at high ionic strength (500 mM NaCl, Supplemental Fig. S5), which produced SEC chromatograms that were largely similar but differed in two key respects. Firstly, the pH-associated changes to the peak position were reduced, which is reasonable because the salt screens repulsive ion-ion interactions that would promote peptide chain radius expansion. Secondly, and more importantly, the soluble aggregate feature we detected in 8E+Asp,Glu is greatly diminished at 500 mM NaCl, implying that these assemblies involve electrostatic interactions that are screened out at high ionic strength.

Beta Sheets Mediate Assembly of Prebiotic Acidic Peptides. To interrogate the nature of the peptide assemblies formed in the 8E+Asp,Glu library, we performed SEC preparatively and isolated the early-eluting peaks (Fig. 5). The peptide mixture was analysed

by CD spectroscopy, and remarkably, this subpopulation evinced clear signature for β-sheets, including a clear negative peak at 215 nm, and positive peak at 192 nm. When the subpopulation was hydrolysed and analysed with amino acid analysis, we saw an enrichment for the aliphatic branched amino acids, such as Val, Leu, and Ile, and a depletion of Pro and the charged amino acids, Asp and Glu. This shift in the amino acid frequencies matches closely with the amino acids known to have high and low propensity to occur in β-sheets. On the other hand, the late-eluting peak, which we can assign to monomeric peptides, possessed a CD spectrum and an amino acid composition relatively similar to the original unseparated peptide library (Fig. 5), though with an even more negative peak at 200 nm (corresponding to loops and coils) and a narrower shoulder at 222 nm (corresponding to α-helices).

With clearer evidence for the existence of β-sheets in the 8E+Asp,Glu library, the broader shoulder apparent in this library's original CD spectrum can be ascribed to a mixture of α- and β- content – the former associated with the more prevalent monomers, and the latter associated with assemblies. Bearing this insight in mind and referring back to the full set of CD spectra (cf. Fig. 2), it is apparent that 8E+Asp,Glu possessed the most dichroism at 215 nm (-14 mdeg) above neutral pH, followed by 8E+Aad,Api (-11 mdeg). For the other alphabets, the lesser degree of negative dichroism at 215 nm (-5 to -8 mdeg) signals less beta content amongst these peptide libraries.

We used mass photometry to estimate the aggregation number of the peptide assemblies in the 8E+Asp,Glu library. These experiments identified a non-Gaussian peak with a maximum at ca. 80 kDa (ca. 32 peptides, Supplemental Fig. S6b) and a fat tail extending to 400 kDa, which is removed entirely in the late-eluting SEC fraction and enriched substantially in the early-eluting SEC fraction (Supplemental Fig. S6c-d). Hence, we conclude that the 8E+Asp,Glu alphabet admits a rather intriguing characteristic: a sub-population of peptides that spontaneously assemble into larger soluble structures consisting primarily of β-sheets. This property emerges in random sequence space, though only with the appropriate alphabet.

Large Language Models Recapitulate Structure-Forming Potential of Acidic Residues. Obtaining structural information on our random peptide libraries is challenging on two levels: (i) the vast compositional diversity of the libraries (essentially every sequence in the mixture is unique) and (ii) these short unevolved 25-mer peptides are expected to be dynamic and conformationally heterogenous. Recently, large language models have been developed for protein structure prediction[22,28]. These models do not rely on multiple sequence alignments, and hence it has been argued that they have learned a more generic relationship between sequence and structure that is not dependent on similarity to biological sequences or even biological amino acid compositions. We therefore carried out ESMFold structure predictions on 50,000 unique and randomly generated 25-mer sequences from four alphabets comprising only canonical amino acids (8E, 8E+Asp,Glu, 8E+Lys,Arg, and 8E+Lys,Arg,Asp,Glu) as ESMFold does not support noncanonical amino acids. Fig. 6a shows the distribution of radii of gyration (R_g), a measure of compactness, for peptides using these alphabets. In consonance with our experimental studies, ESMFold predicts that peptide compositions drawing from the prebiotically plausible 8E+Asp,Glu alphabet (median R_g = 1.63 nm) are slightly more compact than peptides drawing from the basic alphabet (median R_g = 1.82 nm) or the charge-balanced alphabet (median R_g = 1.69 nm).

On the other hand, ESMFold predicts that the 8E library, which possesses no charged residues, generates the most compact structures (Fig. 6a, median R_g = 1.36 nm). At face value, this result appears to be in direct contrast with CD spectroscopy (Fig. 6b) of this peptide

library, which evidences very little secondary structure. However, fluorometric solubility assays show that the 8E library has poor solubility at all pH's (Fig. 6b), which can be ascribed to its very hydrophobic composition (results for all other assays on the 8E peptide library are given in Supplemental Fig. S7). One can reconcile these findings by suggesting that 8E *does* possess many sequences that can fold in principle, but those peptides also aggregate, and are therefore depleted from the soluble fraction on which CD spectra were acquired.

We next assessed the secondary structures in the generated structural models of these random sequences with DSSP4.4[29] (Fig. 6c). In general, coil conformations are the most common followed by helices; β-sheets appear to be quite rare amongst random peptides as is to be expected due to the internal helix stabilisation. The 8E library had the most helical content (19.6%, consistent with its lower R_g values), but surprisingly, the 8E+Asp,Glu sequences had the least helical content (9.7%). On the other hand, 8E+Asp,Glu exhibited lower levels of polyproline II character (1.5%, associated with extended conformations) and higher levels of turns and bends (10.3%, which can help a chain compact by turning on itself), consistent with the radius of gyration calculations[30]. To support these patterns in secondary structure, we also analyzed 100,000 random sequences from all four alphabets with three different sequence-based secondary structure prediction algorithms (Spider3[31], Jnet[32], and Gor4[33]; Supplemental Fig. S8). While the Gor4 results are rather inconclusive, Jnet and Spider3 qualitatively agree with ESMFold structural models in that they also predict the highest secondary structure propensity in the hydrophobic 8E library, and the least secondary structure propensity in the 8E+Asp,Glu library.

ESMFold (Fig. 6c) and the secondary structure prediction algorithms (Supplemental Fig. S8) are most inconsistent with experiments in their prediction for higher levels of α-helices in the 8E+Lys,Arg peptide library, which is not evidenced in CD spectra (cf. Fig. 3). We wondered if perhaps, given its directive to "force" proteins fold, ESMFold predicts extended alpha helices for 8E+Lys,Arg sequences that in reality are unstable in aqueous solution given their lack of hydrophobic packing interactions. To test this, we applied a generative model for protein thermodynamic stability prediction[34]. Correlating predicted ΔG against predicted R_g revealed that 8E+Asp,Glu peptides – whilst not more stable overall – were more likely to positively correlate stability with compaction (Supplemental Fig. S9a-d). In other words, 8E+Asp,Glu prebiotic alphabet admits a clearer structure-stability relationship whereas 8E+Lys,Arg in particular does not. In further support of this idea, when we reconsider CD spectra in which folding was induced with TFE (Supplemental Fig. S4), it is clear that both basic libraries (8E+Lys,Arg and 8E+Dap,Dab) revealed greater inducibility, with lower secondary structure than 8E+Asp,Glu at TFE at 20% (v/v) and below, but then surpassing 8E+Asp,Glu at TFE at 30% (v/v) and higher. In summary, the experiments and generative modeling support the premise that 8E+Lys,Arg has a higher helical propensity in principle, that is not realised in practice in aqueous solution due to its lower capacity to compact. The acidic library on the other hand, can better stabilise secondary structures by adopting more compact conformations.

Discussion

Our research into the properties of prebiotically-plausible amino acid alphabets began with the assumption that a severe reduction in the chemical inventory would imply proteins with more limited folding and function compared to modern proteins with their access to aromatic, basic, and sulfur-containing building blocks. We have so far seen that such a rudimentary alphabet could fulfil simple requirements like the ability to fold[17], bind RNA[13], or

bind organic cofactors[35]. In light of the results described here, our view of the early 8E+Asp,Glu alphabet has further evolved, as here we show that this pool of building blocks possesses unusually adaptive properties fit for serving as a "stepping stone" to the emergence of proteins via self-assembling peptides.

We have seen previously by CD spectroscopy that random peptides using the 8E+Asp,Glu alphabet have a high propensity to form secondary structures. The examination of eight further random peptide libraries spanning the gamut of isoelectric point, hydrophobicity, and side chain lengths have revealed that this property is not common. 8E+Asp,Glu possesses this trait whilst also maintaining high solubility, a delicate "balancing act" that it pulls off, unlike the hydrophobic 8E alphabet and the modern alphabet (cf. Supplemental Fig. S10 and ref. 12,17), which are highly prone to aggregate. The fact that 8E+Aad,Api alphabet recapitulates most of the key biophysical features of 8E+Asp,Glu shows that there is *some* leeway in the chemical space that admits peptides that can fold and assemble. These results suggest that side-chain length is perhaps less critical of a modulator than we had initially hypothesised[17].

Together, these findings suggest that 8E+Asp,Glu was a strong candidate for assisting the emergence of proteins prior to purifying selection. The modern 20-letter alphabet has great functional breadth but owing to specific sequences that have been finessed by selection and frequently assisted by chaperones[12,36]. Random sequences utilising this more advanced amino acid repertoire, on the other hand, suffer from aggregation. An amino acid set that promotes high solubility and foldability across unevolved sequences would have played a key adaptive role during the early stages of life's emergence, for which the early 8E+Asp,Glu alphabet would be more suitable.

Both experiments and AI modelling agree that β-sheets are rare amongst random peptide sequences as monomers. We report here, however, that a subset of 8E+Asp,Glu peptides self-assemble into higher-order oligomers with pronounced β-sheet content, and SEC as well as reanalysis of CD spectra again suggest this property is not common. This finding is particularly germane to the theory that amyloids could have played a key role during the origin of life (the amyloid world hypothesis), with their ability to form complex structures and propagate using short peptides as the monomer[26]. Modern amyloids, however, are enriched with many late amino acids (Asn, Gln, Tyr), and most studies exploring the amyloid world hypothesis use engineered peptides with regular sequence patterns, late amino acids, and *N*-acetyl capping. Our results here constitute, to the best of our knowledge, the first observation of soluble oligomeric β-sheet assemblies arising from random peptide sequences comprising early amino acids, extending previous observations on β-content in insoluble protein[37]. Whilst we would caution from referring to these species as amyloids (their molecular weights span between 80-400 kDa and they are soluble), they provide evidence that beta structures were also available to early peptides, albeit in an oligomeric rather than monomeric form. Importantly, self-assembled peptides stemming from relatively short sequences have been implied as plausible sites of prebiotic catalysis, preceding evolution of longer polypeptides and tertiary folding[38,39,40].

A view of the protein universe based solely on extant proteins might suggest that polypeptide sequences which fold, assemble, and stay soluble are rare, and once discovered, are conserved, thereby explaining the relatively small number of unique fold topologies used in Nature[41]. Nevertheless, this is a view based on the modern amino acid alphabet. An early acidic amino acid alphabet can produce random polypeptides with these properties with higher probability. With this view in mind, the emergence of polymers that can fold and assemble

from a primordial soup rich in α-amino acids begins to appear less of a statistical anomaly, but rather, quite plausible.

Figures

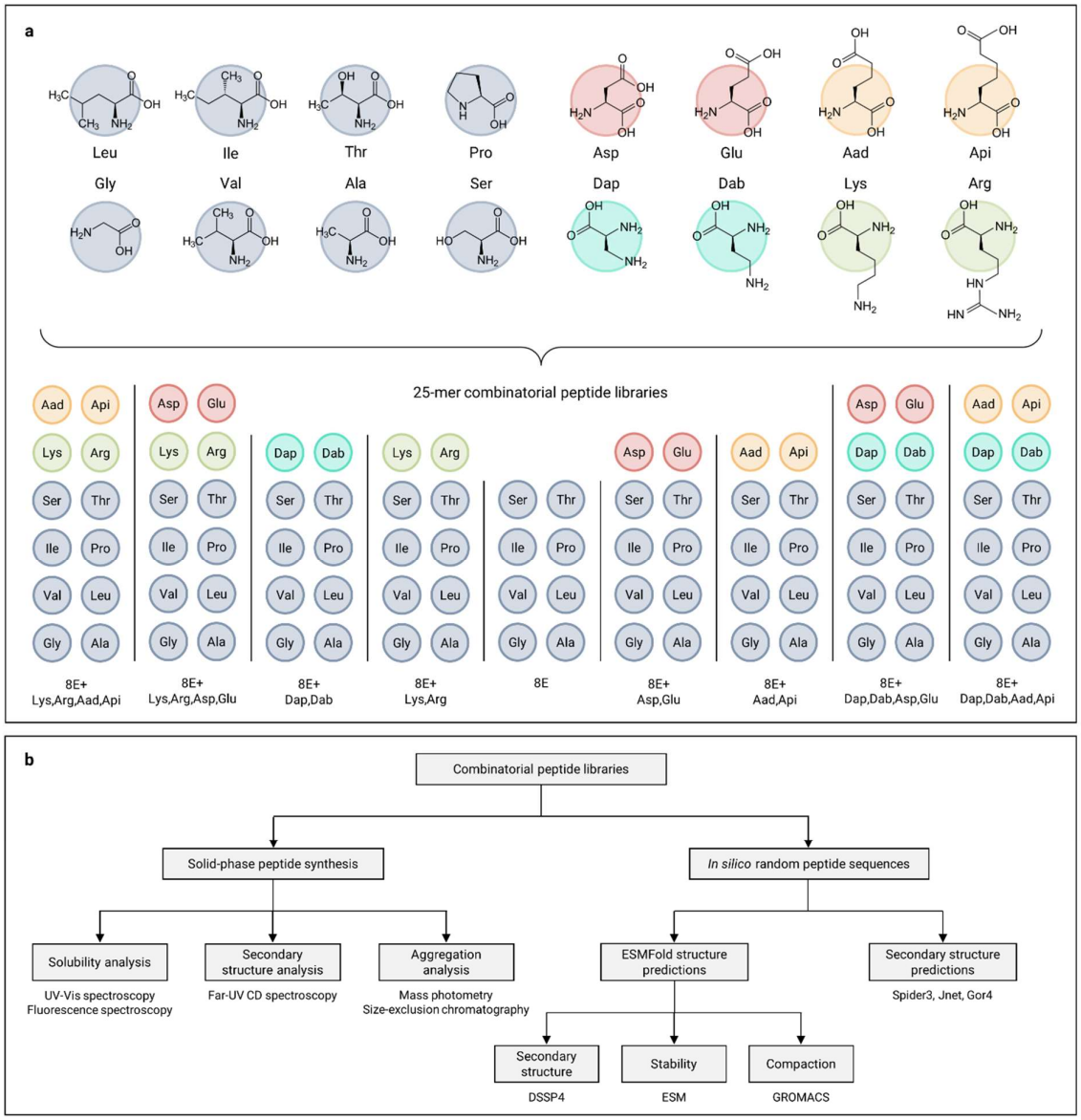


Figure 1. Design and evaluation of random peptide libraries using distinct amino acid alphabets.

(A) Chemical structures of 16 L-α-amino acids assembled through solid phase peptide synthesis to generate random 25-mer peptides utilising one of nine distinct alphabets (collections of amino acids). Dap, 2,3-diaminopropionic acid; Dab, 2,4-diaminobutyric acid; Aad, 2-aminoadipic acid; Api, 2-aminopimelic acid.

(B) Methods used in this study to interrogate random peptide libraries.

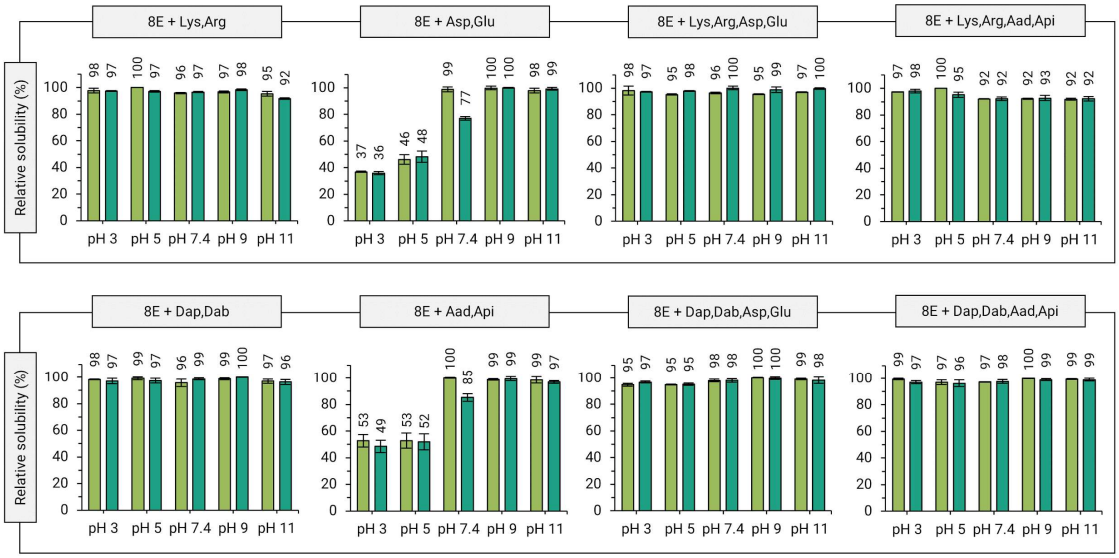


Figure 2. The relative solubility of eight peptide libraries.
Solubility of the indicated libraries was measured in 20 mM ABP buffer at different pH's and ionic strengths by absorption of peptide bonds at 215 nm. Lime bars correspond to medium (50 mM NaCl) and teal bars correspond to high (500 mM NaCl) ionic strength. Error bars represent standard deviations from duplicates with the exception of 8E+Asp,Glu library (triplicates). The data for 8E+Asp,Glu were taken from ref. 17.

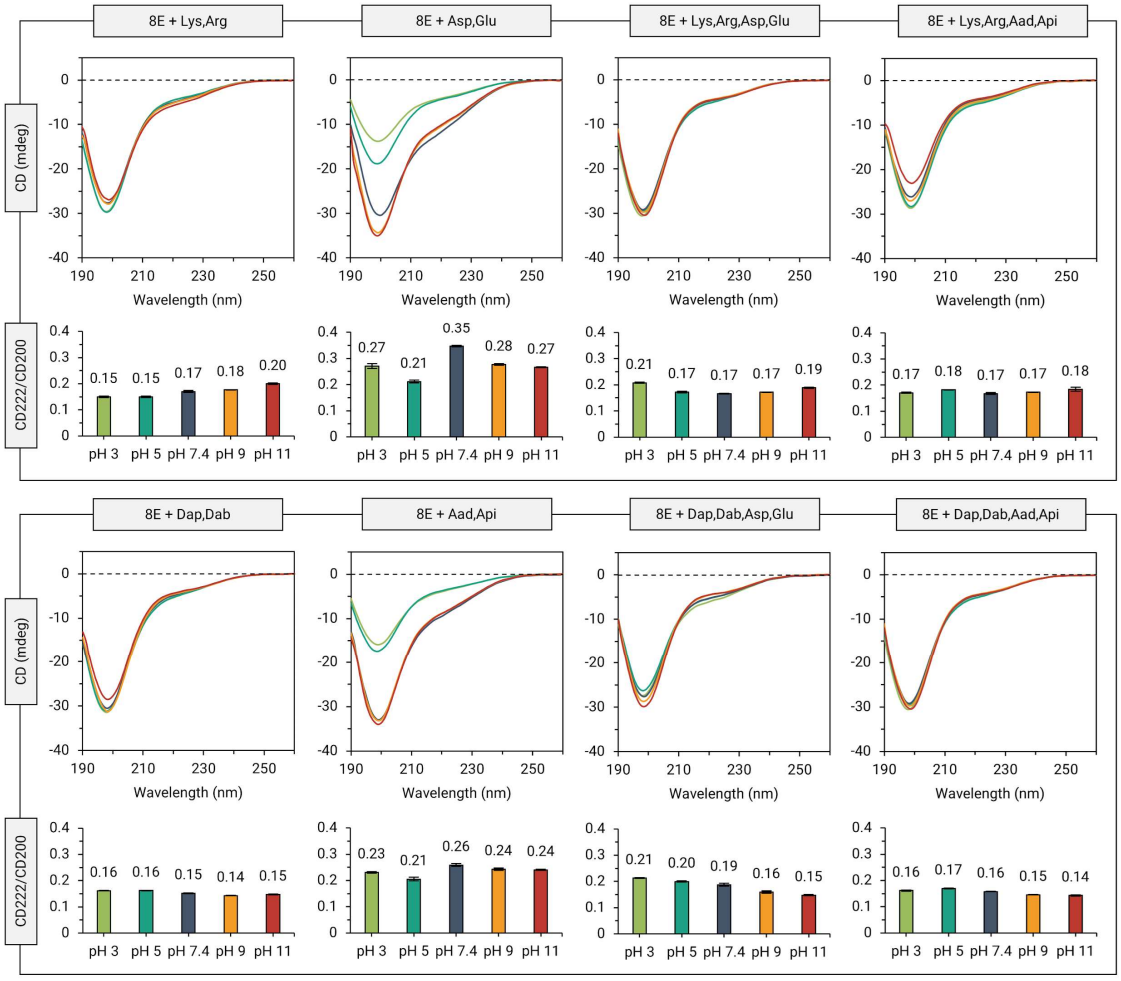


Figure 3. The secondary structure of eight peptide libraries.
Secondary structure of the indicated libraries was measured in 10 mM ABP buffer at different pH's with far-UV circular dichroism (CD) spectroscopy. Far-UV CD spectra are representative from duplicates. Secondary structure content is estimated as the ratio of the ellipticities at 222 nm and 200 nm as shown in the bar charts. Error bars represent standard deviations from duplicates. The data for 8E+Asp,Glu library were taken from ref. 17.

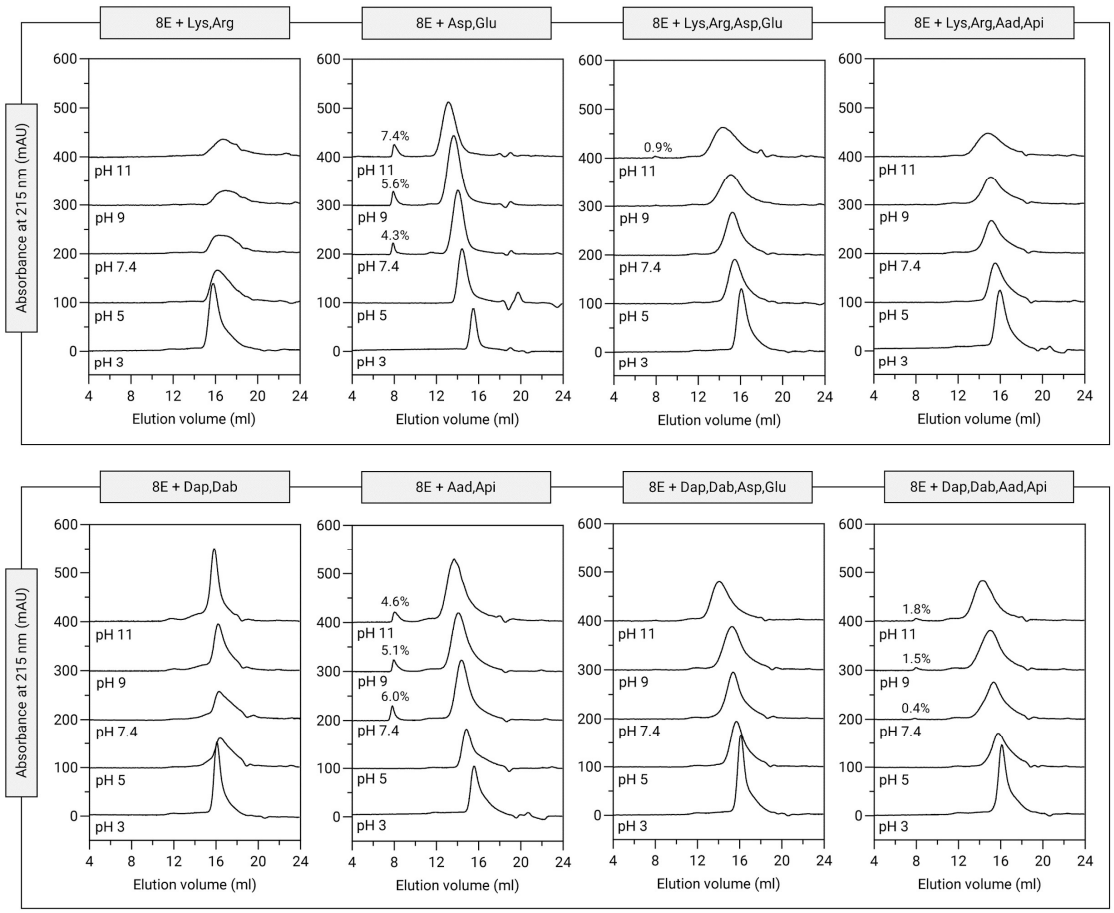


Figure 4. The aggregation propensity of eight peptide libraries.
Aggregation propensity of the indicated libraries was measured in 20 mM ABP buffer at different pH's and at medium ionic strength (50 mM NaCl) with size-exclusion chromatography (SEC). SEC chromatograms report absorbance at 215 nm as a function of elution volume. The data for 8E+Asp,Glu library were taken from ref. 17.

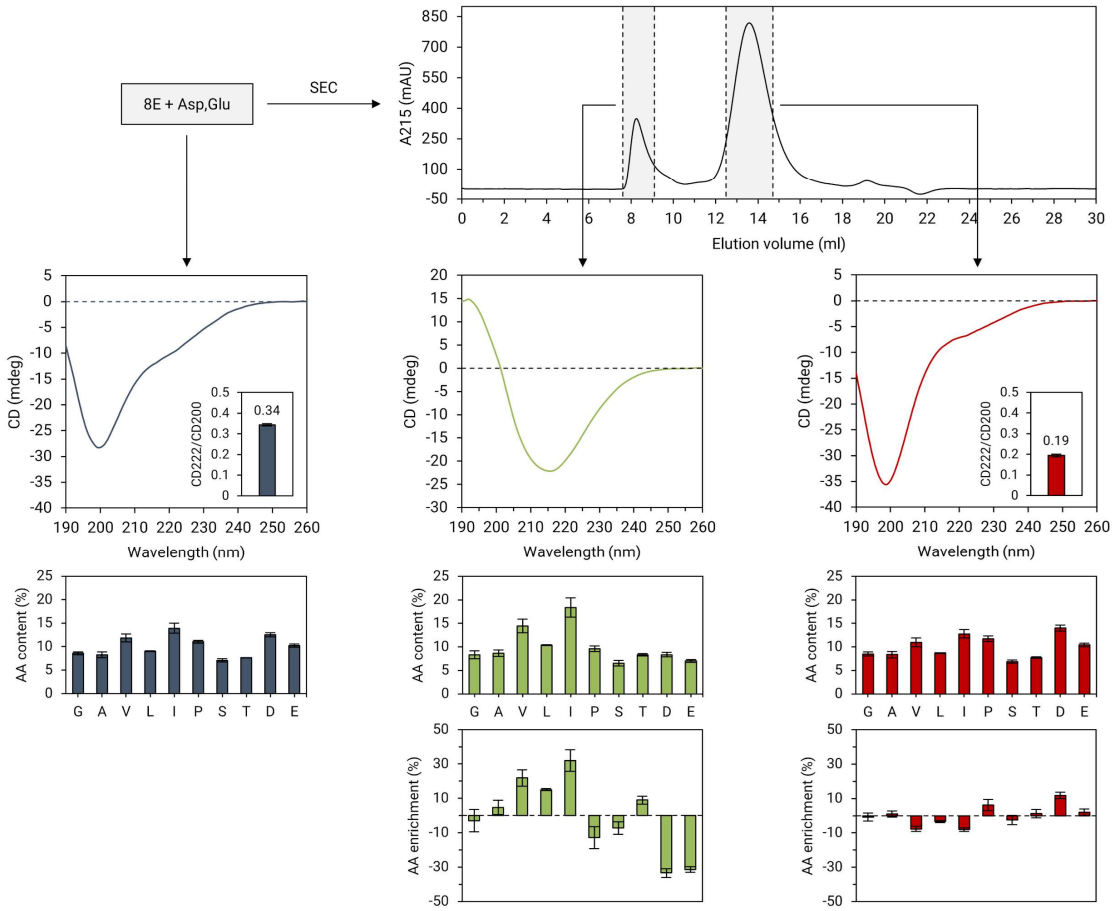


Figure 5. The resolution of a β-sheet rich subpopulation of peptides from the soluble aggregate fraction of the 8E+Asp,Glu library.

The 8E+Asp,Glu library was dissolved in 10 mM ABP buffer (pH 7.4) and resolved by size-exclusion chromatography. The soluble aggregate fraction (lime) at 8.5 ml was isolated and analyzed by far-UV circular dichroism (CD) and HPLC amino acid analysis. The far-UV CD spectrum possesses a negative peak at 215 nm and a positive peak at 192 nm, signalling β-sheet secondary structure. The HPLC amino acid analysis showed enrichment in branched amino acids (Val (V), Leu (L), and Ile (I)) and depletion of Pro (P) and acidic residues (Asp (D), Glu (E)). The monomer fraction (red) at 13.5 ml possesses a far-UV CD spectrum and amino acid content similar to the input peptide library.

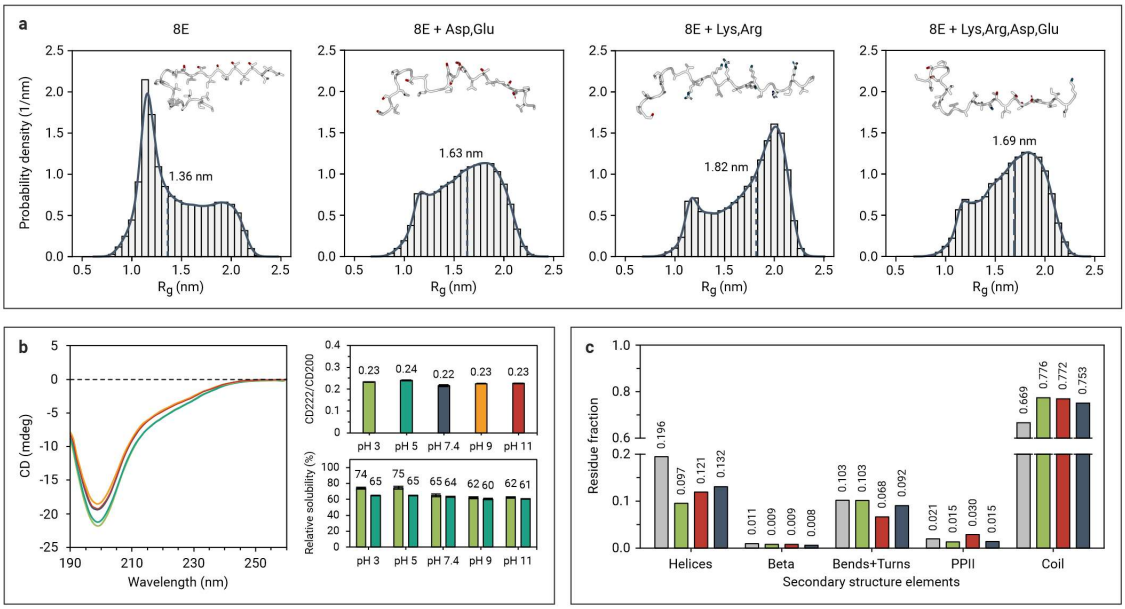


Figure 6. A large language model predicts greater compaction for acidic peptide sequences.

(A, C) The structures of 50,000 unique and randomly generated 25-mer peptide sequences were predicted using ESMFold[22] for each 8E, 8E+Asp,Glu, 8E+Lys,Arg and 8E+Lys,Arg,Asp,Glu libraries.

(A) Distribution of the radius of gyration calculated with GROMACS 2024.2. Representative peptide structures centered at the distributions' medians (dashed lines) are shown above.

(B) Secondary structure and solubility of 8E library at various pH.

(C) Fraction of residues with secondary structure calculated by DSSP 4.4 for 8E (grey); 8E+Asp,Glu (lime); 8E+Lys,Arg (red); and 8E+Lys,Arg,Asp,Glu (blue) libraries. Helices includes α-helix, π-helix, and 3_{10} helix; Beta includes extended β-sheets and β-bridges; PPII, polyproline II; coil, random coil.

Experimental procedures

Solid-phase synthesis.
Synthesis was conducted using the standard Fmoc peptide synthesis methodology using the Spyder Mark IV automated synthesizer (Spyder Institute Praha, Czech Republic). Fmoc-protected amino acids were weighed (see Supplemental Data S1), combined, and then dissolved to form a 0.3 M solution with 0.3 M N-hydroxybenzotriazole (HOBt) in dimethylformamide (DMF). Synthetic protocol on Rink amide AM resin (200 mg, 0.42 mmol/g) for one cycle was as follows: 2 × 1 min washing with DMF, 1 × 1 and 1 × 20 min treatment with 20% piperidine in DMF, 4 × 1 min wash with DMF, 1 × 1 min wash with 0.3 M HOBt, 2 × 60 min coupling with amino acid mixture, and 1 M diisopropylcarbodiimide in DMF. Following 25 synthetic cycles, the resin was thoroughly rinsed with DMF and dichloromethane, then dried under vacuum. The dried resin underwent treatment with 5 ml of Mixture K (trifluoroacetic acid–thioanisole–water–phenol–ethanedithiol, 82.5:5:5:5:2.5, v/v) for 2 h. The resin was then filtered out, and the resulting peptide solution was precipitated with diethyl ether thrice before being vacuum-dried. The resulting pellet was dissolved in 20% acetic acid and subjected to lyophilization. Before conducting subsequent experiments, all peptide libraries were further lyophilized thrice from 1 mM HCl overnight, establishing Cl^- as the counterion.

Quality control.
The molecular weight distributions of combinatorial peptide libraries were confirmed by mass spectrometry using UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Germany) according to the standard procedure. Prior to the experiment, peptide libraries were dissolved in acetonitrile:water (1:1) mixture to the final concentration of 10 mg/ml. 2,5-Dihydroxybenzoic acid (DHB) was used as a matrix.
Prior to the amino acid analysis, the library samples were hydrolyzed in 6 M hydrochloric acid at 110 °C for 20 h and the hydrolysate was evaporated and reconstituted with 0.1 M hydrochloric acid containing the internal standard. Amino acid analysis was performed on Agilent 1260 HPLC (Agilent Technologies, Germany) equipped with an UV-Vis detector using automated ninhydrin derivatization.

Solubility and aggregation analysis.
Prior to analysis, a series of universal Britton-Robinson (ABP) buffers was prepared by mixing 20 mM acetic acid, 20 mM phosphoric acid and 20 mM boric acid and adjusting pH with 5 M NaOH. The ionic strength was adjusted with NaCl to conductivity that corresponds to the conductivity of 50 mM or 500 mM NaCl. Peptide libraries were thoroughly resuspended in autoclaved MilliQ water to the final concentration of 5 mg/ml. The peptide suspensions were diluted into the corresponding ABP buffer to the final concentration of 0.5 mg/ml and incubated for 30 min with mild shaking (500 rpm) at room temperature. After incubation, the insoluble fraction was separated by centrifugation at 21,300 *g* for 30 min at 4 ℃, and the supernatant (cleared sample) was subsequently used for solubility and aggregation analyses.
For all peptide libraries except 8E library, the amount of soluble peptides was estimated by measuring absorption of peptide bonds at 215 nm on a Nanodrop 2000 (Thermo Fisher Scientific, USA).
Due to its low solubility, the relative solubility of 8E peptide library was estimated with a fluorescamine assay[42] in a 96-well plate format. In brief, 90 µl of the peptide library solution was mixed with 30 µl of 3 mg/ml fluorescamine in DMSO, and the resulting mixture was

incubated for 30 min in the dark at room temperature in a dark-bottom 96-well plate. The fluorescence was then recorded on Spark microplate reader (Tecan, Switzerland) using the following parameters: excitation at 365 nm, emission at 470 nm. 0.5 mg/ml solution of 8E peptide library in DMSO was used as a standard. All solubility measurements were conducted in technical duplicate for every library at every pH at two ionic strengths.

The relative amount of soluble aggregates was estimated by size-exclusion chromatography. A 100 µl aliquot of clarified sample (following centrifugation) was loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva, USA) that was pre-equilibrated with two bed volumes of the corresponding 20 mM ABP buffer with either 50 or 500 mM NaCl. The peptides were eluted from the column isocratically on NGC Quest Plus System (Bio-Rad, USA), by one bed volume of the buffer at 0.5 ml/min flow rate at room temperature, and the eluted peptides were detected by absorption at 215 nm. The aggregation analysis was performed by quantifying the relative intensity underneath the peaks at ~8 or ~14 ml elution volume in the acquisition software (ChromLab Software v. 6.1). Size exclusion chromatography runs were conducted once for every library at every pH at two ionic strengths.

Secondary structure analysis.

The CD spectra were recorded using a Chirascan-plus spectrophotometer (Applied Photophysics, UK) over the wavelength range 190–260 nm in steps of 1 nm with an averaging time of 1 s per step. Cleared samples at 0.2 mg/ml nominal concentration in 1 mm path-length quartz cells were placed into a cell holder, and spectra were recorded at room temperature. The CD signal was obtained as ellipticity in units of millidegrees, and the resulting spectra were averaged from two scans and buffer-spectrum subtracted. All CD measurements were performed twice, and the resulting ratio of ellipticity at 222 nm to ellipticity at 200 nm was averaged and used to estimate the secondary structure content in peptide libraries.

To estimate the effect of pH on secondary structure, 5 mg/ml suspensions of peptide libraries in Milli-Q water were diluted into 10 mM ABP buffers at pH 3.0, 5.0, 7.4, 9.0, and 11.0 to the final nominal concentration of 0.2 mg/ml, gently mixed at room temperature for 30 min, and then centrifuged at maximum speed (21,300 *g*) for 15 min at 4 °C in order to remove insoluble fraction.

To estimate the effect of 2,2,2-trifluoroethanol on secondary structure, 5 mg/ml suspensions of peptide libraries in MilliQ water were diluted in a series of 10 mM ABP (pH 7.4) supplemented with 0–90% (v/v) 2,2,2-trifluoroethanol.

To estimate the effect of buffer on secondary structure, 5 mg/ml suspensions of peptide libraries in MilliQ water were diluted in 10 mM ABP (pH 7.4) or 10 mM Tris-HCl (pH 7.4).

Isolation and analysis of total, aggregate and monomer fractions from 8E+Asp,Glu library.

The 800 µl aliquot of cleared 0.5 mg/ml 8E+Asp,Glu library solution in 10 mM ABP (pH 7.4) was loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva, USA) that was pre-equilibrated with two bed volumes of 10 mM ABP buffer (pH 7.4). The peptides were eluted from the column by one bed volume of the buffer at 0.5 ml/min flow rate at room temperature, and the eluted peptides were detected by absorption of peptide bonds at 215 nm. The aggregate (~8.5 ml) and monomer (~13.5 ml) fractions were isolated and concentrated to 0.2 mg/ml using Amicon Ultra centrifugal filters (3 kDa MWCO). The total fraction was prepared by direct dilution of 0.5 mg/ml 8E+Asp,Glu library solution with 10 mM ABP (pH 7.4) to 0.2 mg/ml. The secondary structure analysis and HPLC amino acid analysis of total, aggregate and monomer fractions was performed as described above. The collection of the total,

aggregate, and monomer fractions from the 8E+Asp,Glu library and the HPLC amino acid analysis on those fractions was conducted in technical triplicate. CD measurements and secondary structure analysis were conducted in technical duplicate.

The mass photometry data were collected using a TwoMP system (Refeyn, UK)[43]. Prior to analysis, total, aggregate and monomer fractions were diluted with 10 mM Tris (pH 7.4), 50 mM NaCl to the final concentration of 0.02 mg/ml (~8 μM). 17.5 μl of 10 mM Tris (pH 7.4), 50 mM NaCl was placed in a well of glass casket (Refeyn, UK) to focus the objective and subsequently mixed with 2.5 μl of the sample to achieve the final concentration of 2.5 μg/ml (~1 μM). Mass photometry data was acquired and analyzed using AcquireMP and DiscoverMP software (Refeyn, UK), respectively. Mass calibration was performed using 10 nM bovine serum albumin (BSA) in 10 mM Tris (pH 7.4), 50 mM NaCl as a standard. The mass photometry analysis was performed once.

Radius of gyration, secondary structure annotation, and unfolding free-energy analysis on predicted peptide structures.
Fifty thousand (50,000) unique and random 25-mer peptide sequences were generated *in-silico* by uniformly sampling the amino acid alphabets of each library (8E, 8E+Asp,Glu, 8E+Lys,Arg, 8E+Lys,Arg,Asp,Glu and the full canonical alphabet). The heavy atom coordinates of all two hundred fifty thousand (250,000) peptide sequences were predicted using ESMFold-v1[22]. These structures were used to calculate the radius of gyration (GROMACS 2024.2), annotate their secondary structure (DSSP 4.4), and estimate the unfolding free energy[34]. All histograms and kernel density estimations were calculated using Python v3.12.0 and SciPy v1.13.0.

Gor4, Jnet, and Spider3 secondary structure predictions.
For each of the four libraries (8E, 8E+Asp,Glu, 8E+Lys,Arg, 8E+Lys,Arg,Asp,Glu), one hundred thousand (100,000) unique random 25-mer peptide sequences were generated *in silico* by equimolar random sampling of the amino acids. For these sets of 100,000 sequences, the secondary structure was predicted by the Gor4[33], Jnet[32], Spider3[31] algorithms. In all three cases, only the list of sequences was the input information for the prediction. The secondary structure propensity of each library was then calculated as a mean of the fraction of the amino acids in each sequence predicted to be either α-helix or β-sheet for all 100,000 sequences.

Recource availability

Lead contact
Further information should be directed to and will be fulfilled by the lead contact, Klara Hlouchova (klara.hlouchova@natur.cuni.cz)

Materials availability
This study did not generate new unique reagents or genetic material.

Data and code availability
To facilitate resynthesis, all product codes and loadings for Fmoc-protected amino acids used for solid phase peptide synthesis are reported as Supplemental Data S1. To facilitate reanalysis, raw CD spectral data, solubility data, and amino acid analysis are supplied as

Supplemental Data S2. Computational results including predicted radius of gyration and secondary structure for 50,000 random sequences for 5 alphabets are supplied in Supplemental Data S3.

Acknowledgements

We thank Dr. Radko Souček and Dr. Kateřina Nováková for their technical support during the quality control of the peptide libraries. We thank John Cava for assistance with installation of ESMFold on compute clusters. In addition, we would like to thank Prof. Stephen Freeland for helpful discussions about this work.

Author Contributions

Conceptualization: S.D.F., K.H.; Chemical Synthesis: R.K., M.L.; Experimental Investigation: M.M., T.C.; Computational Investigation: E.M.S., S.D., T.K., J.M.K., S.M.B.; Visualization: M.M., E.M.S., S.M.B.; Writing–original draft: M.M., S.D.F., K.H.; Writing–editing: M.M., R.K., E.M.S., S.D., T.K., S.M.B., S.D.F., K.H.; Supervision: J.K., A.S., S.D.F., K.H.; Project Administration: S.D.F., K.H.; Funding Acquisition: J.K., S.D.F., K.H.

Funding

This work was supported by the Human Frontier Science Program grant HFSP-RGY0074/2019 and HFSP-RGEC27/2023 to S.D.F. and K.H. S.D.F. acknowledges support from the NIH Director's New Innovator Award (DP2-GM140926), a Cottrell Scholar Award, and a Sloan Fellowship for support. E.M.S. thanks the Program in Molecular Biophysics training grant (NIH-T32GM135131). M.M. and K.H. acknowledge CMS CF Biophysical techniques of CIISB, Instruct-CZ Centre, supported by MEYS CR Infrastructure project LM2018127 and European Regional Development Fund-Project, "UP CIISB" (No. CZ.02.1.01/0.0/0.0/18_046/0015974). Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Competing Interests

The authors have no competing interests to declare.

Supplemental information

Document S1. Supplemental Figures S1-S10
Data S1. Details of the material used in solid phase peptide synthesis
Data S2. Raw CD spectral data, solubility data, and amino acid analysis
Data S3. Computational results including predicted radius of gyration and secondary structure for 50,000 random sequences

References

1. Weber, A. L. & Miller, S. L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol*. **17**, 273–284 (1981).
2. Cleaves, H.J. 2nd. The origin of the biologically coded amino acids. *J. Theor. Biol*. **263**, 490–498 (2010).
3. Philip, G.K. & Freeland, S. J. Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology*. **11**, 235–240 (2011).
4. Mayer-Bacon, C. & Freeland, S. J. A broader context for understanding amino acid alphabet optimality. *J. Theor. Biol*. **520**, 110661 (2021).
5. Brown, S. M., Mayer-Bacon, C., & Freeland, S. (2023). Xeno amino acids: a look into biochemistry as we do not know it. *Life*, **13**, 2281 (2023).
6. Miller, S. L. & Urey, H. C. Organic compound synthesis on the primitive earth. *Science*. **130**, 245–251 (1959).
7. Johnson, A. P. et al. The Miller volcanic spark discharge experiment. *Science*. **322**, 404 (2008).
8. McCollom, T. Miller-Urey and beyond: what have we learned about prebiotic organic synthesis reactions in the past 60 years?. *Ann. Rev. Earth Planet. Sci.* **41**, 207–229 (2012).
9. Elsila, J. E. et al. Meteoritic amino acids: diversity in compositions reflects parent body histories. *ACS Cent. Sci*. **2**, 370–379 (2016).
10. Higgs, P. G. & Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*. **9**, 483–490 (2009).
11. Despotović, D. et al. Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry*. **59**, 4456–4462 (2020).
12. Tretyachenko, V. et al. Modern and prebiotic amino acids support distinct structural profiles in proteins. *Open Biol*. **12**, 220040 (2022).
13. Giacobelli, V. G. et al. In vitro evolution reveals noncationic protein-RNA interaction mediated by metal ions. *Mol. Biol. Evol*. **39**, msac032 (2022).
14. Tanaka, J., Doi, N., Takashima, H. & Yanagawa, H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* **19**, 786–795 (2010).
15. Newton, M. S., Morrone, D. J., Lee, K. H. & Seelig, B. Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries. *Chembiochem*. **20**, 8466–856 (2019).
16. Meierhenrich, U. J., Muñoz Caro G. M., Bredehöft, J. H., Jessberger E. K. & Thiemann, W. H. Identification of diamino acids in the Murchison meteorite. *Proc. Natl. Acad. Sci. USA*. **101**, 9182–9186 (2004).
17. Makarov, M. et al. Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc.* **145**, 5320–5329 (2023).
18. Leman, L., Orgel, L. & Ghadiri, M. R. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science*. **306**, 283–286 (2004).
19. Forsythe, J. G. et al. Ester-mediated amide bond formation driven by wet-dry cycles: a possible path to polypeptides on the prebiotic earth. *Angew. Chem. Int. Ed. Engl*. **54**, 9871–9875 (2015).
20. Canavelli, P., Islam, S. & Powner, M. W. Peptide ligation by chemoselective aminonitrile coupling in water. *Nature*. **571**, 546–549 (2019).

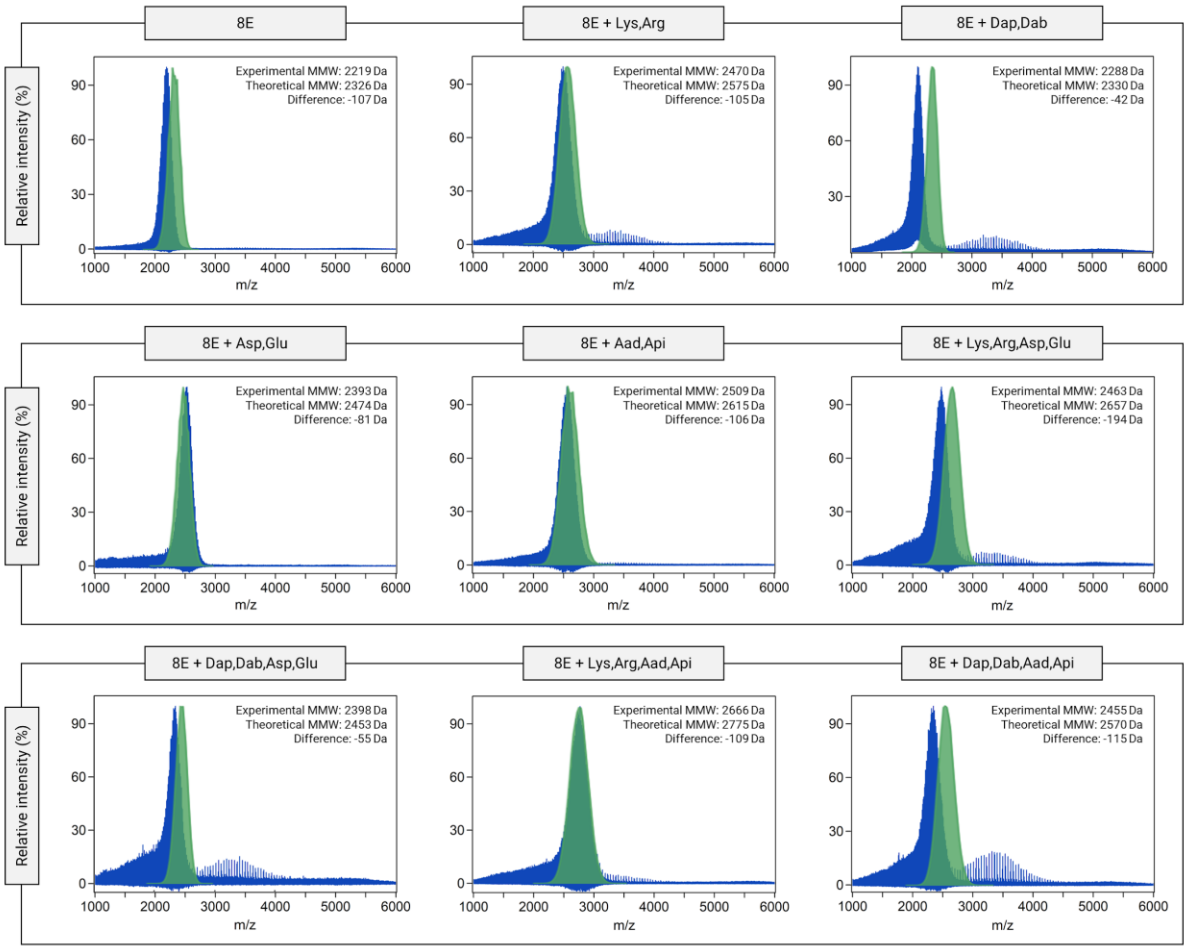
21. Foden, C. S. et al. Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. *Science*. **370**, 865–869 (2020).
22. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. **379**, 1123–1130 (2023).
23. Zhang, Y. & Cremer, P. S. Interactions between macromolecules and ions: The Hofmeister series. *Curr. Opin. Chem. Biol*. **10**, 658–663 (2006).
24. Kherb, J., Flores, S. C. & Cremer, P. S. Role of carboxylate side chains in the cation Hofmeister series. *J. Phys. Chem. B*. **116**, 7389–7397 (2012).
25. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc*. **1**, 2876–2890 (2006).
26. Greenwald, J., Kwiatkowski, W. & Riek, R. Peptide Amyloids in the Origin of Life. *J. Mol. Biol*. **430**, 3735–3750 (2018).
27. Frenkel-Pinter, M., Samanta, M., Ashkenasy, G. & Leman, L. J. Prebiotic peptides: molecular hubs in the origin of life. *Chem. Rev*. **120**, 4707–4765 (2020).
28. Jing, X., Wu, F., Luo, X. & Xu, J. Single-sequence protein structure prediction by integrating protein language models. *Proc. Natl. Acad. Sci. U S A*. **121**, e2308788121 (2024).
29. Joosten, R. P. et al. A series of PDB related databases for everyday needs. *Nucleic Acids Res*. **39**, D411–419 (2011).
30. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Software X*, **1**, 19–25 (2015).
31. Heffernan, R. et al. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem*. **39**, 2210–2216 (2018).
32. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*. **40**, 502–511 (2000).
33. Garnier, J., Osguthorpe, D. J. & Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol*. **120**, 97–120 (1978).
34. Cagiada, M., Ovchinnikov, S. & Lindorff-Larsen, K. Predicting absolute protein folding stability using generative models. Preprint at http://www.biorxiv.org/content/10.1101/2024.03.14.584940v1 (2024).
35. Sanchez-Rocha, A. C., Makarov, M., Pravda, L., Novotný, M. & Hlouchová, K. Preprint at https://elifesciences.org/reviewed-preprints/94174 (2024).
36. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., & Hartl, F. U. Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annu. Rev. Biochem.* **82**. 323–355 (2013).
37. Greenwald, J., Friedmann, M. P. & Riek, R. Amyloid aggregates arise from amino acid condensations under prebiotic conditions. *Angew. Chem. Int. Ed. Engl*. **55**, 11609–11613 (2016).
38. Rufo, C. M. et al. Short peptides self-assemble to produce catalytic amyloids. *Nat. Chem*. **6**, 303–309 (2014).
39. Wieczorek, R., Adamala, K., Gasperi, T., Polticelli, F. & Stano, P. Small and random peptides: an unexplored reservoir of potentially functional primitive organocatalysts. The case of seryl-histidine. *Life (Basel)*. **7**, 19 (2017).
40. Rout, S. K., Friedmann, M. P., Riek, R. & Greenwald, J. A prebiotic template-directed peptide synthesis based on amyloids. *Nat. Commun*. **9**, 234 (2018).
41. Bordin, N., Sillitoe, I., Lees, J. G. & Orengo, C. Tracing evolution through protein structures: nature captured in a few thousand folds. *Front. Mol. Biosci.* **8**, 668184 (2021).

42. Udenfriend, S. et al. Fluorescamine: a reagent for assay of amino acids, peptides, proteins, and primary amines in the picomole range. *Science*. **178**, 871–872 (1972).
43. Young, G, et al. Quantitative mass imaging of single biological macromolecules. *Science* **360**, 423-427 (2018).

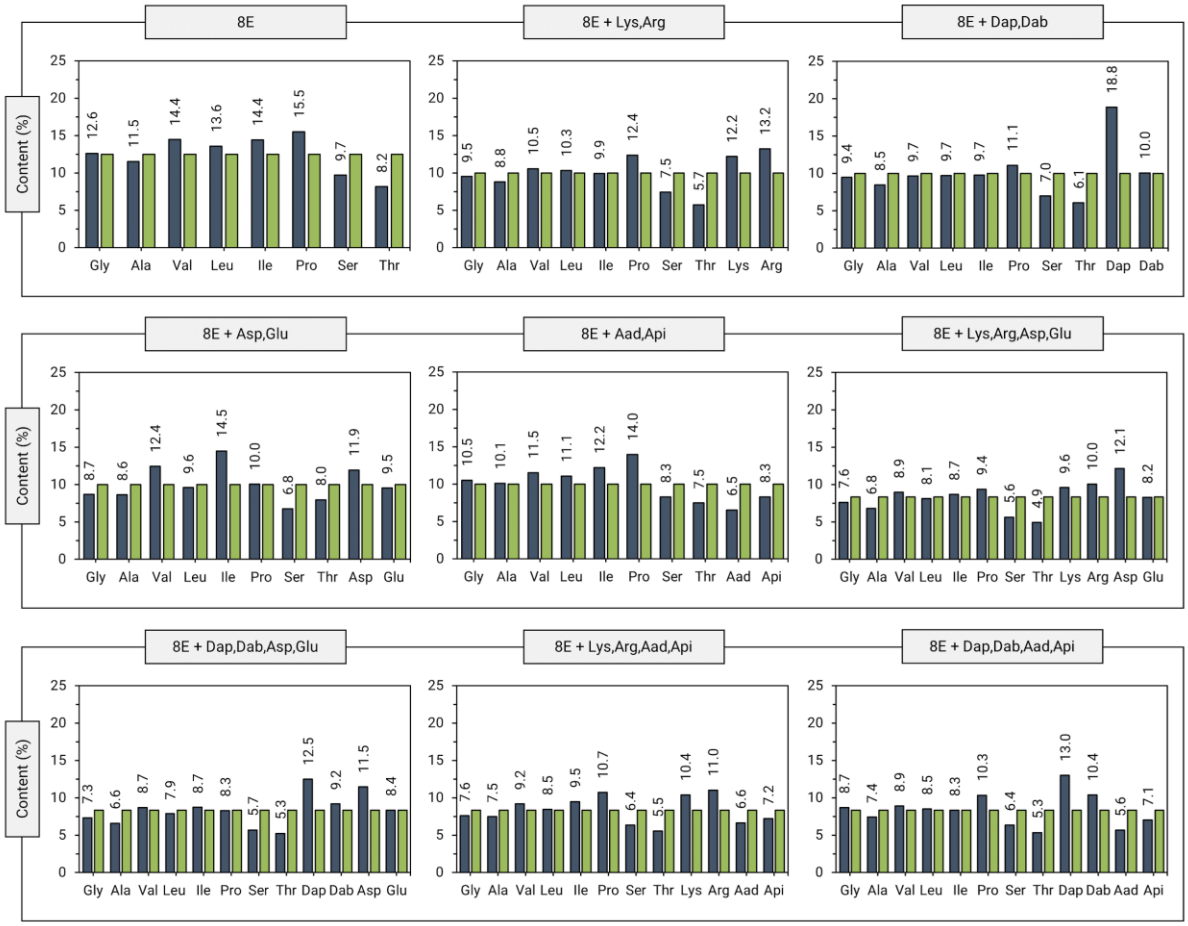
Self-assembly of random primordial acidic peptides as a stepping stone to foldable polymers

Mikhail Makarov, Robin Kryštůfek, Edgar Manriquez-Sandoval, Soumya Dutta, Tadeáš Kalvoda, Ján Michael Kormaník, Tatsiana Charnavets, Michal Lebl, Sean M. Brown, Jan Konvalinka, Abhishek Singharoy, Stephen D. Fried, Klára Hlouchová

Supplemental Figures

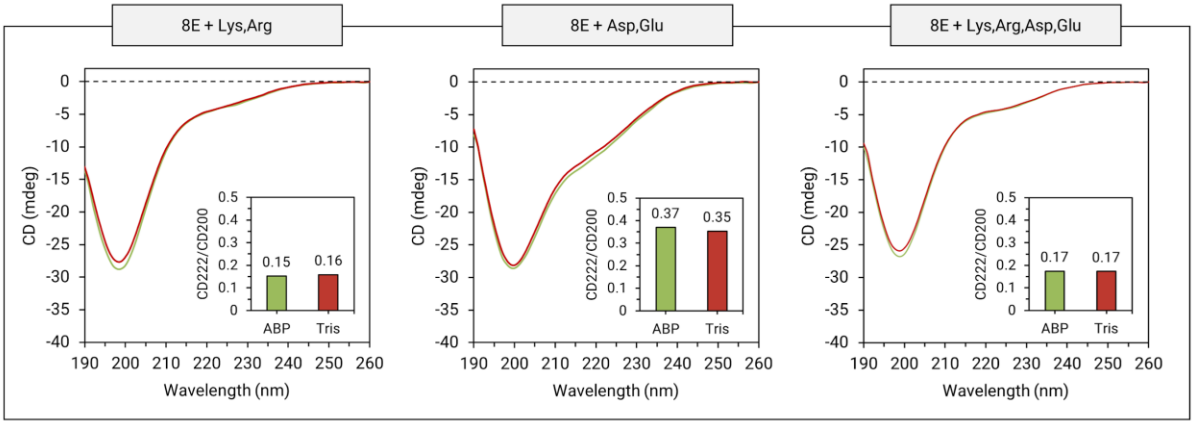


Supplemental Figure S1. Molecular weight distribution of nine indicated peptide libraries as assessed by MALDI.
MALDI spectra (blue) were obtained on 10 mg/ml solutions of peptide libraries dissolved in acetonitrile:water (1:1) mixture. The theoretical distributions (green) were obtained from 10,000,000 randomly generated sequences with equimolar amino acid distribution. For each library, the mean molecular weight (MMW) observed experimentally and the theoretical mean molecular weight are reported.

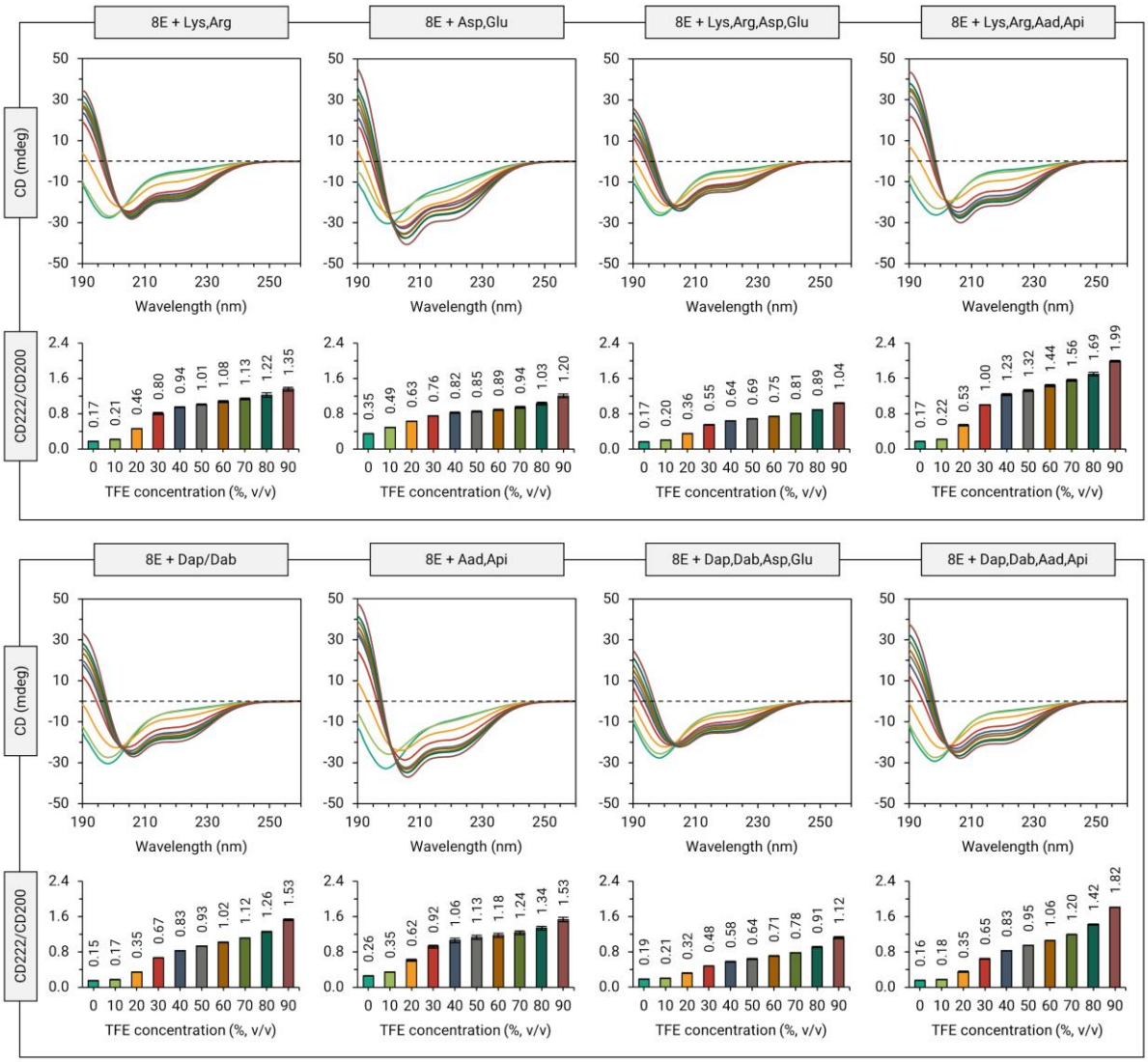


Supplemental Figure S2. The amino acid composition of nine indicated peptide libraries as assessed by HPLC amino acid analysis.

Theoretical amino acid distributions (in green) represent equimolar distribution of amino acids for each library, and experimental distributions (in blue) reflect integrated ion intensities from HPLC chromatograms.

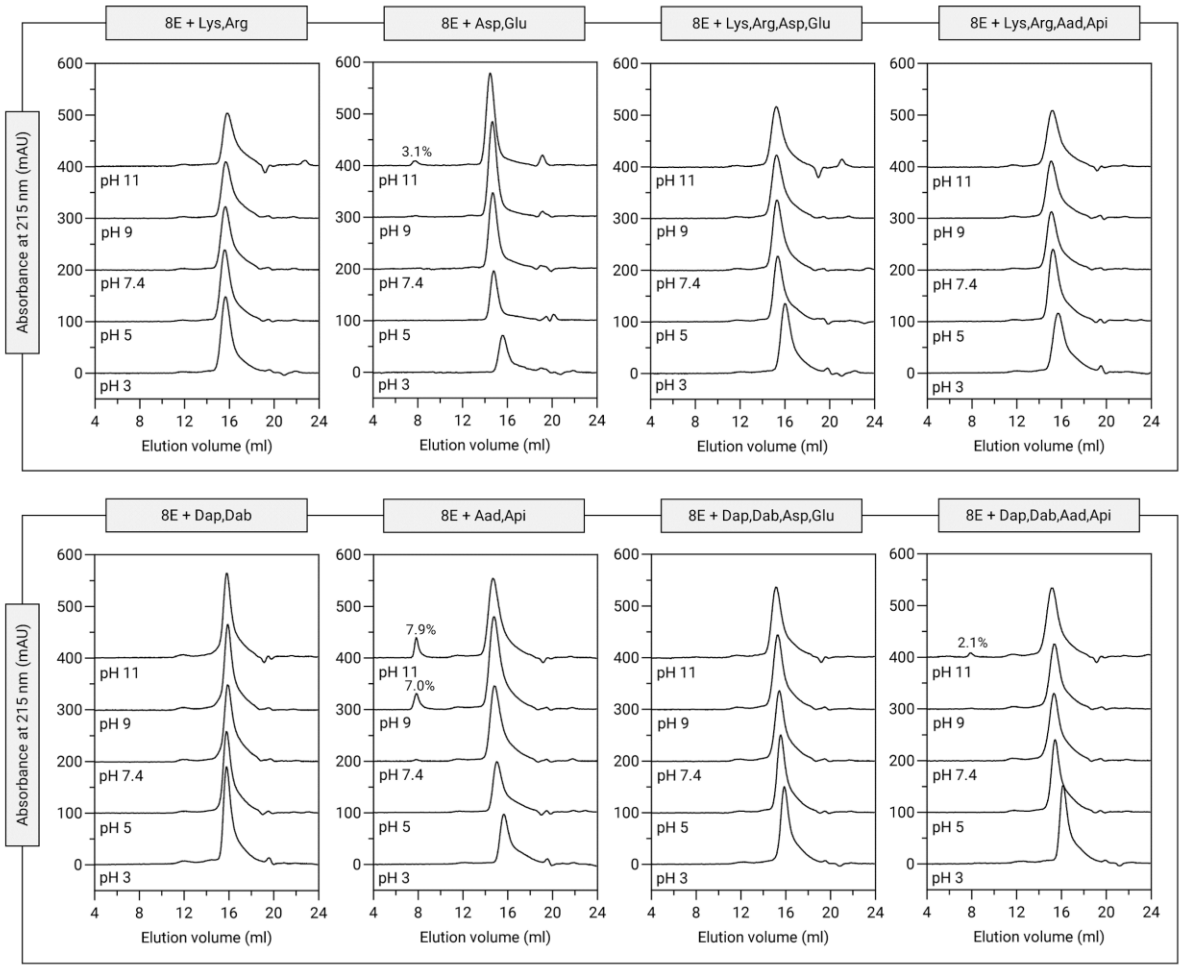


Supplemental Figure S3. The effect of buffer on peptide library secondary structure. The secondary structure of three indicated peptide libraries in 10 mM Tris-HCl (pH 7.4, red) and in 10 mM ABP (pH 7.4, lime) as assessed by far-UV circular dichroism (CD) spectroscopy. Secondary structure content is estimated as the ratio of the ellipticities at 222 nm and 200 nm as shown in the bar charts. No significant differences are observed between the two buffers.



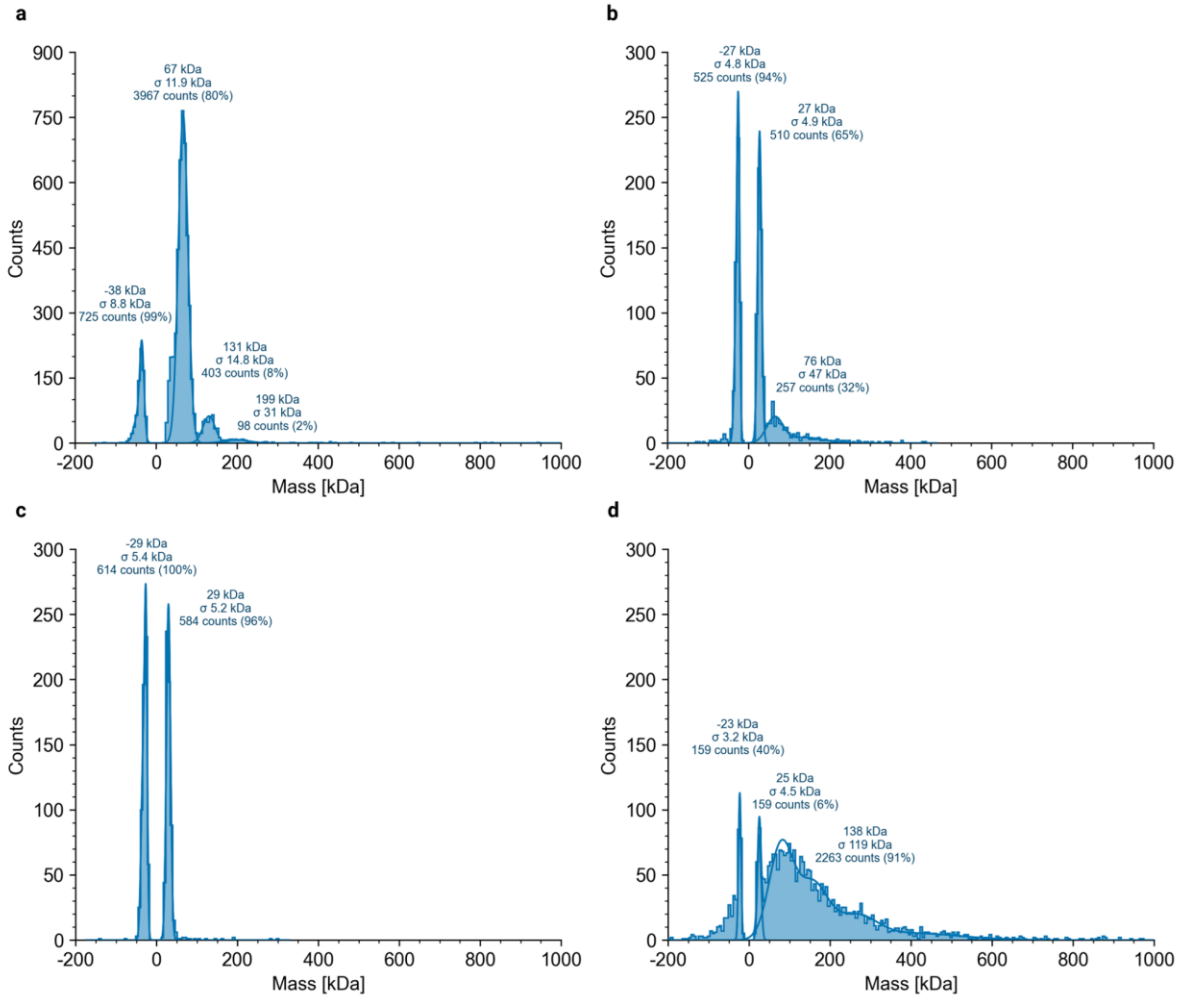
Supplemental Figure S4. Induction of secondary structure with 2,2,2-trifluoroethanol (TFE).

The secondary structure of eight peptide libraries at different TFE concentrations as assessed by far-UV circular dichroism (CD) spectroscopy. Far-UV CD spectra were measured in 10 mM ABP (pH 7.4) with 0-90% (v/v) TFE and are representative from duplicates. The secondary structure content is estimated as the ratio of the ellipticities at 222 nm and 200 nm as shown in the bar charts. Error bars represent standard deviations from duplicates. The data for 8E+Asp,Glu library were taken from ref. 17.

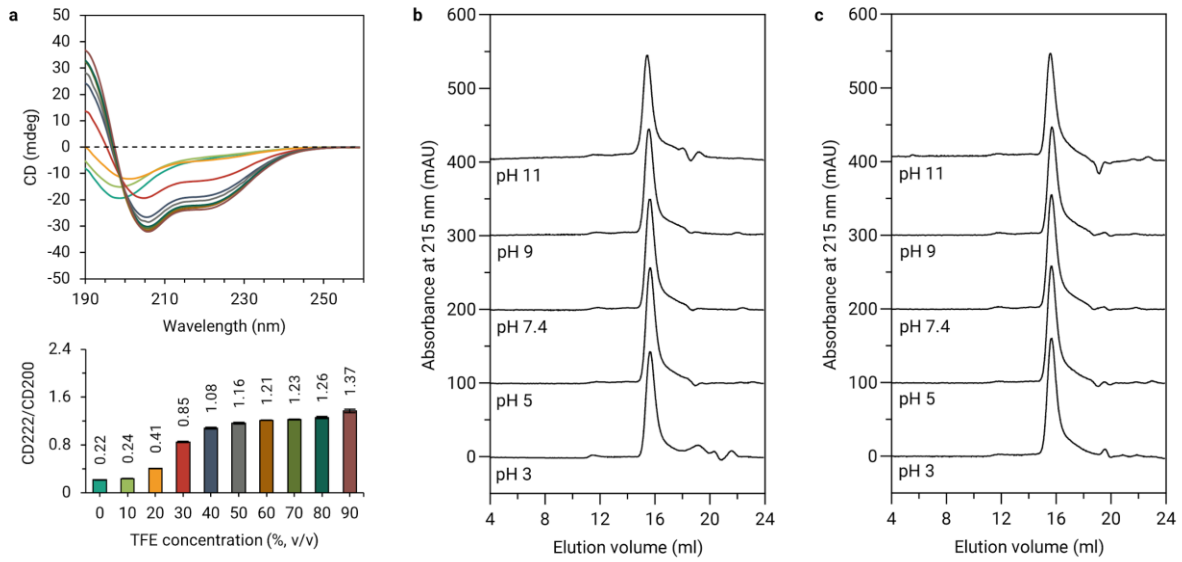


Supplemental Figure S5. The aggregation propensity of eight peptide libraries at high ionic strength.

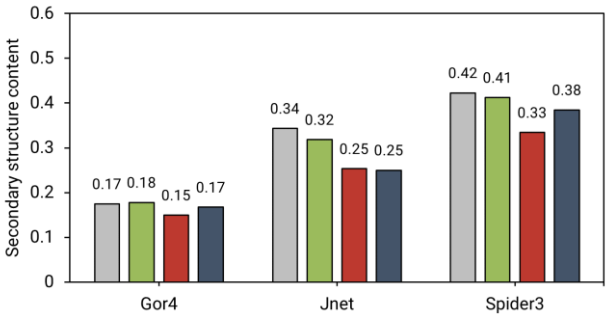
Aggregation propensity of the indicated libraries was measured in 20 mM ABP buffer at different pH's and at high ionic strength (500 mM NaCl) with size-exclusion chromatography (SEC). SEC chromatograms report absorbance at 215 nm as a function of elution volume. The data for 8E+Asp,Glu library were taken from ref. 17.



Supplemental Figure S6. Analysis of soluble aggregates with mass photometry. The molecular weight distribution of 8E+Asp,Glu (and its SEC-resolved fractions) in 10 mM Tris-HCl (pH 7.4), 50 mM NaCl as assessed by mass photometry. Mass histograms of: **a**, BSA standard; **b**, total fraction; **c**, monomer fraction; and **d**, aggregate fraction.

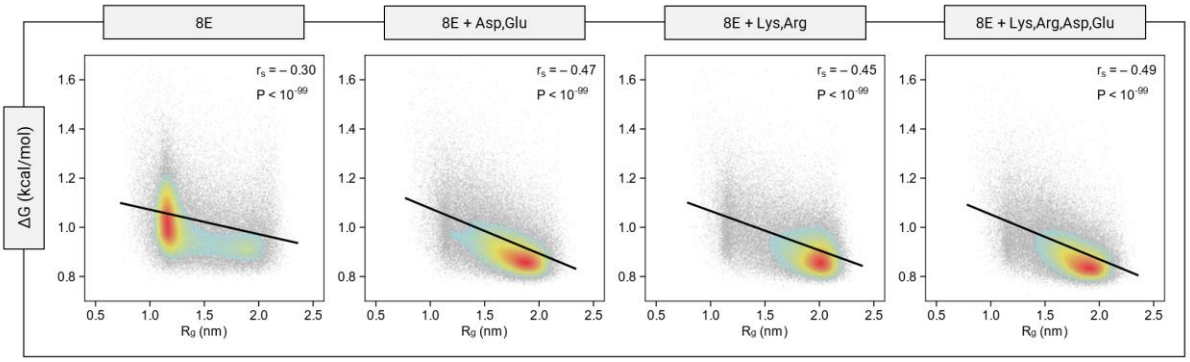


Supplemental Figure S7. Additional assays on the hydrophobic 8E peptide library.
(A) The secondary structure of the 8E peptide library at different 2,2,2-trifluoroethanol (TFE) concentrations as assessed by far-UV circular dichroism (CD) spectroscopy. Far-UV CD spectra were measured in 10 mM ABP (pH 7.4) with 0-90% (v/v) TFE and are representative from duplicates. The secondary structure content is estimated as the ratio of the ellipticities at 222 nm and 200 nm as shown in the bar charts. Error bars represent standard deviations from duplicates.
(B, C) Aggregation propensity of the 8E peptide library was measured in 20 mM ABP buffer at different pH's and at medium ionic strength (50 mM NaCl, B) and at high ionic strength (500 mM NaCl, C) with size-exclusion chromatography (SEC). SEC chromatograms report absorbance at 215 nm as a function of elution volume.



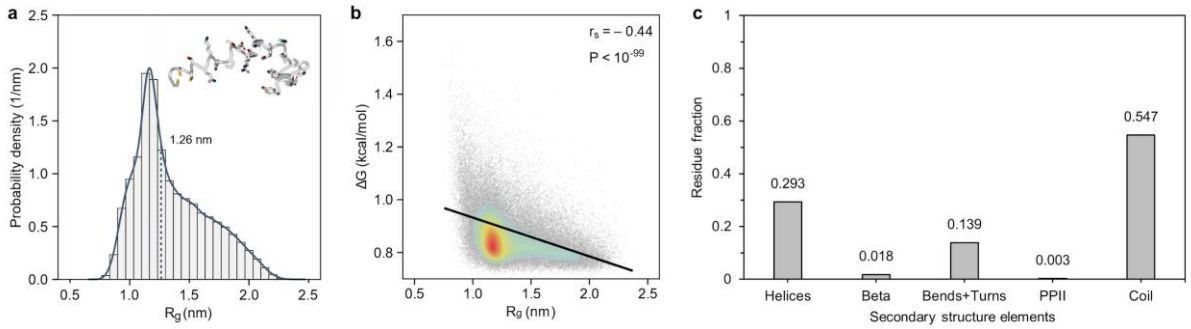
Supplemental Figure S8. Secondary structure assessed by Gor4, Jnet, and Spider3 structure predictors.

Bars showing percentage of residues (amino acids) predicted to have either α-helix or β-sheet propensity, averaged over 100,000 randomly generated 25-mer sequences for each library (8E, grey; 8E+Asp,Glu, lime; 8E+Lys,Arg, red; and 8E+Lys,Arg,Asp,Glu, blue).



Supplemental Figure S9. Correlation of predicted protein stability and radius of gyration.

Scatter plots showing the predicted free energy of unfolding[34] as a function of radius of gyration (R_g; GROMACS 2024.2) for 50,000 unique and randomly generated 25-mer peptide sequences predicted by ESMFold for the 8E, 8E+Asp,Glu, 8E+Lys,Arg and 8E+Lys,Arg,Asp,Glu libraries. Contours show 50% of all data with Spearman correlation coefficients provided.



Supplemental Figure S10. Computational results for random peptides using the full canonical 20-residue alphabet.

The structures of 50,000 unique and randomly generated 25-mer peptide sequences were predicted using ESMFold for a library uniformly using all 20 canonical amino acids.

(A) Distribution of the radius of gyration (R_g). Representative peptide structure centered at the distribution's median (dashed line) is shown.

(B) Scatter plots showing the predicted free energy of unfolding[34] as a function of R_g (GROMACS 2024.2). Contours show 50% of all data with Spearman correlation coefficient provided.

(C) Fraction of residues with secondary structure calculated by DSSP4.4. Helix comprises α-helix, π-helix, and 3_{10} helix; Beta comprises extended β sheets and β bridges; PPII, polyproline II helix.

Nucleic Acids Research, 2024, 1–12
https://doi.org/10.1093/nar/gkae823
RNA and RNA-protein complexes

The interplay between peptides and RNA is critical for protoribosome compartmentalization and stability

Simone Codispoti[1,†], Tomoko Yamaguchi[2,3,†], Mikhail Makarov[3], Valerio G. Giacobelli[3], Martin Mašek[4], Michal H. Kolář[4], Alma Carolina Sanchez Rocha[3], Kosuke Fujishima[2,5,*], Giuliano Zanchetta[1,*] and Klára Hlouchová[3,6,*]

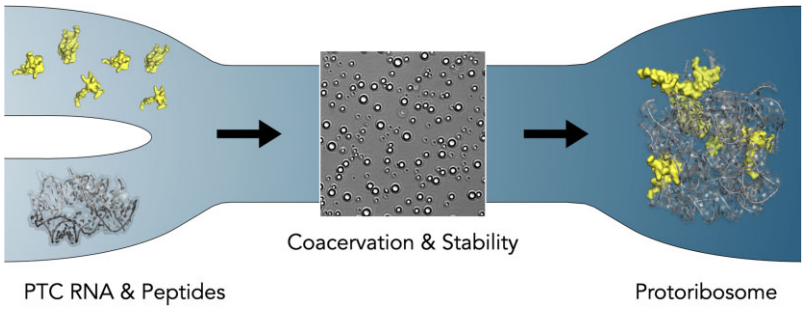
[1]Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università di Milano, Segrate 20054, Italy
[2]Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[3]Department of Cell Biology, Faculty of Science, Charles University, Viničná 7, Prague 12800, Czech Republic
[4]Department of Physical Chemistry, University of Chemistry and Technology, Technicka 5, 16628 Prague, Czech Republic
[5]Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa 252-0882, Japan
[6]Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Flemingovo n. 2, Prague 16610, Czech Republic

[*]To whom correspondence should be addressed. Tel: +420 221951801; Email: klara.hlouchova@natur.cuni.cz
Correspondence may also be addressed to Kosuke Fujishima. Email: fuji@elsi.jp
Correspondence may also be addressed to Giuliano Zanchetta. Email: giuliano.zanchetta@unimi.it
[†]The first two authors should be regarded as Joint First Authors.
Present address: Tomoko Yamaguchi, Department of Structural Biology, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan.

Abstract

The ribosome, owing to its exceptional conservation, harbours a remarkable molecular fossil known as the protoribosome. It surrounds the peptidyl transferase center (PTC), responsible for peptide bond formation. While previous studies have demonstrated the PTC activity in RNA alone, our investigation reveals the intricate roles of the ribosomal protein fragments (rPeptides) within the ribosomal core. This research highlights the significance of rPeptides in stability and coacervation of two distinct protoribosomal evolutionary stages. The 617nt 'big' protoribosome model, which associates with rPeptides specifically, exhibits a structurally defined and rigid nature, further stabilized by the peptides. In contrast, the 136nt 'small' model, previously linked to peptidyltransferase activity, displays greater structural flexibility. While this construct interacts with rPeptides with lower specificity, they induce coacervation of the 'small' protoribosome across a wide concentration range, which is concomitantly dependent on the RNA sequence and structure. Moreover, these conditions protect RNA from degradation. This phenomenon suggests a significant evolutionary advantage in the RNA–protein interaction at the early stages of ribosome evolution. The distinct properties of the two protoribosomal stages suggest that rPeptides initially provided compartmentalization and prevented RNA degradation, preceding the emergence of specific RNA–protein interactions crucial for the ribosomal structural integrity.

Graphical abstract



Introduction

Every cell capable of protein expression contains ribosomes. Due to their omnipresence and high conservation across all life forms, ribosomes qualify as ancient molecular fossils, attracting evolutionary biologists as the best connection to our biological past (1,2). According to several studies, the peptidyl transferase center (PTC) of the ribosome evolved ∼3.8–4.2 billion years ago, before the appearance of the last universal common ancestor. As a nest of RNA capable of polymerizing amino acids, the PTC (also known as protoribosome) is considered by some scientists to be a true cradle of life embedded in the ribosome, bridging prebiotic chemistry and protein-dominated biology (3). Although the PTC is made up of ribosomal RNA (rRNA), it is surrounded by fragments of

Received: February 1, 2024. **Revised:** September 4, 2024. **Editorial Decision:** September 5, 2024. **Accepted:** September 10, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

ribosomal proteins (rProteins) that lack secondary-structure motifs and are considered older than rProteins with globular structures found in the outer layers of the modern ribosome (4,5). A record of polypeptide interaction with the PTC can be found in these inner tails. With the onset of templated synthesis, the sequence of these polypeptides is presumed to have been fixed. They are considered to have expanded into proteins, along with the rRNA, following an accretion process. This is known as the onion model for ribosome evolution (5).

Much of the progress in the study of the protoribosome has resulted from the explosion of resolved ribosomal structures over the last two decades (reflected also in the 2009 Nobel Prize in Chemistry awarded to V. Ramakrishnan, T.A. Steitz and A.E. Yonath). Based on these, several groups have proposed regions of the 23S rRNA in the large subunit, ranging ∼100–600 nucleotides (nt) in length, as the ancestral PTC (3,6–9). Most importantly, two independent groups have recently demonstrated the peptidyl-transferase activity of their ∼70–140 nt rRNA constructs, mimicking the semi-symmetrical PTC pocket (10–12). Unlike some of the larger PTC constructs, these minimal protoribosome models were devoid of peptides, proving the rRNA sovereignty in the catalysis of amino acid polymerization.

Although some scientists consider peptides relevant only after the evolution of templated protein synthesis, peptides have been repeatedly reported prebiotically plausible, supported by the abundance of amino acids and the facile nature of their condensation reaction (13). Some of the ribosomal protein fragments/tails (rPeptides) have been implied to interact with a larger model of the PTC and form a specific RNA–peptide assembly (9). Additionally, ribosomal peptides have been reported to enhance RNA polymerase ribozyme function (14).

Here, we report the key role of ancient rPeptides in protoribosome stabilization and coacervation. These prebiotically important properties depend on the RNA length and secondary structure content but they are predominantly influenced by the interplay between RNA and peptides. In this context, we propose a scenario in which peptides facilitated protoribosome accumulation within liquid droplets, thereby fostering the coevolution of a compartmentalized core for an active RNA–peptide world with a prospect of a true protocell, as proposed by Aleksander Oparin a whole century ago (15,16).

Materials and methods

Synthesis of PTC rRNAs and peptides

We adopted two published models of the PTC from *Thermus thermophilus*, namely WT bPTC (referring to a 'big' 617 nt rRNA construct) and WT sPTC (referring to a 'small' 136 nt rRNA construct) (9–11) (Figure 1). In both cases, the constructs stem from segmented fragments of the 23S rRNA, connected with loops to preserve the 3D structures. WT bPTC is made up of the 505 nt from the 23S rRNA, joined by 11 stem loops (5′- gccGUAAggc-3′); WT sPTC is composed of 116 nt from the 23S rRNA and connected with three stem loops (5′-CUUCGG-3′) (Supplementary Figure S1). As a control, shuffled versions of both PTC rRNAs were designed by randomizing their sequences while keeping the same composition of A, U, G and C (Supplementary Figure S2). WT_bPTC and Sh_bPTC ssDNA oligonucleotides were purchased from Integrated DNA Technologies (IDT), and the remaining ssDNA oligos were purchased from Sigma-Aldrich (Supplementary Table S1).

For the synthesis of WT and Sh_bPTC rRNAs, dsDNA templates were prepared by polymerase chain reaction (PCR) amplification of WT_bPTC and Sh_bPTC ssDNA oligos using Q5 Hot Start High-Fidelity DNA Polymerase (NEB) according to the manufacturer's instructions. PCR amplification was performed using bPTC_F and bPTC_R primers according to the following program: initial denaturation at 98°C for 30 s; denaturation at 98°C for 10 s; annealing at 69°C for 30 s; extension at 72°C for 60 s; final extension at 72°C for 20 s; 32 cycles.

For the synthesis of WT sPTC rRNA, dsDNA template was prepared by annealing of WT_sPTC_F and WT_sPTC_R ssDNA oligos (Supplementary Table S1) followed by DNA synthesis using Klenow Fragment of DNA polymerase I (NEB). For annealing, 15 pmol of WT_sPTC_F and WT_sPTC_R ssDNA oligos (8.1 μg of DNA in total) were mixed in NEB2 buffer to the total volume of 47 μl, denatured at 80°C for 5 min and cooled down to room temperature at 0.1°C/s rate. Then, 1 μl of 5U/μl Klenow Fragment of DNA polymerase I (NEB) and 2 μl of 10 mM dNTP mix (SERVA) were added to the reaction mixture, and the reaction mixture was kept at 25°C for 1 h followed by heat inactivation of enzyme at 50°C for 15 min. For the synthesis of Sh1 and Sh2 sPTC rRNAs, 40 pmol of Sh1_sPTC_R/Sh2_sPTC_R ssDNA and T7_F (Supplementary Table S1) was mixed in H_2O to the total volume of 36 μl, incubated at 95°C for 1 min and then kept at room temperature for 2 min; this was used as DNA template (1.9 μg of DNA in total).

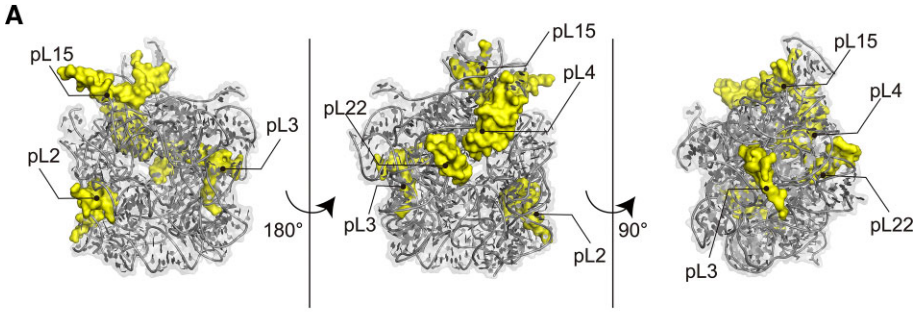
The resulting dsDNA templates were purified from 2% (Tris-acetate-EDTA)-agarose gel using Monarch Gel Dissolving Buffer (NEB) and Monarch PCR&DNA Cleanup Kit (5 μg) (NEB). The purified dsDNA templates were transcribed using HiScribe T7 High Yield RNA Synthesis Kit (NEB) according to the manufacturer's instructions, and the resulting RNAs were isolated from the reaction mixtures by Monarch RNA Cleanup Kit (NEB). The size and purity of the RNA constructs was confirmed by (Tris-Borate-EDTA)-urea PAGE (Supplementary Figure S3). Due to the differences in electrophoretic mobility, the sizes of the RNA constructs were additionally confirmed by matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) (Supplementary Figure S4). All RNA samples were stored at −80°C before use.

The rPeptides were synthesized by the Spyder Institute using standard protocols for solid-phase peptide synthesis. The identities and purities of the peptides were confirmed by mass spectrometry using UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) according to the standard procedure (Supplementary Figure S5).

Peptide structural characterization

Circular dichroism (CD) spectra of rPeptides were recorded using a Chirascan-plus spectrophotometer (Applied Photophysics, Leatherhead, UK) over the wavelength range 190–260 nm in steps of 1 nm with an averaging time of 1 s per step (Supplementary Figure S6A). Cleared peptide samples at 0.2 mg/ml concentration in 10 mM Tris (pH 7.5), 1 mM KCl, 1 mM $MgCl_2$ and 1 mM $CaCl_2$ were placed in 1 mm path-length quartz cells, and spectra were recorded at room temperature. The CD signal was obtained as ellipticity

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024



C

Sequence of peptides

Peptides	Sequence
pL2	GRRPHVRGAAMNPVDHPHGGGEGRAPRGR
pL3	RHPGSIGNRKTPGRVYKG
pL4	YSGRKIWPQKHTGRARHGDIGAPIFV
pL15	PGSGHGKTATRGHKGQKSRSGGLKD
pL22	LKRVLPRARGRADIIKKR

Figure 1. PTC models and rPeptide sequences. (**A** and **B**) The models of WT bPTC (A) and WT sPTC (B), represented as ribbon, and rPeptides (pL2, 3, 4, 15, 22). Both models are extracted from the *Thermus thermophilus* ribosomal structure (PDB ID: 4V51), missing the loops that were added by authors of the designed models (9,11). The two symmetrical A and P regions of the WT sPTC (B) are depicted with lighter and darker shades of grey, respectively. (**C**) Sequences of rPeptides used in this study.

in units of millidegrees, and the spectra were averaged from two scans and buffer-spectrum subtracted. The resulting CD spectra were then expressed as molar ellipticity per residue θ (deg·cm^2·dmol^{-1}). All CD measurements were performed twice for every rPeptide and curves were then averaged.

Characterization of rPeptide–rRNA interaction by microscale thermophoresis

In order to determine the dissociation constant of rPeptide–rRNA complexes, the 3′-end of the RNA constructs was fluorescently labelled with fluorescein-5-thiosemicarbazide (FTSC) using a published protocol (17) with slight modifications. First, 500 pmol of PTC rRNA was oxidized by mixing with 10 nmol $NaIO_4$ (20 eq.) and 100 mM potassium acetate (pH 5.2) to a total volume of 50 µl and incubating for 90 min at room temperature in the dark. The oxidized PTC rRNA was subsequently purified from the reaction mixture with Monarch RNA Cleanup Kit (NEB) and immediately mixed with 150 nmol FTSC (300 eq.) dissolved in DMF and 100 mM potassium acetate (pH 5.2) to a total volume of 50 µl. The reaction mixture was incubated overnight at 4°C in the dark, and the excess of FTSC was removed with Monarch RNA Cleanup Kit (NEB).

The interaction of the five rPeptides with both WT PTC variants was measured using microscale thermophoresis (MST). Experiments were conducted in duplicate on a Monolith NT.115 system (NanoTemper Technologies). To estimate the effect of metal ions, rPeptide solutions were prepared in 20 mM Tris-HCl (pH 7.5), while PTC rRNA solutions were prepared in two conditions: (i) 20 mM Tris-HCl (pH 7.5) or (ii) 20 mM Tris-HCl (pH 7.5), 2 mM KCl, 2 mM $MgCl_2$, 2 mM $CaCl_2$. Prior to the experiment, PTC RNAs were heated at 85°C for 30 s and then slowly cooled to room temperature over 30 min, to allow the molecules to refold.

A 2-fold dilution series of the unlabelled rPeptides was prepared with the concentrations of rPeptides ranging from 1 mM to 61.04 nM. Ten microliters of rPeptide solutions were subsequently mixed with 10 µl of 100 nM fluorescein-labelled PTC rRNA and the samples were incubated for 15 min at room temperature. Following incubation, the samples were filled into standard-treated capillaries (NanoTemper Technologies) and subjected to MST analysis. The MST measurements were performed at 1% LED power (green channel) and 40% MST power, laser-on time 20 s, laser-off time 5 s. The results were analysed by TJump analysis, and the normalized fluorescence values were plotted against the rPeptide concentration. Dissociation constants (Figure 2) and binding stoichiometries were then estimated adopting both a single-site model and a cooperative Hill model to fit the curves (Supplementary Figure S7).

Molecular dynamics simulations

Several molecular systems were studied using all-atom molecular dynamics (MD) simulations, namely WT bPTC, WT

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

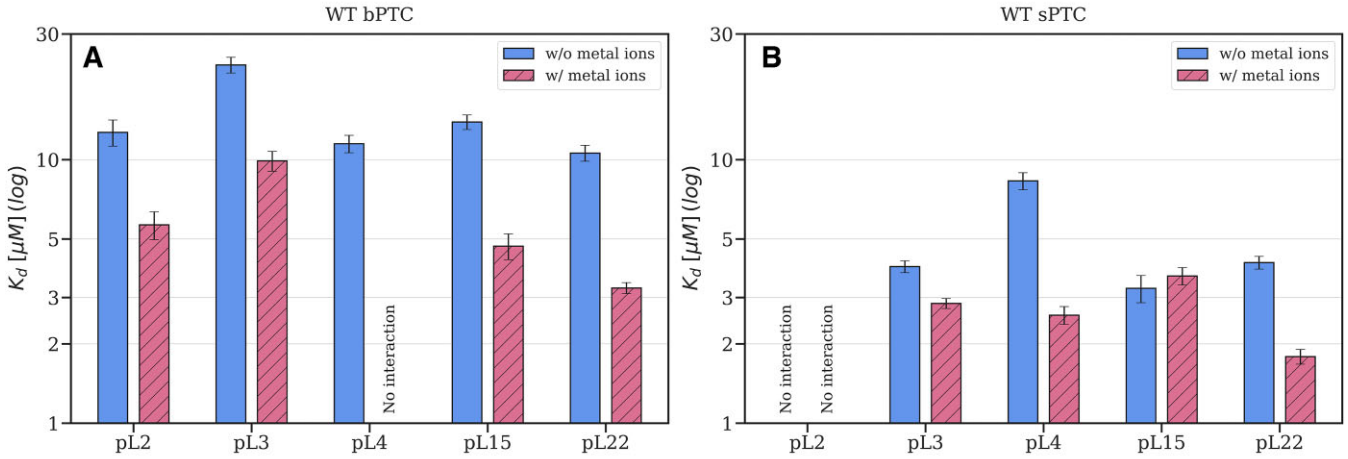


Figure 2. Summary of the WT bPTC (**A**) and WT sPTC (**B**) dissociation constants (K_d) with the rPeptides determined using MST in the absence and presence of metal ions (1 mM Mg^{2+}, K^+ and Ca^{2+}). Error bars represent the semi-dispersion of the two replicates; 'no interaction' indicates the cases where the potential K_d value is outside the measurable range of 30 nM to 500 μM or the dissociation curve could not be fitted.

bPTC + rPeptides (pL2, 3, 4, 15 and 22), WT sPTC, WT sPTC + pL2, WT sPTC + pL3, WT sPTC + pL4 and WT sPTC + rPeptides (pL2, 3 and 4). The initial atomic coordinates were obtained by shaving the X-ray structure of *Thermus thermophilus* ribosome (PDB 4V51) (18) and interconnecting the flanking rRNA ends by stem loops as detailed in (9) and (11). The sequences of rPeptides were adjusted to match our experiments. Structural Mg^{2+} ions were used as found in the X-ray structure within the cut-off distance of 0.5 nm from the selected regions. The systems were placed in a periodic rhombic dodecahedral box ensuring the distance of 1.5 nm between the box face and the solute. The box was filled with explicit water molecules and K^+ and Cl^- ions of the concentration 150 mM. Finally, each box was neutralized by adding enough K^+ ions.

The system was described by the Amber family of force fields: ff10 for rRNA (19,20), ff12SB for peptides (21), Joung and Cheatham ions (22), and SPC/E water (23). The systems were energy minimized in ∼80 000 steps of the steepest descent algorithm and subsequently heated to 300 K during 100 ps simulations. The initial velocities were randomly drawn from the Maxwell-Boltzmann distribution at 10 K. The pressure and density were equilibrated during another 100 ps simulations. During minimization and equilibration, position restraints (k = 1000 kJ mol^{-1} nm^{-2}) were applied to all heavy atoms of the solute to prevent its spurious conformational changes. The unrestrained production runs were carried out at constant temperature of 300 K and pressure 1 bar, using a v-rescale thermostat (24) and Parrinello-Rahman barostat (25). The long-range electrostatics were treated by particle mesh Ewald algorithm; van der Waals interactions were described by the Lennard-Jones potential with the cut-off of 1.0 nm. A virtual-site algorithm was used to hydrogen atoms allowing for the integration time step of 4 fs. Three independent trajectories were obtained for each system differing in the initial set of velocities. We collected trajectories of 1100 ns each with trajectory frames saved every 100 ps.

Prior further analyses, all trajectories were superimposed to the same reference conformation taken from the initial X-ray structure of the modern ribosome. This made bPTC and sPTC constructs better comparable. A structural analysis involved the last 300 ns of each trajectory. The atomic coordinates were averaged over the simulation time and the three independent trajectories, which provided MD-based structural models of the WT constructs. Structural differences of the models were represented by Euclidean distances between equivalent atoms of the MD models. For the fluctuation analysis, the first 300 ns of each trajectory were skipped and the root-mean-square fluctuations (RMSFs) were calculated over the rest of the simulation time. The RMSFs were averaged over the independent trajectories and projected onto the initial structure.

All simulations were performed in the GROMACS package (26). The input data necessary for reproducing our MD simulations are available at: https://github.com/mhkoscience/codispoti-protoribosome.

Coacervation

RNA solutions were prepared in RNase free water in small aliquots (1–20 μl) and at various concentrations, and they were stored at −80°C. Lyophilized rPeptides were suspended in RNase free water to typical concentrations of 50 mM. These stock solutions were further diluted at different concentrations, divided in small aliquots and stored at −20°

Phase separation diagrams were estimated as a function of the $R_{+/-}$ ratio and of the molar charge concentration (in units of the elementary charge) Q of the mixtures (Figure 4 and Supplementary Figures S8 and S9), defined as:

$$R_{+/-} = \left(\sum_{p=1}^{N} C_p\, z_p\right) \;/\; (C_{RNA}\, z_{RNA}) \qquad (1)$$

$$Q = \sum_{p=1}^{N} C_p\, z_p + C_{RNA}\, z_{RNA} \qquad (2)$$

where C_p and C_{RNA} are the molar concentrations of rPeptides and rRNA, respectively, and the summation runs over the different rPeptides employed. The $R_{+/-}$ ratio represents the stoichiometric charge ratio of the molecules in solution (positive over negative charges), while Q denotes the molar concentration of charged units. z_{RNA} is defined as the

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

length of the strand minus 1 (the RNA terminals are not phosphorylated), while the net peptide charges (z_p) are estimated using the pH 7.5 values given by the *Peptide Property Calculator* by NovoPro (https://www.novoprolabs.com/tools/calc_peptide_property): pL2 = +4.1, pL3 = +5.0, pL4 = +4.0, pL15 = +5.0, pL22 = +7.0. The typical $R_{+/-}$ values explored in the phase separation diagrams range from 0.1 to 5, while Q ranged between 0.1 and 2 mM. When performing experiments in which more than one rPeptide was involved, all the peptide concentrations were set equal and the desired $R_{+/-}$ and Q values were tuned by changing the RNA and the overall common peptide concentration.

In order to explore different values of $R_{+/-}$ and Q, mixtures of the WT bPTC and WT sPTC and of pL2, 3, 4, 15 and 22 were prepared at various working concentrations, at room temperature and by keeping the ionic strength fixed. The same buffer was used in all experiments: 20 mM Tris pH 7.5, 1 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$. The ionic strength of the buffer, accounting both for the Tris ionic strength and for the dissociated monovalent K^+ and Cl^- and divalent Mg^{2+} and Ca^{2+} ions, is estimated to be 23 mM in equivalent NaCl concentration. Mixtures of pL22 with either WT bPTC or WT sPTC were prepared by adding, in this given order, 1 µl of 8× buffer, 3.5 µl of peptide and 3.5 µl of RNA, while mixtures of more than one rPeptides with RNA were prepared in total volumes of 8 µl by adding the same amount of each peptide, adjusting with 1 µl of 8× buffer and by final addition of RNA. All the necessary dilutions were carried out in RNase-free water and samples were prepared in small PCR plastic tubes. Before adding it to the peptides-buffer mixture, the RNA was annealed in a thermal cycler from 85°C to 40°C in 30 min in 3°C steps to refold the molecules. The RNA solution was then left reaching room temperature before mixing it to the rPeptide(s).

Liquid–liquid phase separation (LLPS) phase diagrams were characterized through bright field microscopy. For imaging, samples were transferred in a plastic multiwell which was previously treated with a bovine serum albumin (BSA) 3% in mass solution (the BSA solution was added to each well, removed after 20 min and then washed two times with water). Well plates were sealed with PCR plastic films and microscopy measurements were carried out at room temperature. Samples were inspected at the Imaging Methods Core Facility at BIOCEV using a Leica DMi8 WF microscope. Images were taken at different magnifications (10×, 20×, 40×) and the phase state was annotated for further characterization of the phase diagrams.

Testing of rRNA stability

A solution of 2 µM WT sPTC RNA was incubated with or without pL22 (39.0 µM, R = 1.0) or pL22 = (10.5 µM, R = 0.27), or a mixture of pL2, 3, 4, 15 and 22 (10.5 µM of each peptide, R = 1.0) in Buffer A (20 mM Tris-HCl pH 7.5, 1 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$) in the total volume of 10 µl. The 2 µM solution of WT bPTC was incubated with or without pL22 (70.3 µM, R = 0.4) or pL22 = (18.8 µM, R = 0.1), or a mixture of pL2, 3, 4, 15, 22 (18.8 µM of each peptide, R = 0.4) in Buffer A. In order to perform the RNase stability test, mixtures were divided into halves and incubated for 1 h at room temperature. Monarch RNase A (NEB) (0.1 ng/ml in final concentration) was added, and both samples with and without RNase A mixtures were incubated at 37°C for 30 or 40 min.

The reactions were stopped by adding 80 units/µl of 'Proteinase K, Molecular Biology Grade' (NEB) and incubated for 5 min at 37°C. Reaction products were analysed using 10% Urea PAGE. Gels were stained by GelRed (Biotium) and visualized by UV. Band intensities were evaluated by Fiji (Schindelin). The experiments were performed in triplicates and statistical analyses were done using Excel 2023 (Microsoft). Comparisons between datasets were performed using a two-tailed Wilcoxon–Mann–Whitney test with $n = 3$. A P-value of $\leqq 0.2$ and $u = 0$ was considered to be a statistically significant difference.

Results

rPeptides are devoid of structural elements and interact with PTC rRNA

The WT bPTC directly interacts with a number of rPeptides within the ribosome and their parts are intertwined within the RNA structure (Figure 1A). All these rPeptides are highly conserved within the bacterial domain of life and also preserve their positive charge pattern in the sequence beyond the bacterial domain except pL4 (Supplementary Figure S10). The WT sPTC, on the other hand, is much less dependent on these rProtein fragments (Figure 1, panels A and B). To resolve their role during ribosomal evolution, the peptide sequences resembling these rPeptides were synthesized and purified (Supplementary Figure S5). Some of the sequences (pL2, pL3 and pL22) were adapted precisely from a previous study (9), while pL4 and pL15 were truncated based on our structural analysis to best preserve their bPTC-binding properties and to allow for the solid-phase synthesis of all of them (Figure 1C). All the rPeptides are well soluble and their CD spectra confirm a general lack of secondary structure arrangement (Supplementary Figure S6A). Within ribosomal structures, these rPeptides also lack secondary structure motifs, as has been reported by previous studies (27). Only the fragment of pL22 probably undergoes moderate structural transition towards β-sheet formation, within the context of the full ribosomal protein and upon binding to rRNA (Supplementary Figure S6B).

The interaction of the five rPeptides with both WT PTC variants was measured using MST. The 3′ end of the PTC rRNA was labelled by FTSC and purified. To refold its 3D structure, PTCs were kept heated at 85°C for 30 s and gradually cooled to room temperature. The refolded PTCs were mixed with different concentrations of the rPeptides to detect dissociation constants (K_d). This was done in both the absence and presence of 1 mM Mg^{2+}, K^+ and Ca^{2+}; the choice of these ions is based upon their prebiotic abundance and prevalence in extant ribosomal structures. We hypothesized that they may play an important role in tuning the interaction between heavily charged molecules, such as Arg-rich peptides and RNA.

Both PTC variants interact with most of the rPeptides with K_d in µM range (Figure 2 and Supplementary Figure S7). Dissociation constants could not be estimated only for pL4 with WT bPTC and for pL2 with WT sPTC in presence of metal ions. While the binding curves remain almost unchanged for the WT sPTC interaction with rPeptides upon addition of metal ions, the WT bPTC interaction with most of the rPeptides is visibly affected. Based on the shape of

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

the binding curves and Hill coefficient estimates, rPeptides (with the exception of pL4) seem to bind to WT bPTC with a decreased peptide:RNA stoichiometry in the presence of metal ions, with K_d values decreasing 2–3 times (Figure 2 and Supplementary Figure S7).

rPeptides stabilize the bPTC fold while the sPTC is more flexible and structurally different from the bPTC

The structural properties of the PTC constructs and the effects of the rPeptides were further explored by all-atom MD simulations. First, we performed a series of MD simulations of WT constructs to check their conformational flexibility and interactions of their components.

During microsecond-long simulations of WT bPTC with rPeptides, all of the rPeptides remained well associated with the rRNA, suggesting low dissociation rates. We observed no large-scale conformational changes of the rRNA. Fluctuation analysis showed that the most conformationally variable regions are the stem loops, peptide pL15 and some residues in the ancestral ribosome tunnel (Figure 3A, highlighted in red). To reveal the role of peptides in the bPTC conformational dynamics, we performed MD simulations of WT bPTC without all rPeptides and analysed the root-mean-square deviation (RMSD) of phosphorus atoms with respect to the initial conformation. A comparison of the two bPTC systems revealed that the rPeptides reduce RMSD of the rRNA (Figure 3B). Because the reference conformation for the RMSD calculations was the same for both WT bPTC simulations, we conclude that the rPeptides conformationally stabilize the rRNA. However, the largest structural differences of rRNA related to the presence/absence of the rPeptides in the WT bPTC are located mostly on the surface of the construct (Supplementary Figure S11).

Further, we identified the largest common rRNA element (113 nt, Figure 3C), which is present in both bPTC and sPTC, and also in the modern ribosome. It includes the nucleotides nearest to the catalytic centre that form the ancestral exit tunnel and the binding sites for transferRNA (tRNA) predecessors. We calculated the RMSD of phosphorus atoms of the common rRNA element with respect to the initial conformation taken from the modern ribosome. A comparison of the RMSDs illustrates that the catalytic centre is conformationally more diverse in WT sPTC than in WT bPTC (Figure 3D). Because the reference conformation was taken from the modern ribosome, it can also be concluded that the catalytic centre of the bPTC is structurally more similar to the modern ribosome than the catalytic centre of the sPTC.

Finally, the MD simulations reported high conformational variability of the sPTC constructs also in the presence of rPeptides pL2, 3 and 4. Although we observed some difference in the probability distributions of RMSDs, on the microsecond time scale they seem not to be statistically significant.

rPeptides drive coacervation of protoribosome constructs

While WT bPTC and WT sPTC are fully soluble in the buffer when in solution as single species, the addition of equimolar mixtures of rPeptides (verified to be soluble at the same average concentrations used in rPeptides–RNA mixtures) triggers a rich phase behaviour, including precipitation of solid particles, formation of branched aggregates and demixing into liquid-like droplets (Figure 4, micrographs A i, A ii and B i). In particular, we investigated the onset of LLPS, or coacervation, as a function of the total concentration of RNA and peptides on the y-axis (expressed in Figure 4A iv and 4B iv as the inverse of the molar charge concentration) and of their ratio (peptides over RNA) on the x-axis, namely $R_{+/-}$ (expressed as the ratio of positive and negative charges, see 'Materials and Methods', coacervation, equations 1 and 2), as sketched in Figure 4C. For both WT bPTC (Figure 4A, micrograph iv) and WT sPTC (Figure 4B, micrograph iv), coacervation is observed at moderate total concentrations (as low as 2 μM RNA) and at high enough relative concentration of rPeptides ($R_{+/-} > 0.3$ and 0.7, respectively). However, bPTC coacervation is suppressed for $R_{+/-} > 0.75$, while it is observed for sPTC over a wide range of $R_{+/-}$ values. Qualitatively similar results are obtained for mixtures of both protoribosomes with a single rPeptide, namely pL22 (the rPeptide with the highest ratio of positively charged residues), corroborating the idea that the phase behaviour is determined by the different rRNA constructs involved (Supplementary Figure S8). In addition, we observed that WT sPTC coacervates with pL22 are stable up to 90°C: this is consistent with an 'upper critical solution' phase behaviour, where enthalpic interactions are the main drivers of LLPS. The fluidity (and relatively small viscosity) of the condensed phase was confirmed by Fluorescence Recovery After Photobleaching (FRAP) using the TAMRA-labelled pL3 and WT sPTC system. Full recovery of fluorescence in the bleached region was observed in the timescale of ∼20 s (Supplementary Figure S12).

The coacervation of rRNA-rPeptide protoribosomes is dependent on rRNA structure and composition

The complexation of positively charged rPeptides and negatively charged PTC is clearly at the basis of the observed LLPS, as previously reported for a variety of other nucleic acid-amino acid systems (28–31). However, we found that the sequence composition of the employed RNA strands, as well as their secondary and tertiary structures, sensitively affect coacervation propensity. We mixed rPeptides with two sPTC variants of shuffled sequences, so that the predicted fraction of paired bases was either higher or lower than in the original WT sPTC and the symmetric structure of the sPTC was disrupted (namely Sh1 sPTC and Sh2 sPTC, respectively, Supplementary Figure S2). Similarly, a single shuffled variant (Sh bPTC) was designed also by permutation of WT bPTC sequence. In such cases, substantial weakening of the LLPS propensity was found, as shown by the optical micrographs A iii, B ii and B iii compared to A i and B in Figure 4, corresponding to the inner LLPS regions for WT bPTC and sPTC, marked by 'X'. The dissimilarity in the phase behaviour extends well beyond these points, as shown in Supplementary Figure S9 for the Sh bPTC variant: aggregation or precipitation occur at all examined samples, highlighting the subtle dependence of LLPS on RNA secondary and tertiary structure.

To further examine the effect of nucleotide composition, various unstructured single-stranded DNA oligomers with different A-T-C-G compositions were mixed with the pL3 peptide. The utmost LLPS propensity occurs when the nucleotide composition closely resembles the native rRNA nucleotide environment of the pL3 peptide (namely, A = 10.7%, C = 39.3%, G = 32.1% U = 17.9%) and LLPS stability monotonically decreases as the composition deviates from it,

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

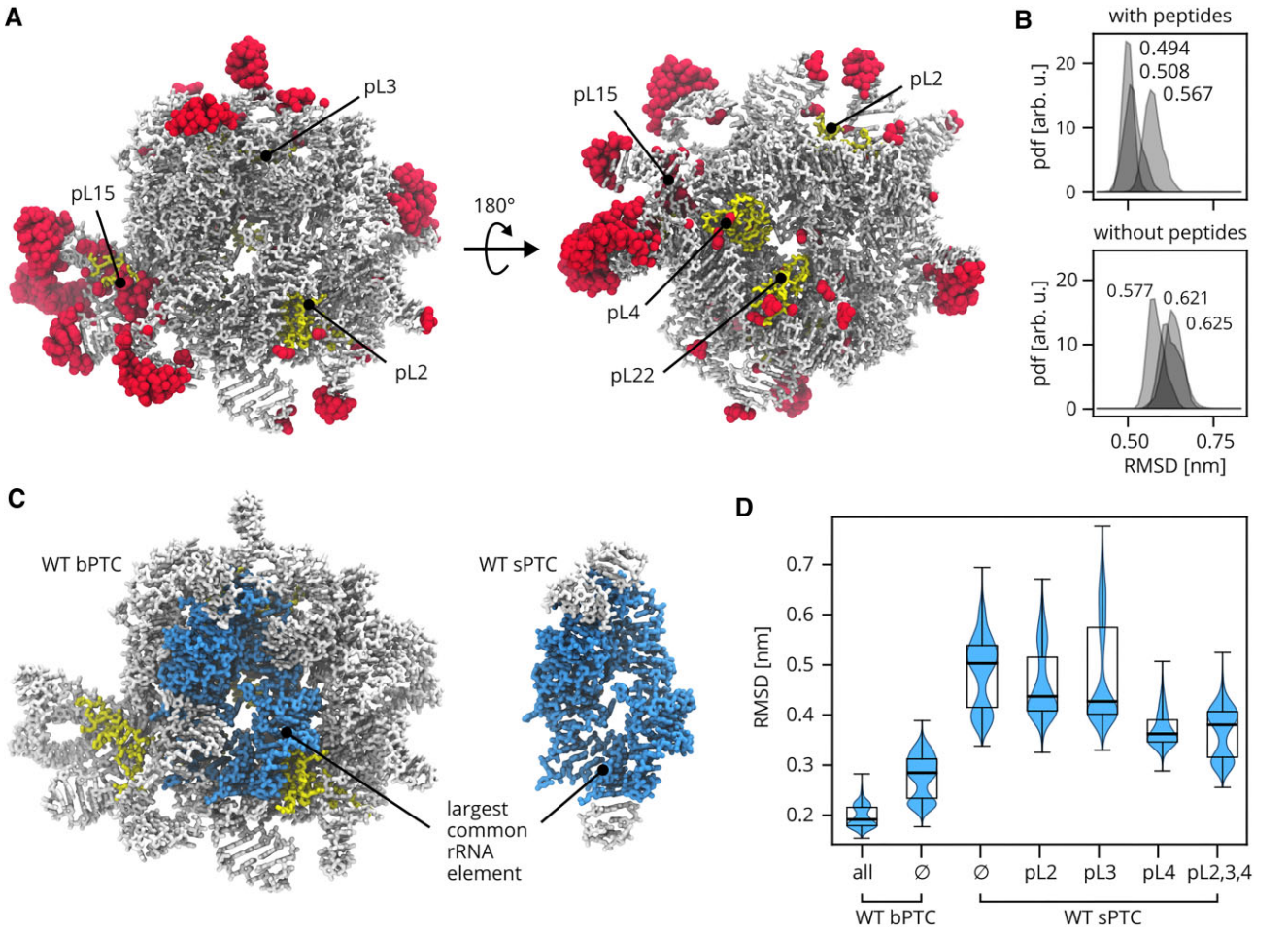


Figure 3. Overview of MD simulations. (**A**) The most flexible residues as identified by RMSF analysis of the last 800 ns of each trajectory. Top 10% residues with the highest RMSF are shown in surface representation, rRNA and rPeptides in stick representation. The two viewpoints show the tRNA-binding site and the ancestral exit tunnel lined by pL4 and pL22. For clarity, the location of pL3 is shown, although it is buried in the rRNA. (**B**) Probability density functions (pdfs) of RMSDs calculated for WT bPTC phosphorus atoms with respect to the initial conformation. The last 300 ns of each independent trajectory were used. Mean values of pdfs are denoted. (**C**) The largest common rRNA element (dark) in the context of WT bPTC and WT sPTC. (**D**) Violin plots of RMSD of the largest common rRNA element phosphorus atoms. The last 300 ns of each independent trajectory were used. Median of pdfs are denoted by the thick line, the box extends form the first to the third quartile, the whiskers extend from the box to the farthest data point lying within 1.5× the inter-quartile range.

as shown in Supplementary Figure S13. This is intriguing, as the rRNA environment sequence composition is different for each rPeptide and generally not very rich in purine bases, which have been previously shown to favour LLPS (31). As an extreme case of the nucleotide composition, the effect was further explored using polyU mixed with the pL22 peptide. Substantial LLPS weakening was observed; the peptide concentration required for inducing LLPS is higher and the coacervate has a lower volume fraction compared to the WT sPTC rRNA (Supplementary Figure S14). On the other hand, replacing ribosomal peptides with other polycations also suppresses LLPS: spermidine is not capable of condensing the RNA on a wide concentration interval, while spermine only promotes precipitation or aggregation (Supplementary Figure S14).

The coacervation effectively increases the stability of the PTC rRNA

Inside the coacervate, PTC and peptides are condensed and separated from a dilute aqueous solution, suggesting that coacervation may help to stabilize or protect RNA from hydrolysis. To investigate the structural integrity of the coacervates, the PTC stability was tested against enzymatic degradation.

RNase A was added to the WT bPTC and sPTC with or without rPeptides in concentration ratios corresponding to the inner LLPS region in the phase diagrams in Figure 4. The rPeptides were either used in combination or the same R ratio (defined by equation 1 in 'Materials and Methods', coacervation) was obtained only using pL22 (previously used for testing LLPS propensity, see Supplementary Figure S8). The potential degradation of PTCs under these conditions was visualized by Urea-PAGE (Figure 5A and B, lanes 1–6; Supplementary Figure S15). While the bPTC construct was not significantly protected by the rPeptides against enzymatic degradation under any of the tested conditions, a robust sPTC stabilization and protection was found when either pL22 alone or rPeptide mixture was used in the condition.

To test the importance of the coacervation-prone molar ratios, the pL22 peptide was used also in a significantly smaller concentration (~4×), representing its molar contribution from the rPeptide combination (Figure 5A and B, lanes

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

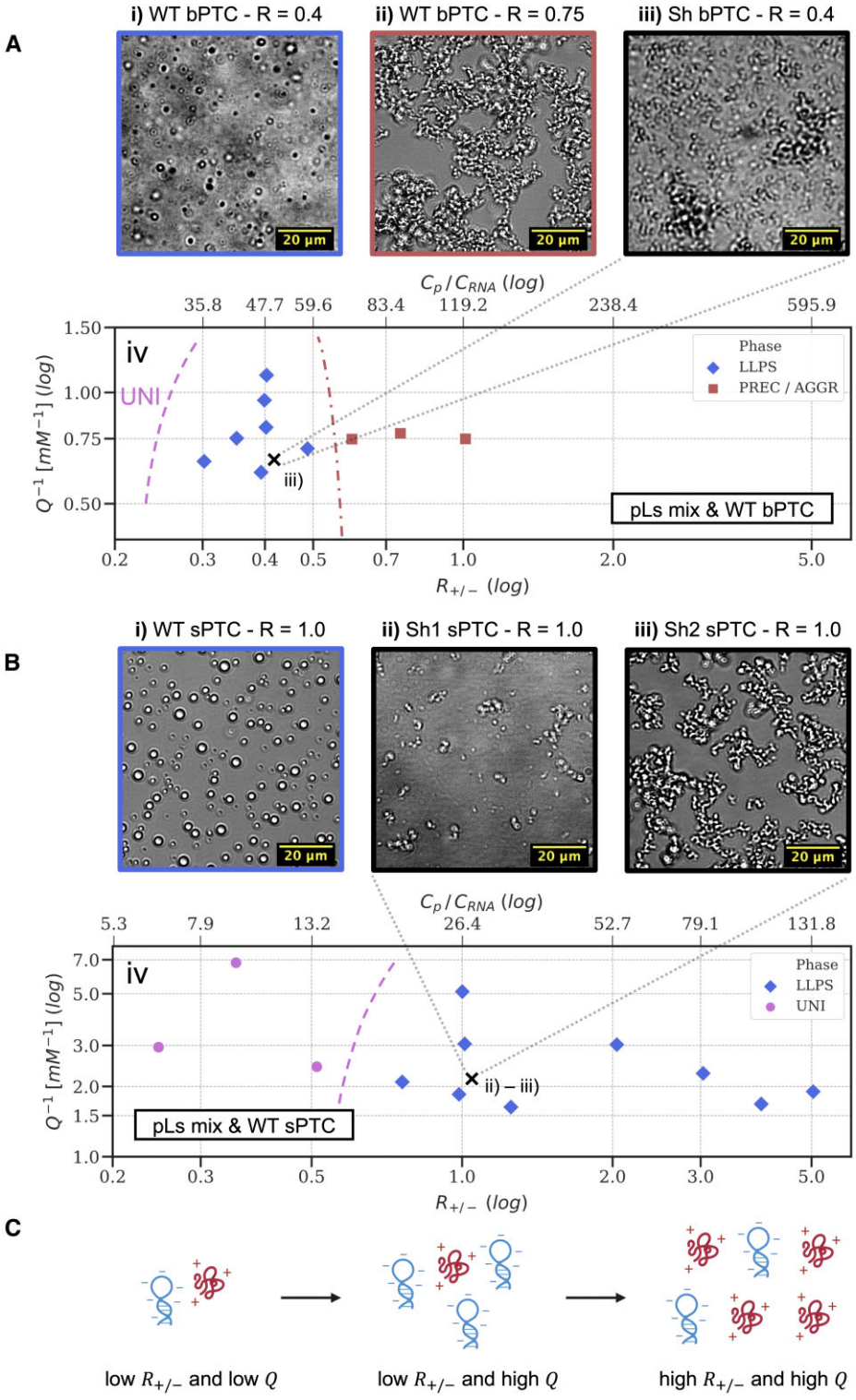
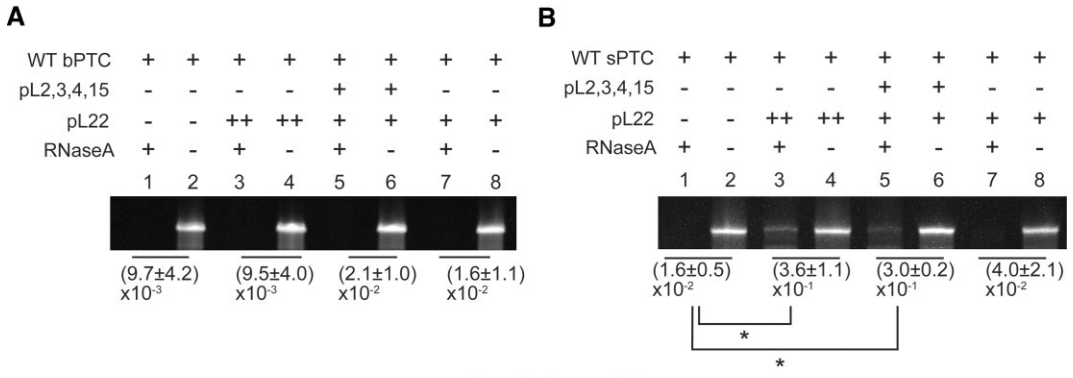


Figure 4. Phase separation of WT bPTC (sector A) and WT sPTC (sector B) with equimolar mixtures of pL2, 3, 4, 15 and 22, compared to the phase behaviour of the shuffled variants Sh bPTC, Sh1 sPTC and Sh2 sPTC (panel iii in sector A and panels ii and iii in sector B, respectively, see Supplementary Figure S2). Phase diagrams (panels iv in sectors A and B) are drawn as a function of the stoichiometric charge ratio $R_{+/-}$ (*x*-axis, see equation 1 in 'Materials and Methods' section, coacervation) and of the inverse of the molar concentration of charged units Q^{-1} (*y*-axis, see equation 2 in ibidem), with both axes in log scale. A second *x*-axis (denoted by C_p/C_{RNA}) on top of the diagrams shows the equivalent of $R_{+/-}$ in terms of stoichiometric ratio between the five rPeptides and WT bPTC or sPTC. Different symbols are associated with different phases, with 'UNI' indicating a uniform solution, 'PREC/AGGR' the presence of precipitates or aggregates (optical micrograph ii in sector A) and 'LLPS' the formation of liquid-like droplets (panels i in sectors A and B). Dashed curved lines in the phase diagrams are arbitrary guidelines for the eye and mark the region where no demixing ('UNI') or precipitation/aggregation is observed. The optical micrographs iii in sector A and ii, iii in sector B show the different phase behaviour of Sh bPTC and Sh1 and Sh2 sPTC as compared to their corresponding WT. The 'X' crosses on the diagrams mark the spots related to the Sh bPTC, Sh1 and Sh2 sPTC measurements. Sector C shows a pictorial representation of the $R_{+/-}$ and Q^{-1} parameters for a mixture of a single peptide (in red, carrying + 3 charges) and an RNA molecule (in blue, carrying -6 charges).

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024



C

R ratio for each lane

lane No.	WT bPTC	WT sPTC
1-2	R=0.0	R=0.0
3-4	R=0.40	R=1.0
5-6	R=0.40	R=1.0
7-8	R=0.11	R=0.27

Figure 5. rRNA stability to RNase A treatment in the absence and presence of rPeptides. (**A**) WT bPTC and (**B**) WT sPTC. The average ratios and standard deviations of the enzymatically treated versus untreated RNAs are listed below the respective bands, along with their significance values where appropriate. (* = *P*-value <0.2, $n = 3$ and $u = 0$ Wilcoxon–Mann–Whitney test). (**C**) Table of *R* ratios in each lane.

7–8). The sPTC was fully degraded under such conditions, supporting our hypothesis that coacervation, and not just interaction, is needed for the protection against enzymatic degradation.

Discussion

Following the past decades of the RNA-first (or other 'Me-First') narratives in the origins-of-life, discussions have gradually shifted towards a greater emphasis on the systems chemistry level of thinking (13,32,33). Several recent studies have succeeded in syntheses of molecules and proto-biological structures that could be nominated for possible routes towards early life (34–39). Nevertheless, to truly understand the origin of *our* life (versus possible routes to life), gaps must be filled between the prebiotic chemistry, protobiology and ancestral life forms. The PTC (or protoribosome) clearly represents one of the most profound molecular fossils of ancestral life. It is one of the farthest points that we can reach by tracing back current molecular life. Understanding how it came about and its route towards early biology therefore provides a unique opportunity to fill these gaps.

Two groups have recently reported the PTC activity using RNA structures reconstructed from the core of extant ribosomes (10–12). The RNA itself was therefore shown to be capable of peptide bond formation between amino acids, albeit with very low efficiency. Nevertheless, due to the prebiotic abundance of amino acids and the facile nature of their condensation (e.g. by wet–dry cycling), short peptide sequences were most probably prebiotically plausible also in the absence of ribosomal synthesis (13,40). The role of peptides in the emergence and evolution of the protoribosome is the subject of this study.

We reproduced two previous models of the PTC rRNA constructs: (i) the sPTC resembling a more compact and ancestral form that has been shown to possess the peptidyl-transferase activity, and (ii) bPTC representing a ∼4.5× larger construct that—within the natural ribosome—interacts with the fragments of several ribosomal proteins. Using MD simulations, we observe that the sPTC displays much broader conformational freedom, while the bPTC is more rigid and structurally similar to the modern ribosome. The bPTC construct includes ∼8% of inserted loops that are not part of the natural rRNA but were designed to join this truncated form in one polymer and stabilize its structure. Our observation of the bPTC structural properties broadly agrees with its previous characterization (9). The insertion does not significantly alter the overall nucleobase composition.

To examine the role of peptides in the different proto-ribosomal scenarios, we further included five rPeptides (fragments of proteins that interact with the PTC in native ribosome) in our study. These peptides contain all canonical amino acids and are enriched in positively charged residues. Diverse lines of evidence suggest that prebiotically available amino acids did not include about half of the extant alphabet, missing e.g. today's positively charged and larger functional groups, such as aromatics (41–43). The five rPeptides therefore do not represent the most prebiotically plausible peptide variants; they can rather be viewed as highly conserved peptide fossils that have been preserved in the ribosomal core. All five rPeptides are devoid of regular secondary structure arrangement.

We observe that most of the rPeptides interact with both rRNA PTC constructs with affinities in μM range. In case of bPTC, the binding curves change and dissociation constants decrease 2–3 times upon addition of metal ions which are (i) found in extant ribosomal structures, (ii) considered prebiotically plausible and (iii) have been previously shown to stabilize the PTC (44). In contrast, the rPeptide-sPTC binding curves are almost unaffected by the metal ion addition. The presence of metal ions may play an important role in minimizing non-specific binding which would be based on the significantly opposite charged components of the interaction partners only. The inclusion of the rPeptides in the MD simulation of the bPTC construct further reduced the RMSD and suggested that the rPeptides could have played a role in stabilization of the rRNA structure in the protoribosome. MD simulations also suggested that the interactions of the rPeptides

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

with the sPTC are less specific than with the bPTC. Although we did not observe the dissociation of the rPeptides from the sPTC within the simulation time, pL2, pL3 and pL4 were all 2- to 3-fold more mobile on the sPTC than on bPTC, as characterized by the RMSD and conformational ensembles of the rPeptides (Supplementary Figure S16).

Similar (although in some cases longer) rPeptide sequences were used previously by Hsiao et al. to characterize the properties of the bPTC construct (9). They observed binding of pL4, pL3, pL15 and pL22, where only pL4 binding was tested in the peptide form and the remaining peptides were prepared as fusions with the maltose-binding protein. Intriguingly, in that study all the binding tests were performed in the absence of metal ions, i.e. conditions which implied less specific interaction properties in our study (9). In our experimental setup, the pL4 peptide (shorter by 15 amino acids than in the previous study) is the only one that does not display any interaction upon metal ion addition.

Finally, the addition of rPeptides to the PTC rRNA constructs was observed to induce LLPS of both construct variants. However, the bPTC liquid demixing is limited to a narrow area of the phase diagram, while the sPTC construct displays LLPS under a broader range of peptide/rRNA concentrations and is qualitatively more prominent. Using the shuffled or mutated nucleic acid sequences, this effect was found to depend on both their sequence specificity and secondary/tertiary structure. We hypothesized that the PTC coacervation may have a stabilizing effect on the rRNA moiety and confirmed that the sPTC rRNA construct was protected against enzymatic degradation. This was true when the experiments followed the LLPS-prone concentration ratios, and the effect was therefore LLPS-based rather than interaction specific. While enzymatic degradation would be an unlikely source of prebiotic RNA instability, it was selected here as a general model of the coacervate diffusion vulnerability. In agreement with the lower prominence of the bPTC coacervation propensity, the bPTC is degradation prone under the same conditions, despite its higher structural integrity. In the absence of harsh conditions, the size and structure of bPTC (further supported by specific interaction with rPeptides) would probably have gained sufficient stability towards its molecular sovereignty. On the other hand, stabilizing and compartmentalizing conditions would likely be essential for the earliest versions of the protoribosome. The protoribosome (sPTC) was previously shown to possess the peptidyltransferase activity (11). Here, we show that the arguably most ancient fragments of the ribosomal proteins trigger the protoribosome coacervation under a wide range of concentrations, providing compartmentalization and stabilization of the sPTC construct against degradation. Our results demonstrate a clear example of a 'peptide-RNA' co-evolution using the very fossil of our biological life as an example. Several studies have recently evoked Oparin's century-old theory that coacervate microdroplets could have played major roles during life's origins (16,28,45). Because of the high RNA–peptide LLPS propensity, coacervates have been proposed as a unique environment supporting co-evolution of those species (16). Current studies are now researching the impact of different environments as well as the polymer properties and lengths on the coacervation propensities (45,46). The bigger construct used in our study, bPTC, is structurally more rigid and interacts with the rPeptides with higher specificity. Compared with sPTC, it has only weak LLPS propensity and the LLPS conditions do not provide additional stability to bPTC.

Based on the data presented here, we speculate that during origins of life, coacervation could be more prominent with flexible structures and based on interactions that could have lacked specificity. Such compartmentalization could lead to concentration of reactants and enhance the activity as previously demonstrated using different ribozyme examples (47,48). It could also lead to evolution of ribosomal peptides towards sequences that could interact with the gradually accrued rRNA structures more specifically and to help in such structure stabilization and sovereignty. Nevertheless, to provide the true link between prebiotic chemistry and the biological past, it will be of key importance to test these hypotheses using prebiotically plausible pre-ribosomal peptides as well as compare the potential impact of pre-/ribosomal peptides on the PTC activity.

Data availability

The data underlying this article are available in the article and in its online Supplementary Data. The data necessary for reproducing our MD simulations are available at https://github.com/mhkoscience/codispoti-protoribosome and https://doi.org/10.5281/zenodo.13729610.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Ing. Dalibor Pánek, Dr Kateřina Nováková and Dr Tilo Schorn for their technical support and Dr Vyacheslav Tretyachenko for helpful discussions about this work. We acknowledge CF Biophysics of CIISB, Instruct-CZ Centre, supported by MEYS CR (LM2023042) and European Regional Development Fund-Project 'UP CIISB' (No. CZ.02.1.01/0.0/0.0/18_046/0 015 974). We further acknowledge the Imaging Methods Core Facility at BIOCEV campus, supported by the MEYS CR (LM2023050 Czech-BioImaging, https://imcf.natur.cuni.cz/IMCF/). The use of confocal microscope was supported by the Department of Excellence Project 'SCALE UP' funded by the Italian Ministry of University and Research (MUR), Department of Molecular Biotechnology and Translational Medicine.

Funding

Charles University; TOYOBO Biotechnology Foundation (to T.Y.); Human Frontier Science Program [RGEC27/2023]; Univerzita Karlova v Praze [4EU+/22/F4/25, 4EU+/23/F4/17, SEED4EU+/2023/F4/316]; NINS [AB301003, AB311001]; Ministerstvo Školství, Mládeže a Tělovýchovy [LM2023042, LM2023050, e-INFRA CZ (ID:90254)]; European Regional Development Fund [CZ.02.1.01/0.0/0.0/18_046/0015974]; JSPS [21J12128 to T.Y.]. Funding for open access charge: HFSP.

Conflict of interest statement

None declared.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

References

1. Bowman,J.C., Petrov,A.S., Frenkel-Pinter,M., Penev,P.I. and Williams,L.D. (2020) Root of the tree: the significance, evolution, and origins of the ribosome. *Chem. Rev.*, **120**, 4848–4878.
2. Petrov,A.S., Gulen,B., Norris,A.M., Kovacs,N.A., Bernier,C.R., Lanier,K.A., Fox,G.E., Harvey,S.C., Wartell,R.M.,, Hud,N.V., *et al.* (2015) History of the ribosome and the origin of translation. *Proc. Natl Acad. Sci. USA*, **112**, 15396–15401.
3. Davidovich,C., Belousoff,M., Bashan,A. and Yonath,A. (2009) The evolving ribosome: from non-coded peptide bond formation to sophisticated translation machinery. *Res. Microbiol.*, **160**, 487–492.
4. Lupas,A.N. and Alva,V. (2017) Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol.*, **198**, 74–81.
5. Hsiao,C., Mohan,S., Kalahar,B.K. and Williams,L.D. (2009) Peeling the onion: ribosomes are ancient molecular fossils. *Mol. Biol. Evol.*, **26**, 2415–2425.
6. Hury,J., Nagaswamy,U., Larios-Sanz,M. and Fox,G.E. (2006) Ribosome origins: the relative age of 23S rRNA Domains. *Orig. Life Evol. Biosph.*, **36**, 421–429.
7. Bokov,K. and Steinberg,S.V. (2009) A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, **457**, 977–980.
8. Noller,H.F. (2012) Evolution of protein synthesis from an RNA world. *Cold Spring Harb. Perspect. Biol.*, **4**, a003681.
9. Hsiao,C., Lenz,T.K., Peters,J.K., Fang,P.-Y., Schneider,D.M., Anderson,E.J., Preeprem,T., Bowman,J.C., O'Neill,E.B., Lie,L., *et al.* (2013) Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucleic Acids Res.*, **41**, 3373–3385.
10. Bose,T., Fridkin,G., Bashan,A. and Yonath,A. (2021) Origin of life: chiral short RNA chains capable of non-enzymatic peptide bond formation. *Isr. J. Chem.*, **61**, 863–872.
11. Bose,T., Fridkin,G., Davidovich,C., Krupkin,M., Dinger,N., Falkovich,A.H., Peleg,Y., Agmon,I., Bashan,A. and Yonath,A. (2022) Origin of life: protoribosome forms peptide bonds and links RNA and protein dominated worlds. *Nucleic Acids Res.*, **50**, 1815–1828.
12. Kawabata,M., Kawashima,K., Mutsuro-Aoki,H., Ando,T., Umehara,T. and Tamura,K. (2022) Peptide bond formation between aminoacyl-minihelices by a scaffold derived from the peptidyl transferase center. *Life (Basel)*, **12**, 573 .
13. Fried,S.D., Fujishima,K., Makarov,M., Cherepashuk,I. and 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.
14. Tagami,S., Attwater,J. and Holliger,P. (2017) Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat. Chem.*, **9**, 325–332.
15. Oparin,A.I. (1959) The Origin of Life on the Earth. (Pergamon, Oxford)
16. Ghosh,B., Bose,R. and Tang,T.-Y.D. (2021) Can coacervation unify disparate hypotheses in the origin of cellular life? *Curr. Opin. Colloid Interface Sci.*, **52**, 101415.
17. Zearfoss,N.R. and Ryder,S.P. (2012) End-labeling oligonucleotides with chemical tags after synthesis. *Methods Mol. Biol.*, **941**, 181–193.
18. Selmer,M., Dunham,C.M., Murphy,F.V., Weixlbaumer,A., Petry,S., Kelley,A.C., Weir,J.R. and Ramakrishnan,V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935–1942.
19. Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
20. Zgarbová,M., Otyepka,M., Šponer,J., Lankaš,F. and Jurečka,P. (2014) Base pair fraying in molecular dynamics simulations of DNA and RNA. *J. Chem. Theory Comput.*, **10**, 3177–3189.
21. Maier,J.A., Martinez,C., Kasavajhala,K., Wickstrom,L., Hauser,K.E. and Simmerling,C. (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, **11**, 3696–3713.
22. Joung,I.S. and Cheatham,T.E. (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, **112**, 9020–9041.
23. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
24. Bussi,G., Donadio,D. and Parrinello,M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.*, **126**, 014101.
25. Parrinello,M. (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.*, **52**, 7182.
26. Abraham,M.J., Murtola,T., Schulz,R., Páll,S., Smith,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25.
27. Kovacs,N.A., Petrov,A.S., Lanier,K.A. and Williams,L.D. (2017) Frozen in time: the history of proteins. *Mol. Biol. Evol.*, **34**, 1252–1260.
28. Poudyal,R.R., Pir Cakmak,F., Keating,C.D. and Bevilacqua,P.C. (2018) Physical principles and extant biology reveal roles for RNA-containing membraneless compartments in origins of life chemistry. *Biochemistry*, **57**, 2509–2519.
29. Shakya,A. and King,J.T. (2018) DNA local-flexibility-dependent assembly of phase-separated liquid droplets. *Biophys. J.*, **115**, 1840–1847.
30. Fraccia,T.P. and Zanchetta,G. (2021) Liquid–liquid crystalline phase separation in biomolecular solutions. *Curr. Opin. Colloid Interface Sci*, **56**, 101500.
31. Lu,T., Nakashima,K.K. and Spruijt,E. (2021) Temperature-responsive peptide-nucleotide coacervates. *J. Phys. Chem. B*, **125**, 3080–3091.
32. Preiner,M., Asche,S., Becker,S., Betts,H.C., Boniface,A., Camprubi,E., Chandru,K., Erastova,V., Garg,S.G., Khawaja,N., *et al.* (2020) The future of origin of life research: bridging decades-old divisions. *Life (Basel)*, **10**, 20.
33. Krishnamurthy,R. (2020) Systems chemistry in the chemical origins of life: the 18th Camel Paradigm. *J Syst Chem*, **8**, 40.
34. Frenkel-Pinter,M., Bouza,M., Fernández,F.M., Leman,L.J., Williams,L.D., Hud,N.V. and Guzman-Martinez,A. (2022) Thioesters provide a plausible prebiotic path to proto-peptides. *Nat. Commun.*, **13**, 2569.
35. Makarov,M., Sanchez Rocha,A.C., Krystufek,R., Cherepashuk,I., Dzmitruk,V., Charnavets,T., Faustino,A.M., Lebl,M., Fujishima,K.,, Fried,S.D., *et al.* (2023) Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc.*, **145**, 5320–5329.
36. Chandru,K., Jia,T.Z., Mamajanov,I., Bapat,N. and Cleaves,H.J. (2020) Prebiotic oligomerization and self-assembly of structurally diverse xenobiological monomers. *Sci. Rep.*, **10**, 17560.
37. Müller,F., Escobar,L., Xu,F., Węgrzyn,E., Nainytė,M., Amatov,T., Chan,C.-Y., Pichler,A. and Carell,T. (2022) A prebiotically plausible scenario of an RNA-peptide world. *Nature*, **605**, 279–284.
38. Toparlak,Ö.D., Karki,M., Egas Ortuno,V., Krishnamurthy,R. and Mansy,S.S. (2020) Cyclophospholipids increase protocellular stability to metal ions. *Small*, **16**, e1903381.
39. Rout,S.K., Friedmann,M.P., Riek,R. and Greenwald,J. (2018) A prebiotic template-directed peptide synthesis based on amyloids. *Nat. Commun.*, **9**, 234.
40. Forsythe,J.G., Yu,S.-S., Mamajanov,I., Grover,M.A., Krishnamurthy,R., Fernández,F.M. and Hud,N.V. (2015) Ester-mediated amide bond formation driven by wet-dry cycles: a possible path to polypeptides on the prebiotic earth. *Angew. Chem. Int. Ed*, **54**, 9871–9875.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

41. Higgs,P.G. and Pudritz,R.E. (2009) A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*, **9**, 483–490.
42. Cleaves,H.J. (2010) The origin of the biologically coded amino acids. *J. Theor. Biol.*, **263**, 490–498.
43. Zaia,D.A.M., Zaia,C.T.B.V. and De Santana,H. (2008) Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.*, **38**, 469–488.
44. Rivas,M. and Fox,G.E. (2020) Further characterization of the pseudo-symmetrical ribosomal region. *Life (Basel)*, **10**, 201.
45. Spruijt,E. (2023) Open questions on liquid-liquid phase separation. *Commun. Chem.*, **6**, 23.
46. Cannelli,S.M.C., Gupta,R., Nguyen,T., Poddar,A., Sharma,S., Vithole,P.V. and Jia,T.Z. (2023) A compositional view comparing modern biological condensates and primitive phase-separated compartments. *Pept. Sci.*, **115**, e24331.
47. Poudyal,R.R., Guth-Metzler,R.M., Veenis,A.J., Frankel,E.A., Keating,C.D. and Bevilacqua,P.C. (2019) Template-directed RNA polymerization and enhanced ribozyme catalysis inside membraneless compartments formed by coacervates. *Nat. Commun.*, **10**, 490.
48. Le Vay,K., Song,E.Y., Ghosh,B., Tang,T.-Y.D. and Mutschler,H. (2021) Enhanced ribozyme-catalyzed recombination and oligonucleotide assembly in peptide-RNA condensates. *Angew. Chem. Int. Ed*, **60**, 26096–26104.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae823/7786163 by guest on 22 October 2024

Received: February 1, 2024. **Revised:** September 4, 2024. **Editorial Decision:** September 5, 2024. **Accepted:** September 10, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

SUPPLEMENTARY INFORMATION

The interplay between peptides and RNA is critical for protoribosome compartmentalization and stability

Simone Codispoti[1†], Tomoko Yamaguchi[2,3†$], Mikhail Makarov[3], Valerio G. Giacobelli[3], Martin Mašek[4], Michal H. Kolář[4], Alma Carolina Sanchez Rocha[3], Kosuke Fujishima[2,5,6*], Giuliano Zanchetta[1*], Klára Hlouchová[3,7*]

[1]*Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università di Milano, Segrate, 20054, Italy*

[2]*Earth-Life Science Institute, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8550, Japan*

[3]*Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague, 12843, Czech Republic*

[4]*Department of Physical Chemistry, University of Chemistry and Technology, Technicka 5, 16628 Prague, Czech Republic*

[5]*School of Life Science and Technology, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan*

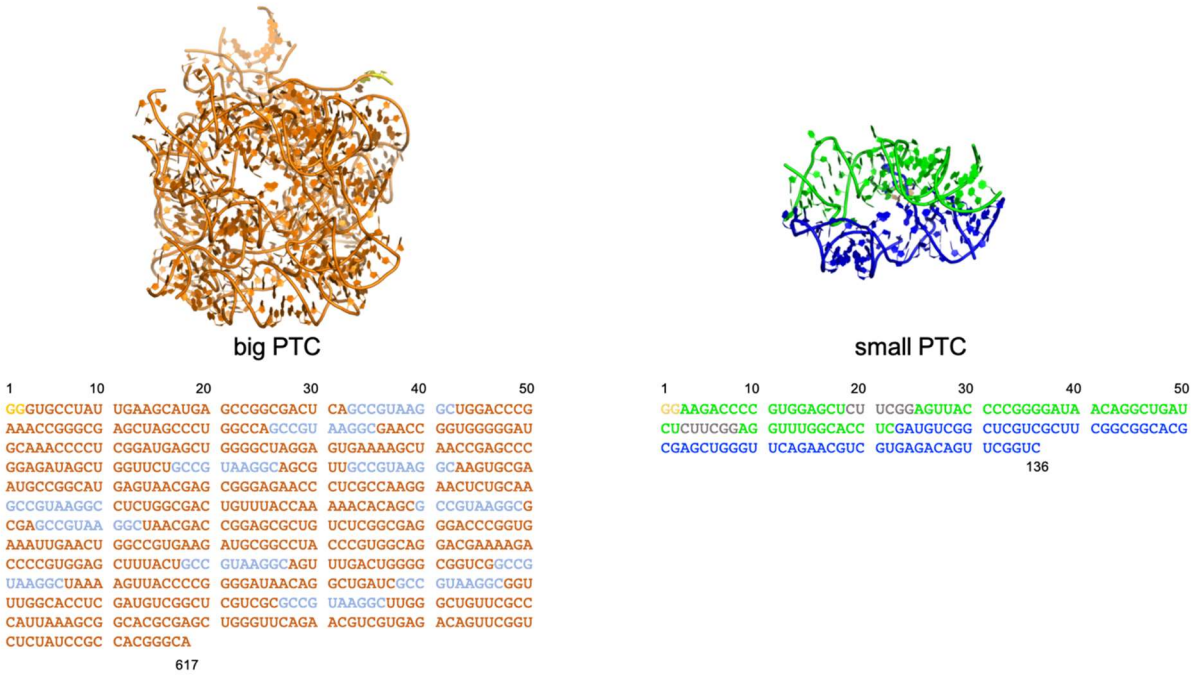
[6]*Graduate School of Media and Governance, Keio University, Fujisawa, 252-0882, Japan*

[7]*Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, 16610, Czech Republic*

[†]These authors contributed equally to this work

[$]Current address: Department of Structural Biology, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan

[*]To whom correspondence should be addressed: fuji@elsi.jp, giuliano.zanchetta@unimi.it, klara.hlouchova@natur.cuni.cz

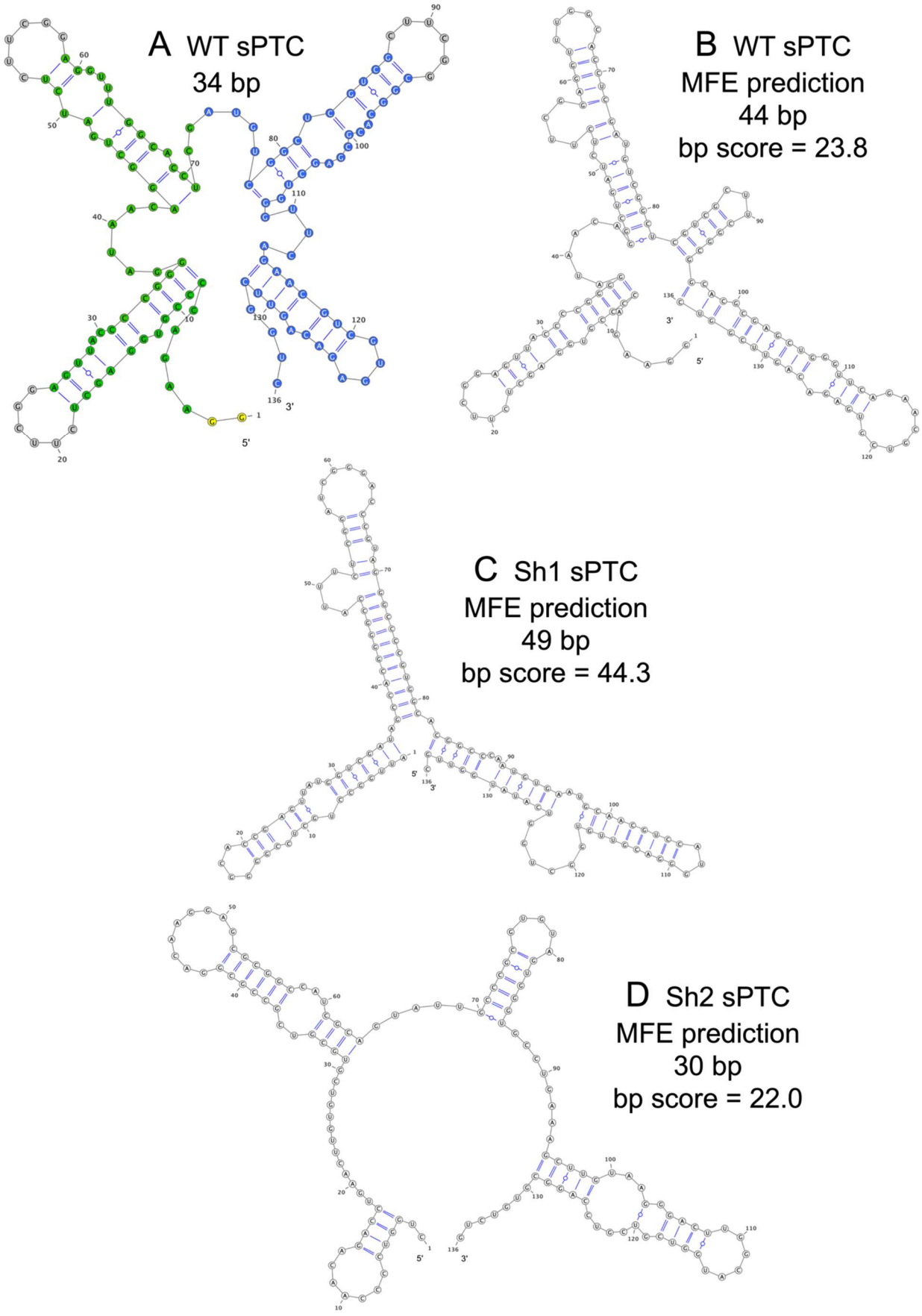


Supplementary Fig. 1. Tertiary structures derived from PDB 4V51 and related sequences of the two WT PTC constructs employed in this study: the 617 nt bPTC (*left panel*) and the 136 nt sPTC (*right panel*). WT bPTC is made up of the 505 nt from the 23S rRNA (*orange*), joined by 11 stem loops (*light blue*). The WT sPTC is composed of 116 nt from the 23S rRNA (A region in *green* & P region in *blue*) and connected with 3 stem loops (*grey*). The extra GG for transcription is marked in *yellow*.

Supplementary Table 1. DNA oligos used for preparation of PTC RNA constructs.

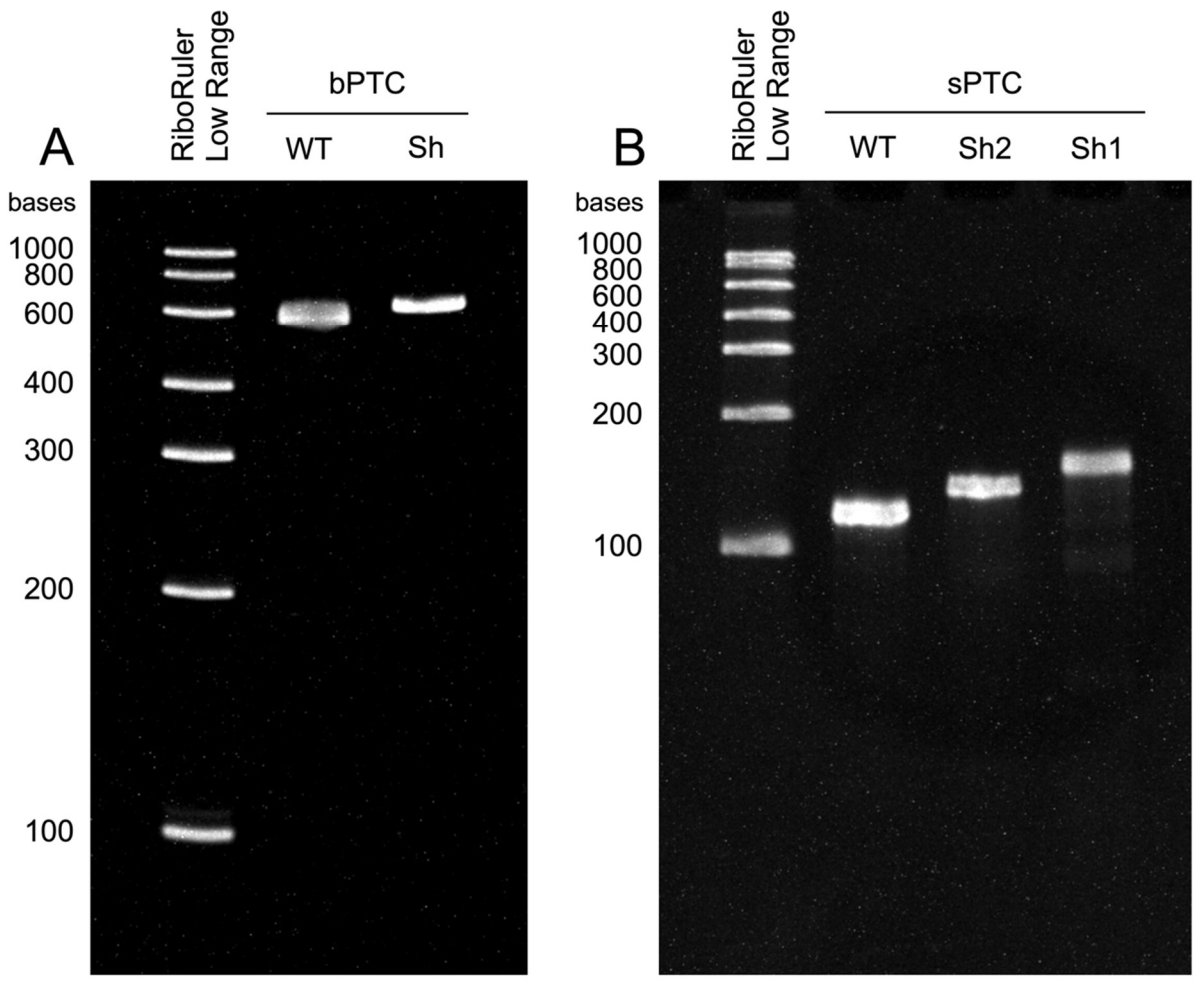
Name	Sequence	Note	Reference
WT_bPTC	AAAAAGTGGCCTGTAATACGACTCACTATAGGGTGCC TATTGAAGCATGAGCCGGCGACTCAGCCGTAAGGCTG GACCCGAAACCGGGCGAGCTAGCCCTGGCCAGCCGTA AGGCGAACCGGTGGGGGATGCAAACCCCTCGGATGAG CTGGGGCTAGGAGTGAAAAGCTAACCGAGCCCGGAGA TAGCTGGTTCTGCCGTAAGGCAGCGTTGCCGTAAGGC AAGTGCGAATGCCGGCATGAGTAACGAGCGGGAGAAC CCTCGCCAAGGAACTCTGCAAGCCGTAAGGCCTCTGG CGACTGTTTACCAAAAACACAGCGCCGTAAGGCGCGA GCCGTAAGGCTAACGACCGGAGCGCTGTCTCGGCGAG GGACCCGGTGAAATTGAACTGGCCGTGAAGATGCGGC CTACCCGTGGCAGGACGAAAAGACCCCGTGGAGCTTT ACTGCCGTAAGGCAGTTTGACTGGGGCGGTCGGCCGT AAGGCTAAAAGTTACCCCGGGGATAACAGGCTGATCG CCGTAAGGCGGTTTGGCACCTCGATGTCGGCTCGTCG CGCCGTAAGGCTTGGGCTGTTCGCCCATTAAAGCGGC ACGCGAGCTGGGTTCAGAACGTCGTGAGACAGTTCGG TCTCTATCCGCCACGGGCA	Template of WT bPTC	(1)
Sh_bPTC	AAAAAGTGGCCTGTAATACGACTCACTATAGGCCTCG GCGGACCCACCCCAACCTCGAGTAGGGGACCTACACA GAGCATTATCAAACGACAGCCTTCTCAGGTGCATTAA GGACACCAAAAGATCCAGCCGCCGGATGGGATGAGTG ACCGACCGCTACTAGGCCGCATGAGACCCCAATATCC CGTGCAGACAGGGAAAGGCACAGACGAGGTCAGTCAA ACTGGTCCGCTTCACAGGGGGAGTCACTGCAAGGTCG TATACTGCGTGGTTGGGTACCGGTCGGGTCTGCTGAC CAGTATGAGGTATTCCCTACTCTAGGGATGACCGAAG GGTGTTAGTTCCCGAGGAGATTCCCGCCCAGGAATGC TAGATGGACGCGGGAGCGTGTAGTCAGGGGTAAGACT TAGACACGGGTCCACGTAGCGGTGGCACTGATAGGTG CGTTATGCGCCGCGACGCTGGGCTGGGCCAGGTGACG TCGAGCAGTGCTGACGGCGAAGGGCGGCGCCGCGGGC GAGGATCTAATCAGGCTTTACTGTGCAGTAGGACGGG GATGGCCCGGGCACGCTATACGGCGACGCCCGCCCAC GGTAAGCGGGAGAGATGCGACCATCGAGCAGCCGAGT GGTCTATCCGCCACGGGCA	Template of Sh bPTC	This study

bPTC_F	AAAAAGTGGCCTGTAATACGACTCA	Forward primer for preparing dsDNA of bPTC by using DNA polymerase	This study
bPTC_R	TCTATCCGCCACGGGCA	Reverse primer for preparing dsDNA of bPTC by using DNA polymerase	This study
WT_sPTC_F	GCGTAATACGACTCACTATAGGAAGACCCCGTGGAGCTCTTCGGAGTTACCCCGGGGATAACAGGCTGATCTCTTCGGAGGTTTGGCAC	Forward primer for preparing dsDNA of WT sPTC by using Klenow enzyme	(2)
WT_sPTC_R	GACCGAACTGTCTCACGACGTTCTGAACCCAGCTCGCGTGCCGCCGAAGCGACGAGCCGACATCGAGGTGCCAAACCTCCGAAGAG	Reverse primer for preparing dsDNA of WT sPTC by using Klenow enzyme	(2)
Sh1_sPTC_R	GCAACCATATGACCAGCCACAACGTCCCATGGACGTTGCATTCACATTGGGCCGTGCCACGGGGCCCTACGGGTCCCGATCCGAGAAATGGCCCCGTGGCTATCGACCATAACTCGGTGCCCCGGAGCAGGCCAATCCTATAGTGAGTCGTATTACGC	Template of Sh1 sPTC	This study
Sh2_sPTC_R	CAGACACGCCTGGACGACGACCATGCCAAGTCCCTTACAAGCTTTCAGGCACCCACTACACGCGGGCAATACTGCGATGGCCGCGCTCCTTGTCCGCGGCGACGCACGACACAAGTTCAGGTCTGTTGGGGACCAGCCTATAGTGAGTCGTATTACGC	Template of Sh2 sPTC	This study
T7_F	GCGTAATACGACTCACTATAGG	Forward primer for sPTCs	This study

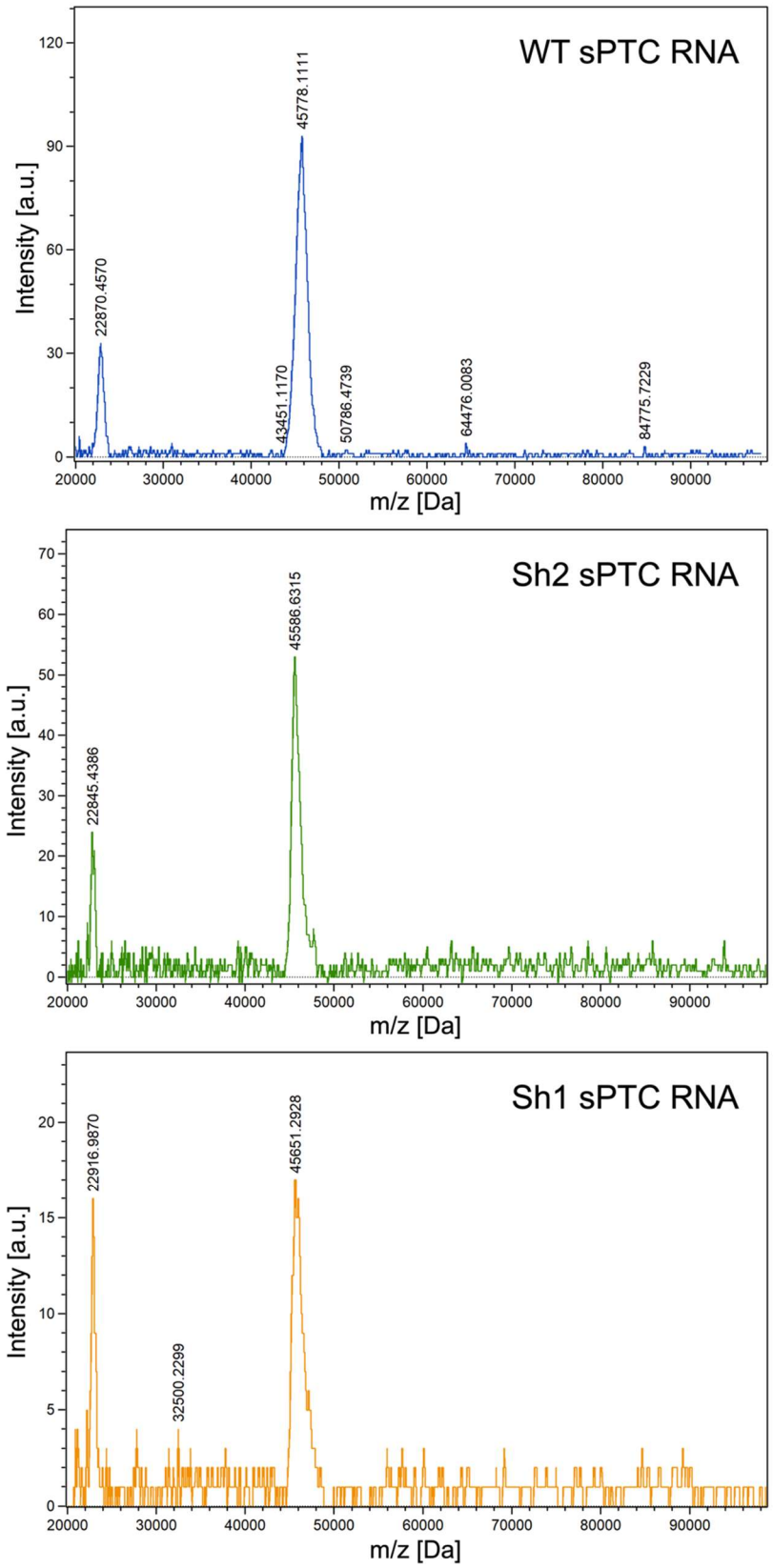

A WT sPTC
34 bp
B WT sPTC
MFE prediction
44 bp
bp score = 23.8
C Sh1 sPTC
MFE prediction
49 bp
bp score = 44.3
D Sh2 sPTC
MFE prediction
30 bp
bp score = 22.0

Supplementary Fig. 2. Secondary structures of WT sPTC and its two shuffled variants.
(A) Original WT sPTC based on PDB 4V51, characterised by 34 total number of base pairs (bp); blue-green colour scheme matches the two symmetrical regions depicted in **Supplementary Fig. 1**, while the additional 3 CUUCGG loops and the starter GG are also inserted and coloured in grey and yellow, respectively.
(B) NUPACK MFE prediction for the WT sPTC, characterised by 44 bp and a “bp score” (see below) of 23.8.
(C - D) NUPACK MFE prediction for the Sh1 and Sh2 sPTC, having respectively more and less structure content compared to the WT sPTC: Sh1 sPTC is characterised by 49 predicted bp and a “bp score” (see below) of 44.3, while Sh2 PTC has 30 predicted bp and a “bp score” of 22.
All predictions (B-D) were performed at the standard 1M NaCl and 37°C conditions adopted by the NUPACK algorithm. The number of base pairs (bp) in the Minimum Free Energy (MFE) is reported, together with a “bp score” computed as the sum of the probabilities that each nucleotide adopts the state of the MFE.

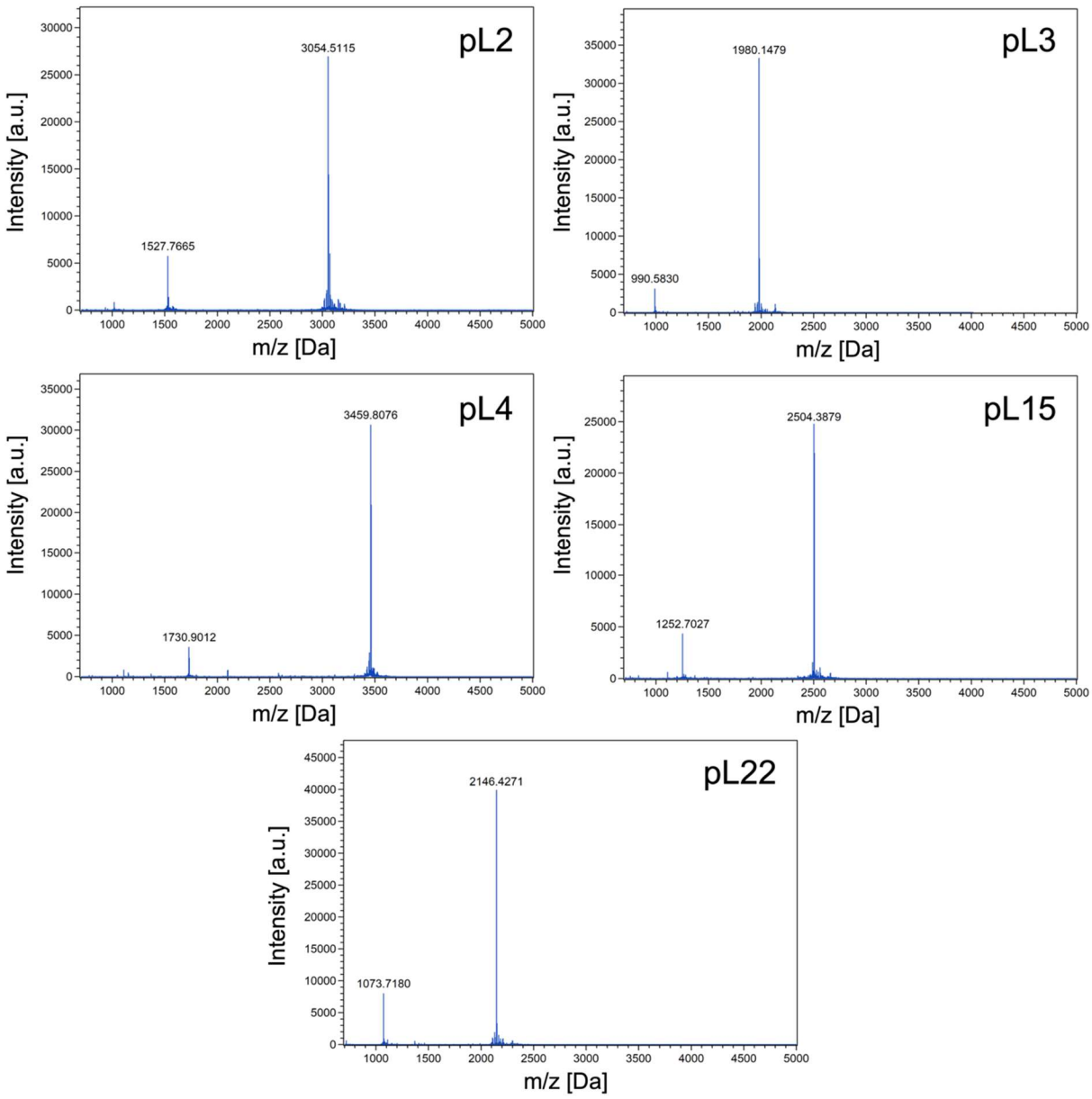
Even if the nominal conditions are different from the ones used in the experiments, it is guaranteed that the two variants assume different conformations, with less and more base pairs at equilibrium with respect to WT sPTC. This is due to the fact that: (i) the predicted structure of the original WT sPTC at 1 M NaCl and 37°C (B) lies in between the structures of Sh1 sPTC and Sh2 sPTC; (ii) monovalent salts and temperature have a monotonic effect on the stability of RNA.



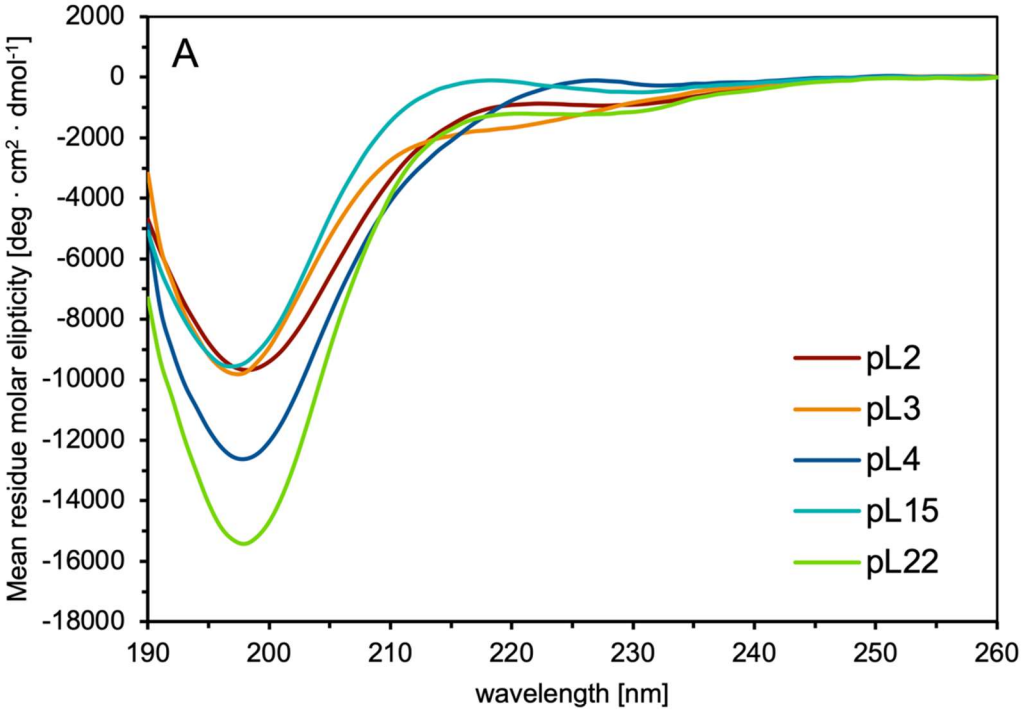
Supplementary Fig. 3. Denaturing gels of the RNA constructs.
(A) 5% TBE-urea gel of WT and Sh bPTC RNA constructs.
(B) 8% TBE-urea gel of WT and Sh sPTC RNA constructs.
Prior to electrophoresis, 200 ng RNA samples were mixed with RNA Gel Loading Dye (2X) (ThermoFisher Scientific) and denatured at 70°C for 10 min. RiboRuler Low Range RNA Ladder (ThermoFisher Scientific) was used as a reference. After electrophoresis, gels were soaked in TBE buffer supplemented with GelRed Nucleic Acid Gel Stain (Sigma Aldrich) for 10 min and visualised by UV transilluminator.



Supplementary Fig. 4. MALDI-TOF spectra of sPTC RNA constructs.



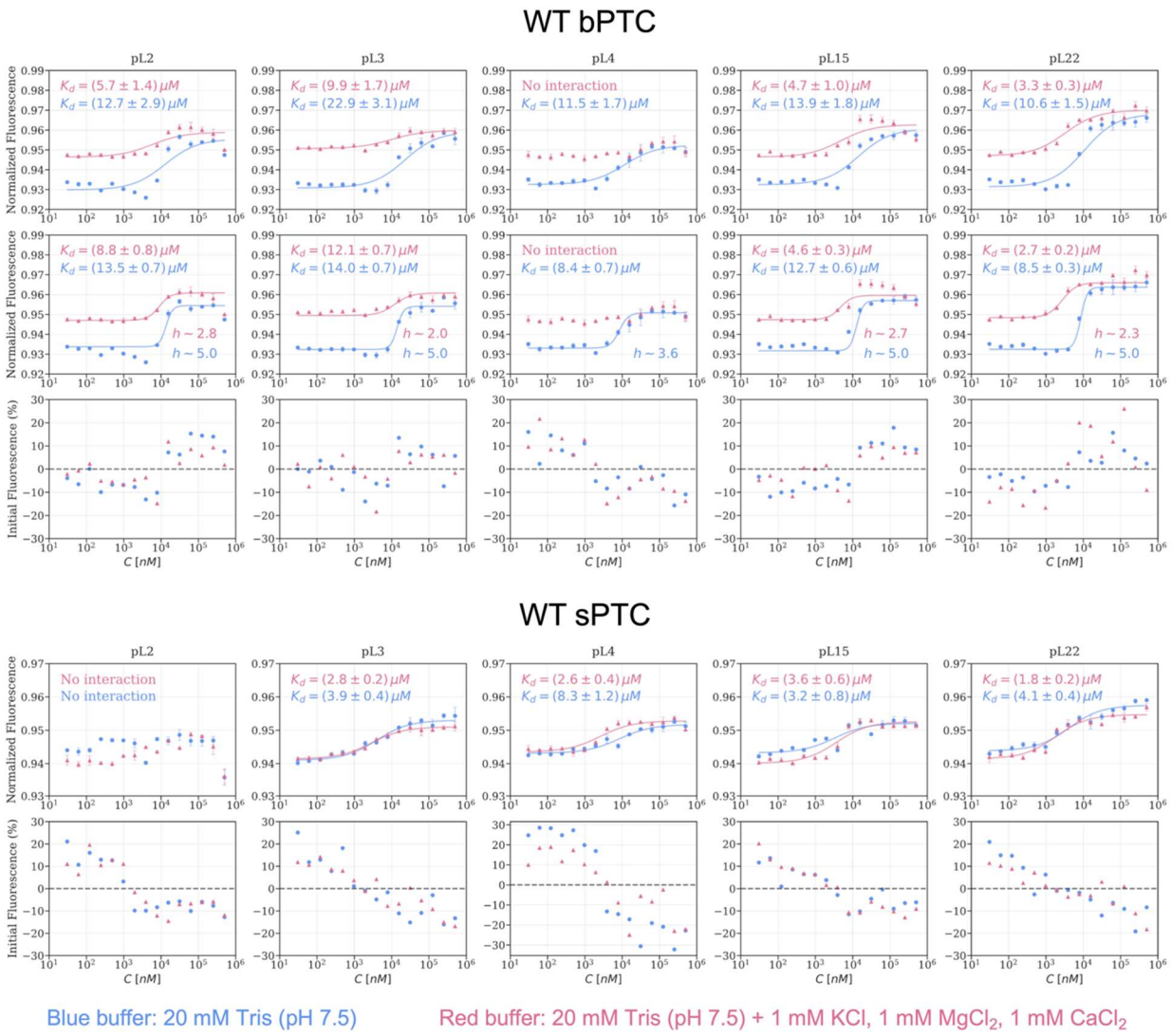
Supplementary Fig. 5. MALDI-TOF spectra of rPeptides (pL2, 3, 4, 15, 22).



B

rPeptide	rRNA bound state	AlphaFold2
pL2	Coil	Coil
pL4	Coil	Coil
pL3	Coil	Coil
pL15	Coil	Coil
pL22	Beta-hairpin	Alpha-helix region

Supplementary Fig. 6. Experimental & computational secondary structure assessment of rPeptides (pL2, 3, 4, 15, 22).
(A) Far-UV CD spectra of rPeptides; spectra were collected twice for every peptide in 10 mM Tris-HCl (pH 7.5), 1 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$; the average spectra are shown.
(B) Comparison between rPeptides structures as found in the rRNA bound state (PDB 4V51) and as predicted by AlphaFold2.



Supplementary Fig. 7. The binding curves (*upper rows*) and fluorescence traces (*lower rows*) for WT bPTC (*top panel*) and sPTC (*bottom panel*) with rPeptides pL2, 3, 4, 15, and 22 (*columns from left to right*), with (*red*) and without (*blue*) metal ions supplemented in the buffer. Solid curves in the first row of both panels represent the best fit to mass-action kinetics (*see eq. M below*). For WT bPTC, differences in the binding curves were observed in the presence vs absence of metal ions: the additional middle row in the WT bPTC panel shows the best fit to Hill functions (*see eq. H below*), where the approximate best Hill coefficients (denoted by *h*) were estimated.

Binding curves were analysed based on mass-action or Hill equations. sPTC curves are well fitted with a simple 1:1 binding reaction model:

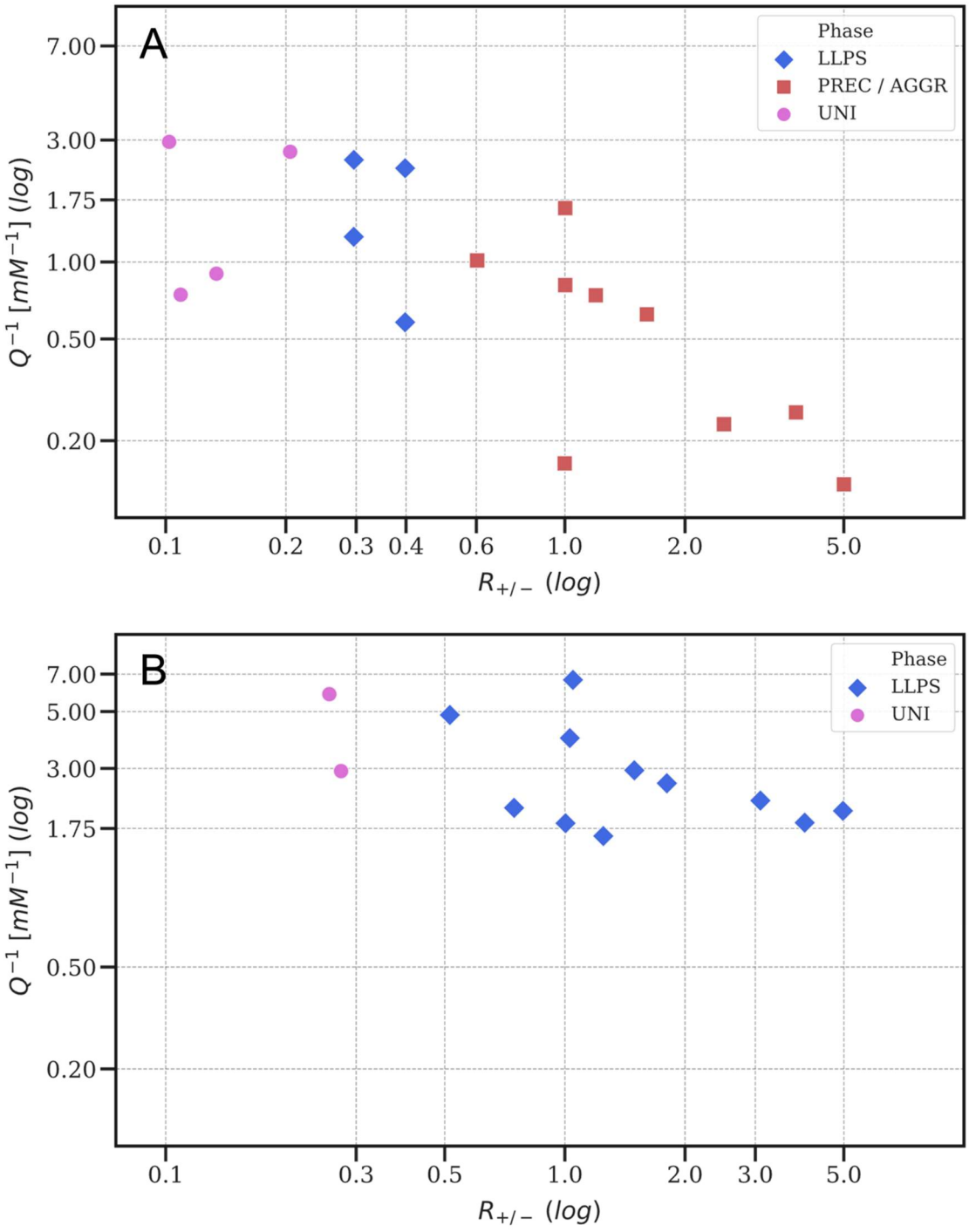
$$F_n([L];K_d) = F_n^{(u)} + \frac{F_n^{(b)} - F_n^{(u)}}{2}\frac{[L] + [T_0] + K_d - \sqrt{([L] + [T_0] + K_d)^2 - 4[L][T_0]}}{[T_0]} \qquad (eq.M)$$

where F_n is the normalized MST signal as a function of $[L]$, the ligand molecule concentration (pL2, 3, 4, 15 or 22, x axis in the figure above), $[T_0]$ is the fixed labelled molecule concentration (here the 50 nM WT sPTC or bPTC), K_d is the dissociation constant and $F_n^{(u)}$ and $F_n^{(b)}$ are, respectively, the normalized fluorescence values of the unbounded and bounded states; binding curves were fitted against $[L]$ with free parameters $F_n^{(u)}$, $F_n^{(b)}$ and K_d.

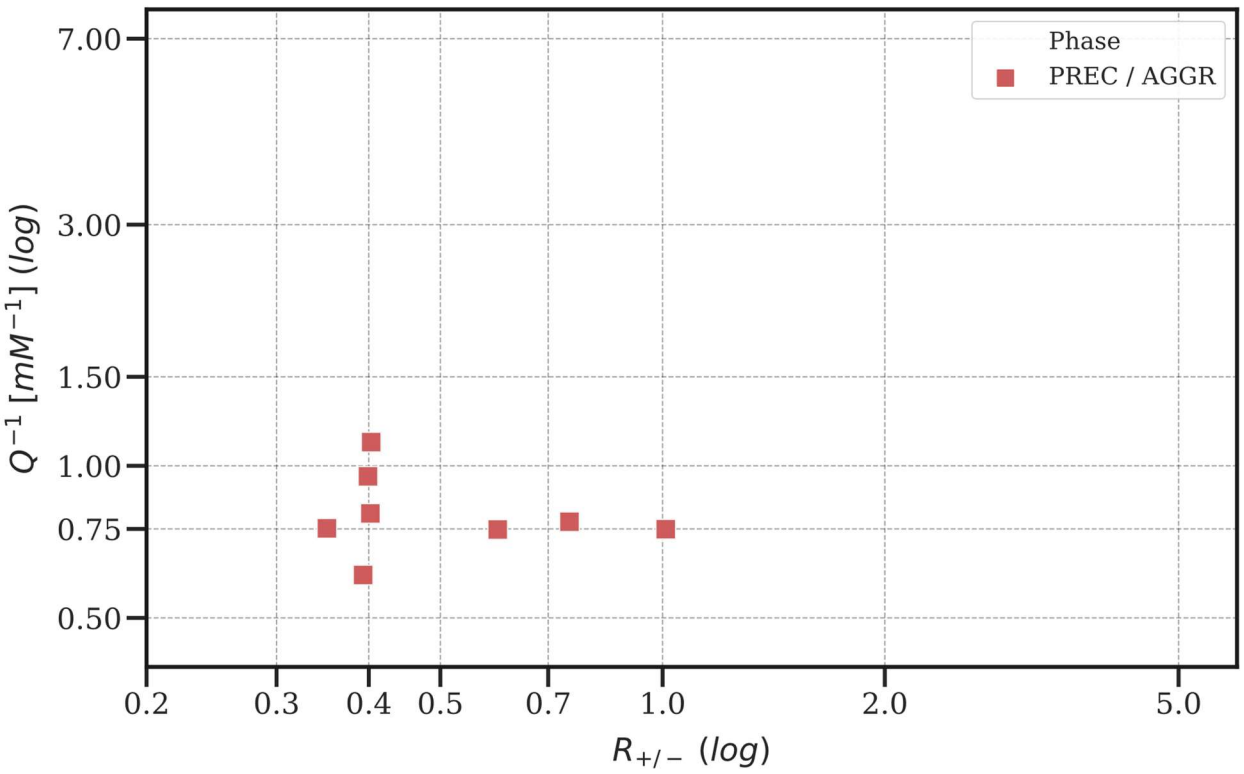
On the contrary, the shape of the bPTC curves suggests binding with peptide:RNA stoichiometry significantly different from 1:1, which can be accounted for in the most simple terms by the cooperative model represented by Hill equation:

$$F_n([L]; K_d, h) = F_n^{(u)} + \frac{F_n^{(b)} - F_n^{(u)}}{1 + (K_d/[L])^h} \quad (eq. H)$$

where $[L]$, $F_n^{(u)}$, $F_n^{(b)}$ and K_d have the same meaning as in $(eq. M)$, while h represents the Hill exponent; binding curves were fitted against $[L]$ with free parameters $F_n^{(u)}$, $F_n^{(b)}$, K_d and h.



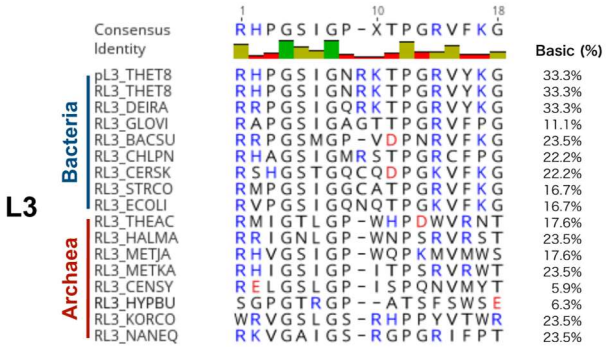
Supplementary Fig. 8. Phase diagrams of pL22 with WT bPTC (*panel A*) and WT sPTC (*panel B*) as a function of the stoichiometric charge ratio $R_{+/-}$ (x axis, see *eq. 1* in **Mat. & Meth., *coacervation***) and of the inverse of the molar concentration of charged units Q^{-1} (y axis, see *eq. 2* in ***ibidem***). Both axes are in log scale. Different colours and markers are associated to different phases, with "UNI" indicating mixed states, "PREC/AGGR" the presence of precipitates or aggregates (**Figure 4B**) and "LLPS" the formation of liquid-like droplets (**Figure 4A & F**).

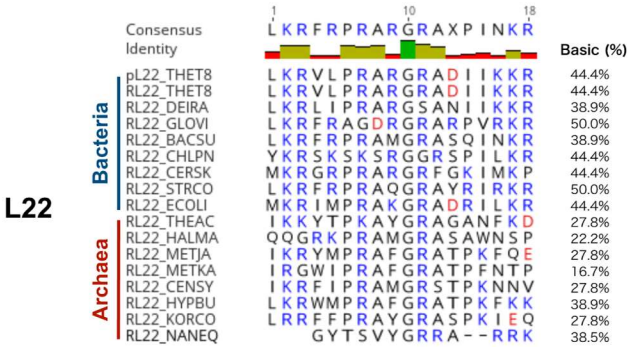


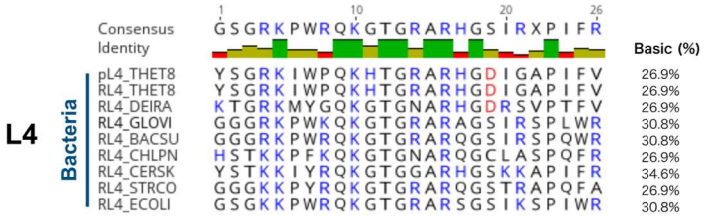
Supplementary Fig. 9. Phase diagram of mixtures of pLs (2, 3, 4, 15 and 22) with Sh bPTC as a function of the stoichiometric charge ratio $R_{+/-}$ (x axis, see *eq. 1* in **Mat. & Meth., *coacervation***) and of the inverse of the molar concentration of charged units Q^{-1} (y axis, see *eq. 2* ***ibidem***). Both axes are in log scale. Every sample examined contained some precipitates or aggregates (**Figure 4B**), thus the phase is indicated with "PREC/AGGR", with the same colour and marker code as in **Figure 4** and in **Supplementary Fig. 8**.











Bacteria

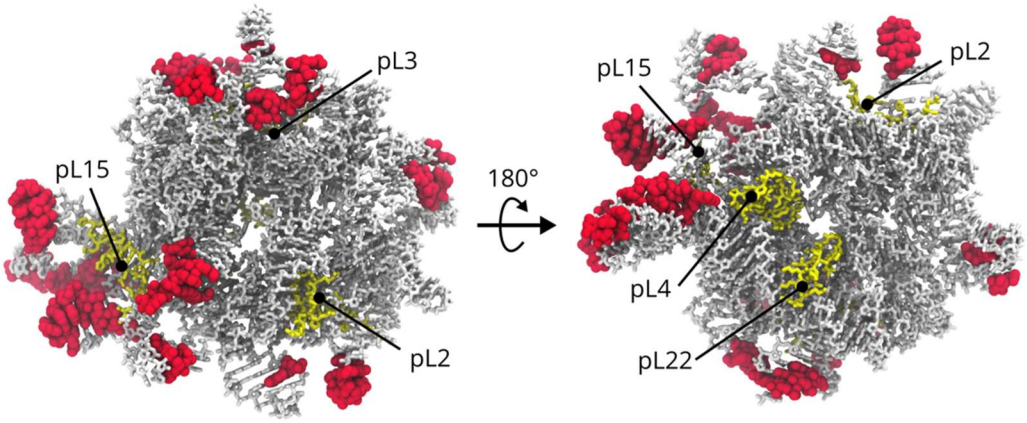
THET8: *Thermus thermophilus*
DEIRA: *Deinococcus radiodurans*
GLOVI: *Gloeobacter violaceus*
BACSU: *Bacillus subtilis*
CHLPN: *Chlamydia pneumoniae*
CERSK: *Cereibacter sphaeroides*
STRCO: *Streptomyces coelicolor*
ECOLI: *Escherichia coli*

Archaea

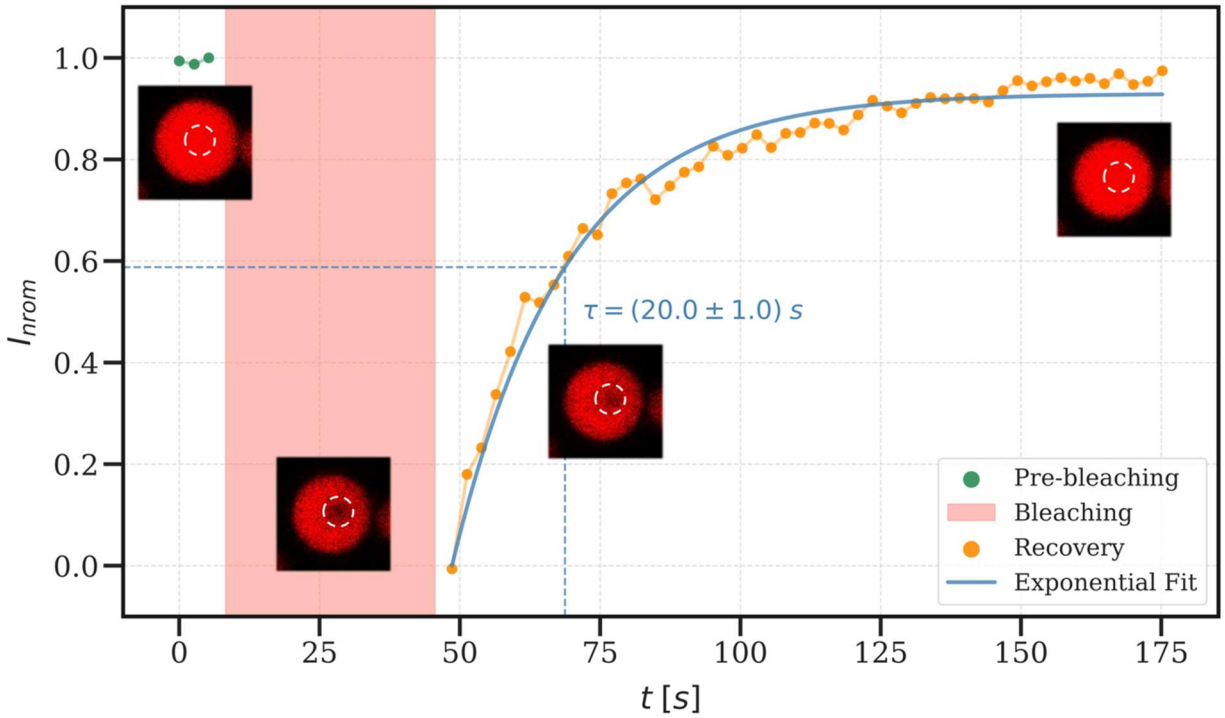
THEAC: *Thermoplasma acidophilum*
HALMA: *Haloarcula marismortui*
METJA: *Methanocaldococcus jannaschii*
METKA: *Methanopyrus kandleri*
CENSY: *Cenarchaeum symbiosum*
HYPBU: *Hyperthermus butylicus*
KORCO: *Korarchaeum cryptofilum*
NANEQ: *Nanoarchaeum equitans*

Supplementary Fig. 10. Multiple sequence alignment of rPeptide residues with diverse bacterial and archaeal ribosomal proteins.

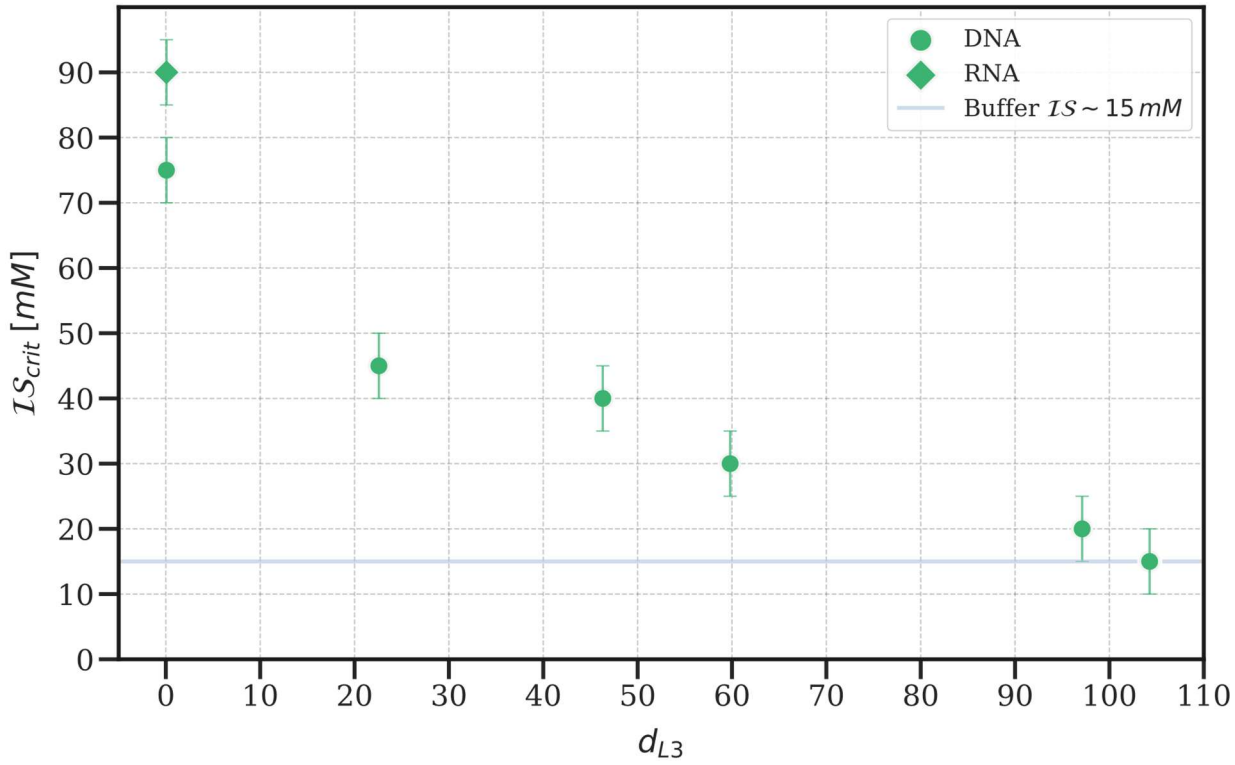
Each rPeptide (pL2, pL3, pL4, pL15 and pL22) sequence was aligned with representative ribosomal protein sequences from 8 bacterial and 8 archaeal species. Multiple sequence alignment was conducted using Clustal Omega with default settings. Consensus and identity (highlighted to represent sequence conservation) for the aligned sequences are shown. Only the bacterial rL4 sequences are shown due to the lack of sequence conservation with archaeal rL4 proteins. Ratio of basic amino acids (R, K and H) is denoted next to each sequence. The full scientific names of the species are shown at the bottom right.



Supplementary Fig. 11. The effect of rPeptides on the structure of WT bPTC represented by RNA residues identified as the most structurally different between MD models obtained with and without rPeptides. Top 10% residues with the largest structural difference are in red, rRNA in white, rPeptides in yellow. The two viewpoints show the tRNA binding site (*left*) and the ancestral exit tunnel lined by pL4 and pL22 (*right*). For clarity, the location of pL3 is shown, although it is buried in the rRNA.



Supplementary Fig. 12. Fluorescence Recovery After Photobleaching (FRAP) experiment on the WT sPTC (2.36 μM) and pL3 (64.8 μM) mixture, with 1:200 TAMRA-labelled peptide (corresponding to R = 1.0 & Q = 0.64 mM). A small region (*dashed white circle*) inside a coacervate droplet was irradiated with a White Light Laser of a Leica Stellaris 8 DLS confocal microscope at high power for about 40 s. The plot shows that the fluorescence intensity, normalised with respect to the initial signal, fully recovers on a time scale of about 20 seconds, indicating a fluid phase.

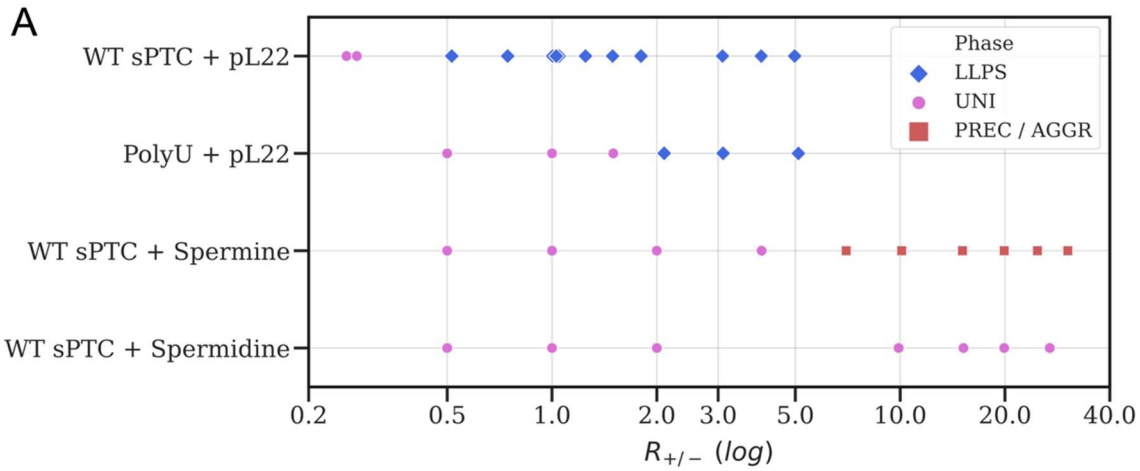


Supplementary Fig. 13. Scatter plot of the critical ionic strength (y axis) as a function of the distance (x axis) between the composition of the pL3 environment and that of several ssDNA unstructured oligomers and one RNA strand (see the legend and the "RNA" and "DNA" tags).

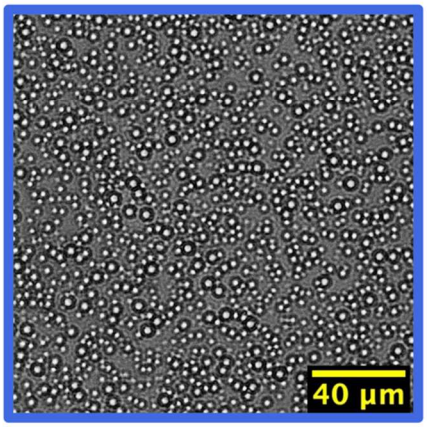
The critical ionic strength, measured at the same fixed DNA/RNA and peptide concentration and at ambient temperature for all the peptide-nucleic acid pairs, is defined as the highest NaCl concentration at which coacervation is still observed under a bright field microscopy setting. The distance d_{L3} is defined using the euclidean metric as:

$$d_{L3} = \sqrt{(A - A_{L3})^2 + (C - C_{L3})^2 + (G - G_{L3})^2 + (T - T_{L3})^2}$$

where A, C, G, T indicate the nucleobases percentage composition of the oligomers, while the same letters with "L3" subscript refers to the percentage composition of the pL3 *T. Thermophilus* rRNA environment, measured directly locating the 5 Angstrom closer residues to pL3 in PDB 4V51. Error bars account for a 10 mM experimental uncertainty on the measured critical ionic strength, while the grey horizontal line indicates the approximate buffer ionic strength (20 mM Tris, pH 7.6), which constitutes the experimental lower limit.

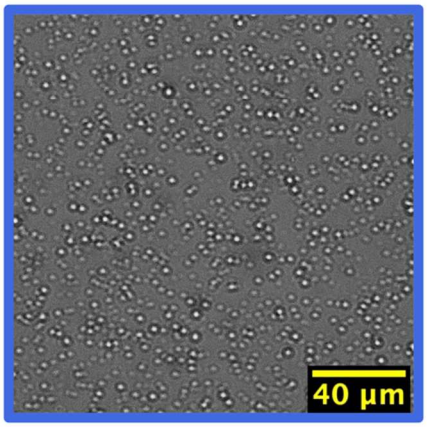


B) WT sPTC + pL22



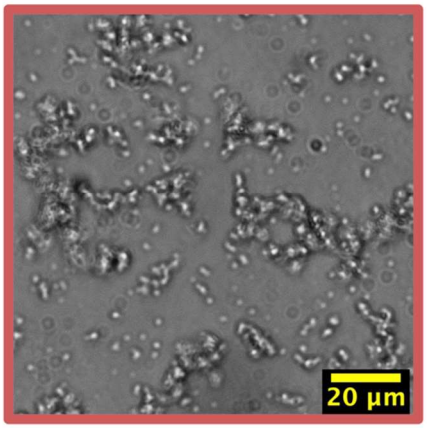
R = 1.0 – Q = 0.68 mM

C) PolyU + pL22



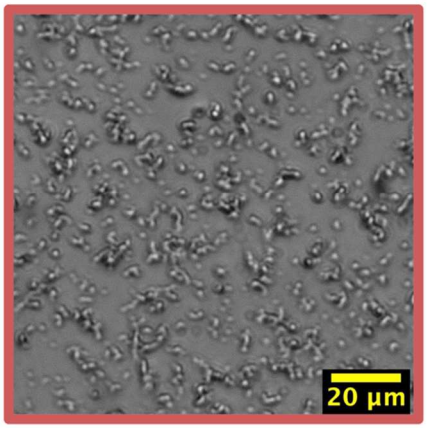
R = 5.1 – Q = 1.2 mM

D) WT sPTC + spermine



R = 10.1 – Q = 2.1 mM

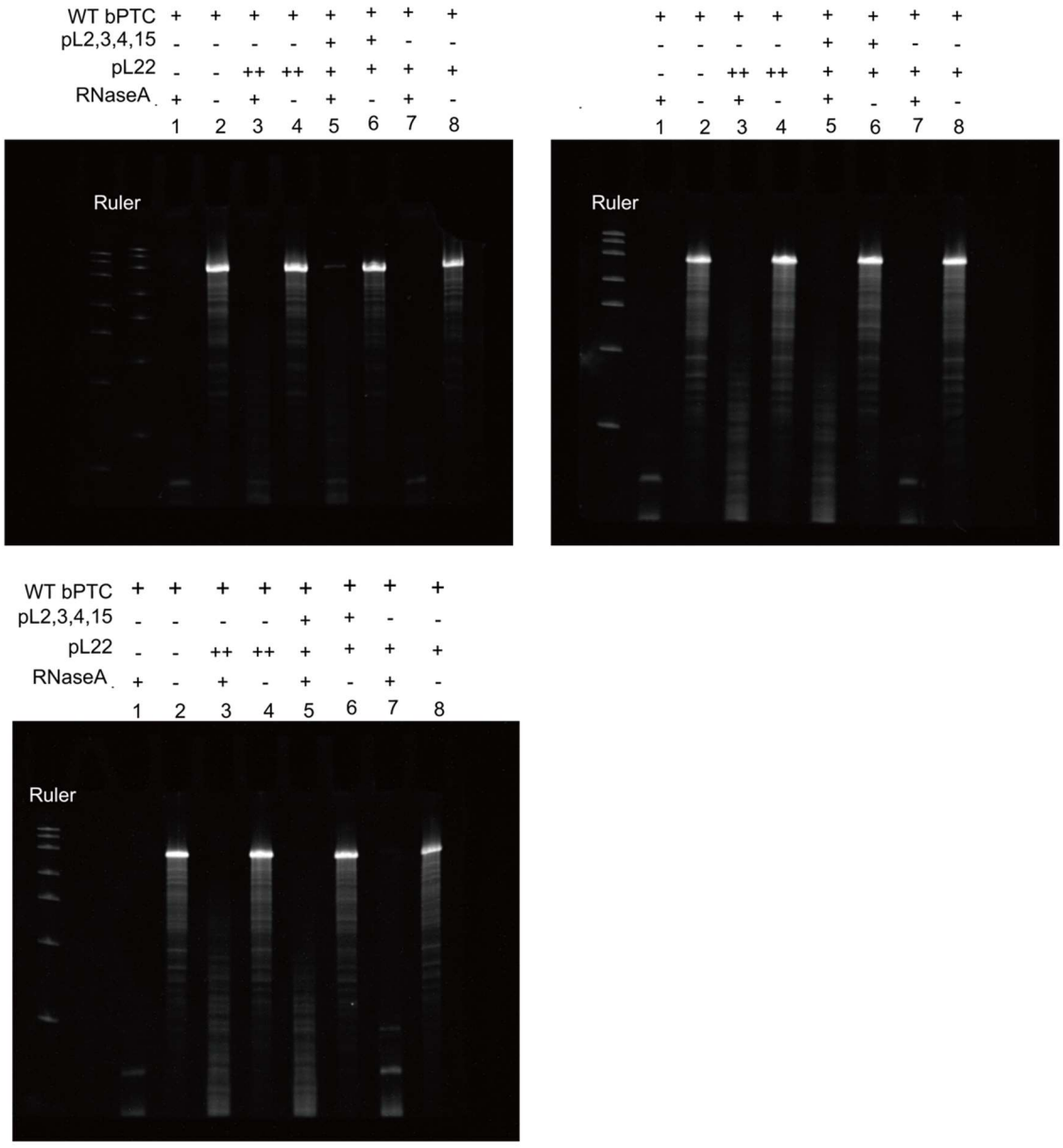
E) WT sPTC + spermine

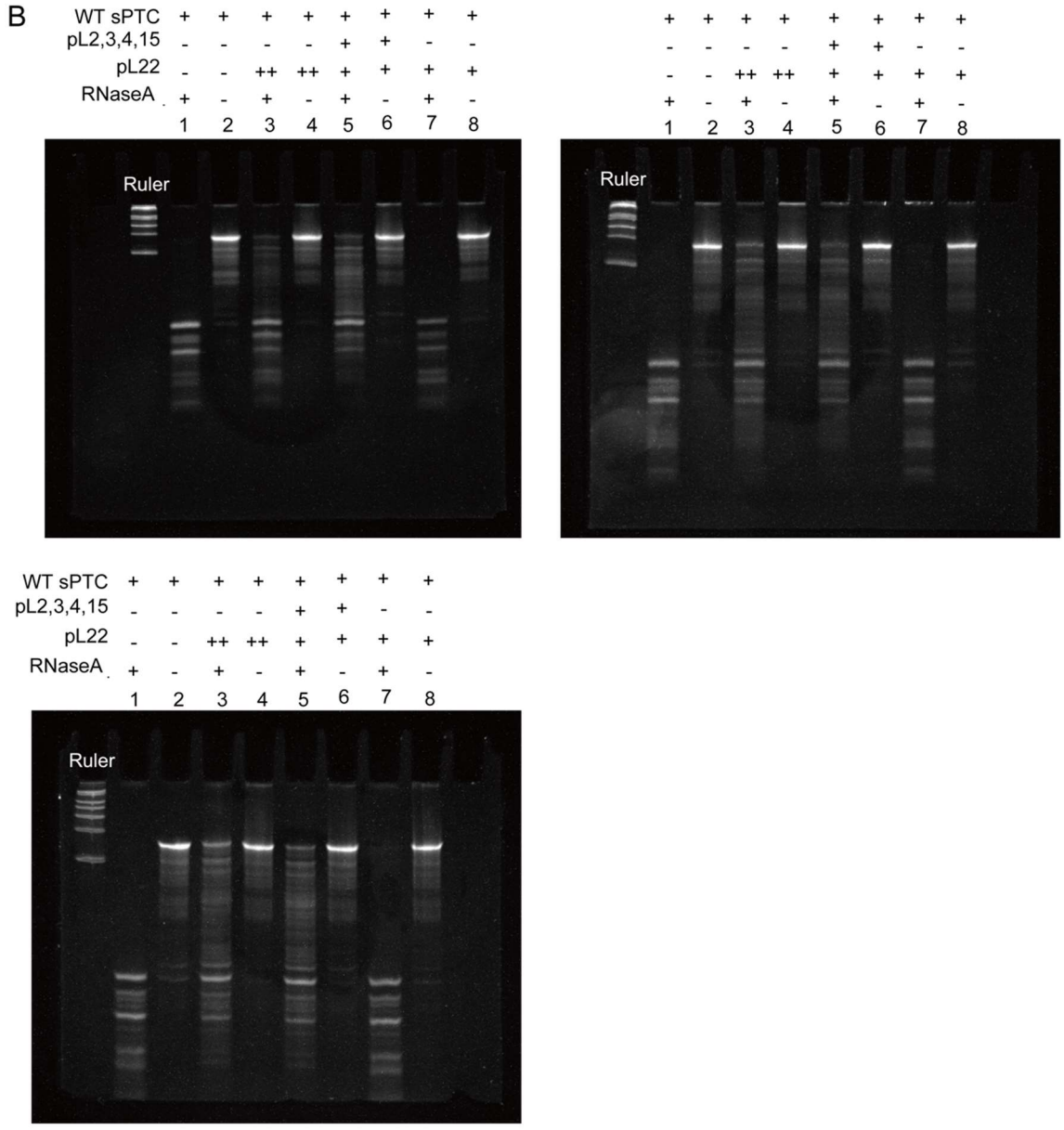


R = 19.9 – Q = 4.0 mM

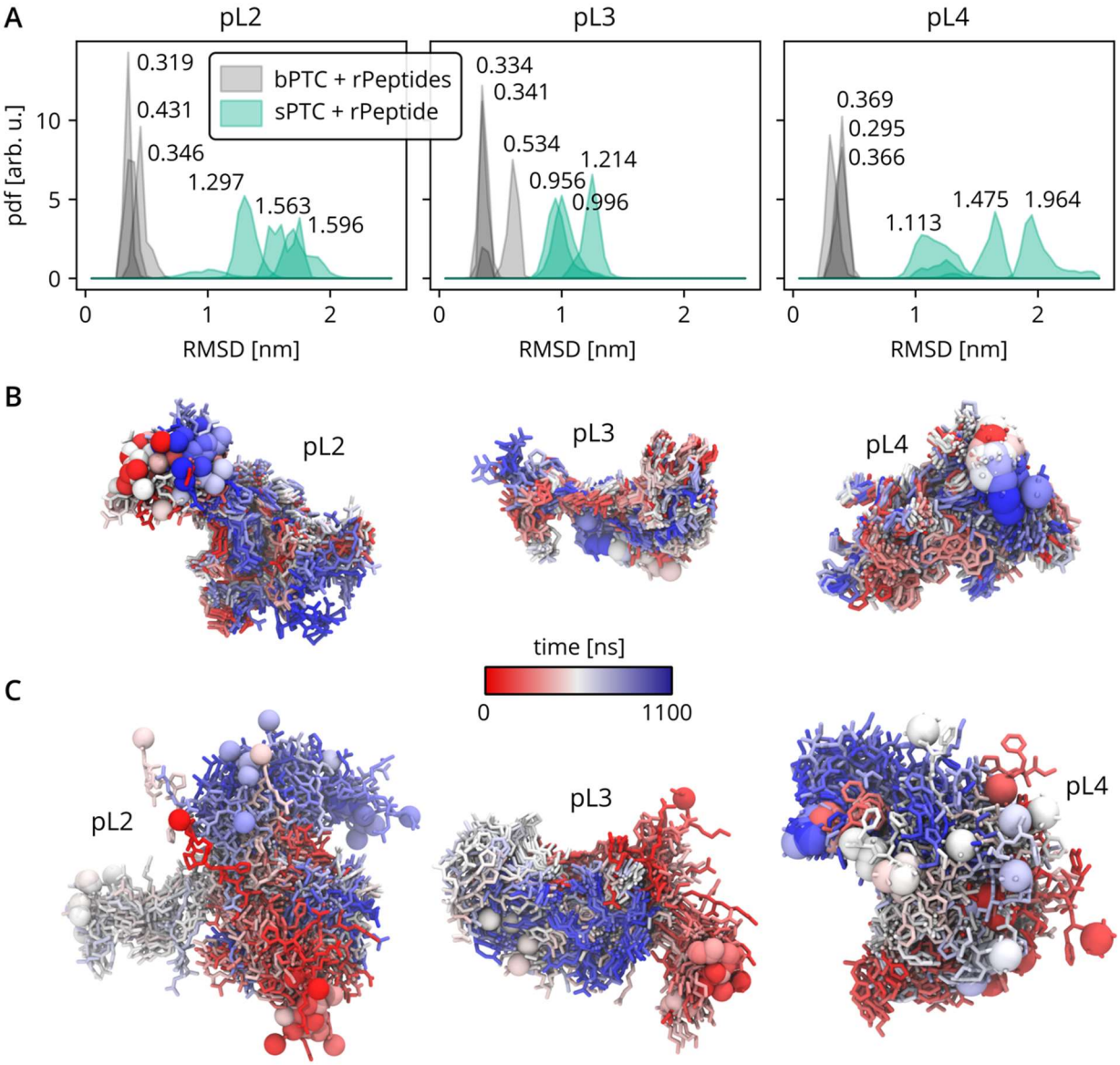
Supplementary Fig. 14. Comparison between the phase behaviours of various RNA + polycation systems, as assed by bright field microscopy.
(A) Observed phases as a function of $R_{+/-}$ (see *eq. 1* in **Mat. & Meth., *coacervation***) for the 4 mixtures investigated: WT sPTC + pL22 (same data as in **Supplementary Fig. 8** projected on the x axis), polyU + pL22, WT sPTC + spermine (+4 charges) and WT sPTC + spermidine (+3 charges). The RNA component (either WT sPTC or PolyU) was kept at fixed concentration (close to values promoting LLPS for WT sPTC with pL22) and different $R_{+/-}$ values have been tested by changing the polycation concentration. Different colours and markers are associated to different phases, with "UNI" indicating mixed states, "PREC/AGGR" the presence of precipitates or aggregates and "LLPS" the formation of liquid-like droplets.
(B - E) Optical micrographs showing the WT sPTC + pL22 mixture at R = 1.0 and Q = 0.68 mM (B), the PolyU + pL22 mixture at R = 5.1 and Q = 1.2 mM (C), the WT sPTC + spermine mixture at R = 10.1 and Q = 2.1 mM (D) and the WT sPTC + spermine mixture at R = 19.9 and Q = 4.0 mM (E).

A


WT bPTC + + + + + + + +
pL2,3,4,15 - - - - + + - -
pL22 - - ++ ++ + + + +
RNaseA + - + - + - + -
1 2 3 4 5 6 7 8
Ruler
+ + + + + + + +
- - - - + + - -
- - ++ ++ + + + +
+ - + - + - + -
1 2 3 4 5 6 7 8
Ruler
WT bPTC + + + + + + + +
pL2,3,4,15 - - - - + + - -
pL22 - - ++ ++ + + + +
RNaseA + - + - + - + -
1 2 3 4 5 6 7 8
Ruler



Supplementary Fig. 15. Full size Urea-gel analysis of the WT bPTC (*A*) and sPTC (*B*) RNaseA stability (see **Figure 5**). “Ruler” corresponds to RiboRuler ™ Low Range RNA ladder ready to use, Thermo Fisher Scientific.



Supplementary Fig. 16. Analysis of selected rPeptides (pL2, pL3 and pL4).
(A) The root-mean-square deviation (RMSD) of rPeptides with respect to their initial position after least-square superposition of the largest common rRNA element obtained from MD simulations of WT bPTC containing all rPeptides (*grey*) or WT sPTC with one rPeptide (*green*). The mean values of the probability distribution functions (pdfs) are denoted.
(B) Representative rPeptide conformations on the bPTC surface (*not shown*), color-coded according to the simulation time from red through white to blue. The spheres represent the C-terminal $C\alpha$ atoms.
(C) Same as (*B*) but for the MD simulations of the sPTC + rPeptide.

References

1. Hsiao,C., Lenz,T.K., Peters,J.K., Fang,P.-Y., Schneider,D.M., Anderson,E.J., Preeprem,T., Bowman,J.C., O'Neill,E.B., Lie,L., *et al.* (2013) Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucleic Acids Res.*, **41**, 3373–3385.

2. Bose,T., Fridkin,G., Davidovich,C., Krupkin,M., Dinger,N., Falkovich,A.H., Peleg,Y., Agmon,I., Bashan,A. and Yonath,A. (2022) Origin of life: protoribosome forms peptide bonds and links RNA and protein dominated worlds. *Nucleic Acids Res.*, **50**, 1815–1828.