Protein structure, stability, and dynamics in non-terrestrial systems

Prepared by Daniel P. Molland



Supervised by Professor Jason Schnell and Professor Jon Wade

Schnell Lab, Department of Biochemistry, St John's College, Oxford

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy at the University of Oxford

Protein structure, stability, and dynamics in non-terrestrial systems

Word Count: 40,815

Schnell Lab, Department of Biochemistry, University of Oxford

St John’s College, Oxford,

Hilary Term 2025

Abstract

Understanding how proteins and peptides assembled from abiotically synthesised amino acids could function under prebiotic conditions is central to assessing planetary habitability (Des Marais *et al.*, 2008). In this thesis, I investigate whether proteins composed solely of the "abiotic amino acid set" can adopt stable folds, coordinate metals, and exhibit catalytic potential. Through comparative analyses of early Earth, Mars, and Enceladus conditions, I evaluate how pH, salinity, temperature, and redox states constrain the folding and function of early proteins.

First, by mutating the metal-binding cysteine residues of rubredoxin to aspartate or serine, I show that non-thiol amino acids cannot replicate cysteine's role in coordinating iron and other transition metals in this model system, demonstrating the essentiality of cysteine for early transition metalloprotein function. I then expand on this finding through a large-scale computational screening of abiotic-like proteins, revealing that all tested abiotic-protein-like modern proteins adopt molten globule-like structures and are all sensitive to acidic pH. I observed that cysteine offered marginal structural gains to these proteins, but its most significant value came in the metal binding functionality it offered.

Subsequently, I demonstrate that cysteine in alkaline, reducing conditions substantially enhances surface-catalysed peptide synthesis. Cysteine's thiol moiety both increases peptide yields and introduces an L-stereoselective bias during synthesis. The peptides produced from this reaction go on to enable a novel thiol-peptide selective peptide ligation pathway, enabling the rapid

formation of extended cysteine-bearing proteins that may have played an essential role in the formation of the first biocatalysts. Finally, I compare these experimentally derived constraints to predicted early planetary environments. While acidic, oxidising conditions on Mars appear unfavourable, the alkaline, metal-rich waters of Archean Earth and Enceladus could have supported functional, cysteine-based protein emergence. Taken together, these results constraint the minimal environmental and amino acids needed to enable abiotic protein synthesis and folding, offering stronger criteria for evaluating planetary habitability throughout the solar system.

Acknowledgements

I would like to acknowledge the team of people without which this research would not have been possible. Principally I want to thank Professor Jason Schnell and Professor Jon Wade, my supervisors, who for the past three years have helped me in delivering a one-of-a-kind research project that pushed each of us into a new and interdisciplinary field. I would also like to thank the European Space Agency, who funded this project and who (for whatever reason) endorsed and trusted me with over £65,000 worth of research funding when I had little more than a few cool ideas and a willingness to try them. I would also like to thank Dr Julie Keeble who supported me in my previous degree and who I first pitched this research proposal to nearly 4 years ago.

To execute this research, which was cross divisional and cross departmental by its very design, I worked with an array of researchers across the university. I would like to thank Professor Christina Redfield (NMR Facility, Department of Biochemistry), Dr Majorie Fournier (Proteomics Facility, Department of Biochemistry), Dr David Staunton (Biophysics Facility, Department of Biochemistry), Dr Rod Chalk (Mass Spectroscopy Facility, Centre for Medicines Discovery), Dr Robert Jacobs (Surface Analysis Facility, Department of Chemistry), Dr Elizabeth Philips (Department of Chemistry) Dr Richard Makin (Thin Film Facility, Department of Physics), Dr Nick Rees (NMR Facility, Department of Chemistry), Dr Julie Cosmidis (Department of Earth Sciences) and Phil Holdship (ICP-MS Facility, Department of Earth Sciences) for their advice and use of their research facilities to execute this project.

I also want to thank those who maintained my high morale throughout this project. Firstly, to my fellow riflemen at 7th Battalion the Rifles and UK Space Command where I served as an enhanced commitment reservist while conducting this research. I also want to thank my colleagues within the ground floor Biochemistry labs, Dr Oliver Adams, Dr Finaritra Raoelijaona, Sigurbjorn Markusson, Madeline Eaton and Isabella Rhyu for their good chats and ability to be jointly entertained by our many shared failures and successes (although mainly the former).

I would also like to acknowledge the other educators whose support helped me go from a high school drop-out to completing my DPhil. Thank you, Mike Neville, and Elaine Johnston for believing that someone first predicted for a U in A level Chemistry could go on to do a doctorate.

Finally, I would like to acknowledge my family and my partner Saffie. Thank you for your patience and bearing with me when I got locked into the 1000th hyperfocus session while analysing data or working on a new idea. I will buy you some Lego (or wine) to make up for it! (I hope)

"What stands in the way becomes the way"

Table of Contents

Protein structure, stability, and dynamics in non-terrestrial systems ... 1
Abstract ... 3
Acknowledgements ... 5
Table of Contents ... 7
Abbreviations ... 10
Key definitions ... 12
List of Tables ... 14
List of Figures ... 15
Conceptual Framework ... 18
Chapter 1: Introduction ... 23
A new lens for assessing planetary habitability. ... *24*
Planetary habitability ... *24*
Abiogenesis ... *28*
Viable routes for organic synthesis ... *31*
Definition of the solar systems organic inventory ... *39*
Folding nucleation of abiotic proteins and peptides by metal ions ... *47*
Transition Metal Coordination: Influence of Ionic Size, Charge, and Crystal Field Effects ... *51*
Abiotic solvent environments ... *51*
AlphaFold in Protein Structure Prediction ... *57*
Predictions of protein function with abiotic solvents ... *67*
Aim and Scope of the Thesis ... *72*
Hypothesis Statement ... *73*
Chapter 2: Elucidating the basis of abiotic metal ion coordination using an ancient metalloprotein model system. ... 75
Introduction ... 76
Study rationale ... *76*
Modern metalloproteins and metal-binding architectures ... *77*
Selectivity of amino acid side chains for transition metal coordination ... *80*
Criteria for selecting a model metalloprotein. ... *81*
Approach ... *82*
Methods ... 85
Purification ... *85*
Protein solution NMR spectroscopy ... *90*
Biophysical Analysis ... *92*
Proteomics ... *94*
Results ... 97
Conservation of rubredoxin structure and design of variants ... *97*
Purification and characterisation of rubredoxin variants ... *97*
Structural analysis using NMR spectroscopy. ... *104*
pH-dependent stability of RbdD ... *111*
Discussion ... 121
Assignment of apo-rubredoxin ... *121*
pH-dependent folding in RbdD ... *122*
Metal on coordination in the absence of cysteine ... *123*
Limitations and model constraints ... *124*

Conclusion *127*
Chapter 3: Abiotic protein stability and solubility 129
Introduction 130
Study rationale *130*
Abiotic proteins and the challenge of limited sidechain chemistry *131*
Previous attempts to characterise abiotic proteins. *133*
A broader search *134*
Approach *135*
Methods 137
Abiotic-like protein selection *137*
Protein expression and purification *140*
Circular Dichroism (CD) spectroscopy and stability assays *142*
Protein solution NMR spectroscopy *143*
Results 144
Abiotic protein secondary structure propensities *144*
Abiotic-like protein identification and structural classification *144*
Abiotic-like protein analysis *153*
Discussion 175
Abiotic-like proteins are pH sensitive and possess molten globular structures. *176*
Evaluating the hypothesis of metal-guided protein folding *179*
Limitations *180*
Conclusion *182*
Chapter 4: Investigating the role of cysteine in stereoselective surface-catalysed peptide synthesis 183
Introduction 184
Study rationale *184*
Abiotic peptide synthesis and limitations *184*
Cysteine's potential in abiotic peptide synthesis *185*
Thiol-peptide selective peptide ligation *186*
Approach *188*
Methods 190
Peptide synthesis *190*
Peptide purification and analysis *192*
Probing cysteine silicate surface interactions *195*
Monte Carlo simulations *198*
Peptide Ligation *199*
Results 203
Increased peptide yields with cysteine are due to its thiol group. *210*
Stereospecific reorientation of the amino group in bivalently silicate adsorbed D-amino acids *223*
Cysteines reactivity makes the abiotic synthesis of CXXC motif bearing peptides extremely likely. *223*
Metal ions catalyse ligation of thiol heptapeptides. *227*
Discussion 234
Cysteine as the enabler of high yield abiotic peptide synthesis in aqueous conditions *234*
Stereoselective peptide synthesis by bivalent mineral association *234*

Synthesised peptide lengths are inherently limited in mineral catalysed peptide synthesis. *236*
Reducing conditions enabled thiol-peptide selective peptide ligation. *237*
Limitations..... *238*
Conclusion..... *241*
Chapter 5: Evaluation of the functional planetary habitability of Earth, Mars, and Enceladus..... 243
Recap *244*
Experimentally derived abiotic protein folding constraints. 246
Study 1..... *246*
Study 2..... *248*
Study 3..... *250*
Summary *251*
Habitability of plausible abiotic solvents 255
Earth *255*
Mars..... *256*
Enceladus..... *258*
Comparison with recent works..... *261*
Conclusion 263
Chapter 6: Synthesis and future directions..... 265
Enabling the genesis of information transfer. 266
Future directions 270
Conclusion 271
Appendices..... 272
Appendix A – Figures..... *272*
Appendix B – Tables..... *278*
Appendix C – Code..... *287*
Bibliography..... 311

Abbreviations

AEX – Anion Exchange Chromatography

ANOVA – Analysis of Variance

BLAST – Basic Local Alignment Search Tool

CD – Circular Dichroism

CXXC – Cys-X-X-Cys Motif

DSF – Differential Scanning Fluorimetry

EDTA – Ethylenediaminetetraacetic Acid

ESA – European Space Agency

HCD – Higher-Energy Collisional Dissociation

HSAB – Hard and Soft Acids and Bases

HSQC – Heteronuclear Single-Quantum Coherence (NMR)

IPTG – Isopropyl β-D-1-thiogalactopyranoside

JAXA – Japan Aerospace Exploration Agency

LC-MS – Liquid Chromatography–Mass Spectrometry

MALS – Multi-Angle Light Scattering

NASA – National Aeronautics and Space Administration

NMR – Nuclear Magnetic Resonance

OSIRIS-REx – Origins, Spectral Interpretation, Resource Identification, Security, Regolith Explorer

PR-IRRAS – Polarisation-Resolved Infrared Reflection Absorption Spectroscopy

qNMR – Quantitative Nuclear Magnetic Resonance

RNA – Ribonucleic Acid

SPE – Solid Phase Extraction

SDS-PAGE – Sodium Dodecyl Sulfate–Polyacrylamide Gel Electrophoresis

SEC – Size-Exclusion Chromatography

SEC-MALS – Size-Exclusion Chromatography-Multi-Angle Light Scattering

TCA – Trichloroacetic Acid

TCEP – Tris(2-carboxyethyl)phosphine

TFA – Trifluoracetic Acid

TIMS-TOF – Trapped Ion Mobility Spectrometry–Time-of-Flight

TOCSY – Total Correlation Spectroscopy

TSPL – Thio-peptide selective peptide ligation

UV-Vis – Ultraviolet–Visible Spectroscopy

Key definitions

Abiogenesis

The process by which life arose from non-living matter, requiring the formation and evolution of the first biologically active molecules.

Pre-biotic

Compounds that emerged before the origin of life.

Abiotic

Compounds that can emerge without biological involvement.

Abiotic Amino Acid Set

10 amino acids (Gly, Ala, Asp, Glu, Val, Ile, Leu, Pro, Ser, Thr) found in abiotic environments.

Biotic Amino Acid Set

The remaining 10 amino acids (Asn, Gln, Cys, Met, Phe, Trp, Tyr, Arg, His, Lys) found in modern organisms.

Molten Globule

A partially folded protein conformation with well-formed secondary structure but loosely packed tertiary structure.

Peptide Ligation

The process of forming peptide bonds between two peptide polymers.

Peptide Synthesis

The process of forming peptide bonds between two monomers or a polymer and a monomer.

Planetary Habitability

The capacity of an environment to support the emergence and sustainment of life. Defined by Des Marais et al 2008 as locations containing habitable environments which 'provide extended regions of liquid water, conditions favourable for the assembly of complex organic molecules, and energy sources to sustain metabolism'.

Planetesimal

A solid, km-scale non-differentiated (i.e. not yet massive enough to form a molten planetary core system) that formed in a protoplanetary disk through the coagulation of dust and ice, serving as a building block for planetary.

Proteogenic

The twenty amino acids that are incorporated into proteins by the modern ribosomal machinery.

Redox Potential

A measure of an environment's ability to gain or lose electrons.

Stereoisomer

Molecules that have the same molecular formula but differ in the three-dimensional orientations of their atoms when coordinated to a chiral centre with four different bonded groups.

Stereoselective

A property of a chemical reaction that leads preferentially to an increased use or yield of one stereoisomer over another (e.g., L- vs. D-amino acids preference in ribosomal synthesis).

Thiol

A functional group containing an -SH moiety.

List of Tables

Table 1.01 – Summary of abiotic solvent conditions described. 56
Table 3.01 – Summary table of observed characteristics in abiotic-like proteins 178
Table 5.01 – The protein function ‘goldilocks’ zone 254
Appendix B - Table 01 – Proteins identified from the Cys- and Cys+ run. 279
Appendix B - Table 02 – Statistical analysis of data presented in Figure 4.03a. 282
Appendix B - Table 03 – Statistical analysis of data presented in Figure 4.03b 283
Appendix B - Table 04 – Compositions of peptides detected by TIMS-TOF MS 284
Appendix B - Table 05 – Statistical analysis of data presented in Figure 4.08 285
Appendix B - Table 06 – Upper and lower bound cysteine CXXC probability 286

List of Figures

Figure 1.01 – The abiotic amino acid set. 21
Figure 1.02 – Planets and moons with surface or subsurface liquid water oceans. 26
Figure 1.03 – Surface catalysed peptide synthesis 38
Figure 1.04 – Various missions have defined the solar systems abiotic amino acid inventory. 42
Figure 1.05 – Metal guided abiotic protein folding provides an alternate route for early proteins to fold around soluble metal ions. 50
Figure 1.06 – Liquid water was present on Earth and Mars up to 4.1 billion years ago. 61
Figure 1.07 – Enceladus and Europa contain subsurface liquid oceans that are actively ejecting into space. 66
Figure 2.01 – Rubredoxin possesses the simplest Fe/S cluster structure observed in modern biology. 79
Figure 2.02 – Rubredoxin shows complete conservation of its melting binding site and hydrophobic core. 98
Figure 2.03 – AlphaFold3 predictions of consensus Rubredoxin and its variants 99
Figure 2.04 – Rubredoxin purifies readily without requiring affinity labelling based purification. 100
Figure 2.05 – RbdD purifies similarly to RbdC. 102
Figure 2.06 – RbdD and RbdC are monomeric and nearly pure in solution following purification. 103
Figure 2.07 – RbdDS1 and RbdDS2 purifies similarly to RbdD. 105
Figure 2.08 – Apo-RbdC is structured in solution. 106
Figure 2.09 – Metal binding causes localised chemical shift perturbations in apo-RbdC. 107
Figure 2.10 – Melting binding increases rubredoxin melting temperature 109
Figure 2.11 – Rubredoxin has limited flexibility in the apo and holo state. 110
Figure 2.12 – RbdD undergoes global unfolding at neutral pH. 112
Figure 2.13 – RbdD shows a salt dependent titration curve. 113
Figure 2.14 – RbdD aspartate deprotonation drives protein unfolding. 116
Figure 2.15 – RbdDS1 and DS2 remain folded at pH 6.2. 117
Figure 2.16 – Rubredoxin variants do not bind to transition metals at 2:1 ratio. 119
Figure 2.17 – Rubredoxin variants exhibit surface localised CSP's at 10:1 ligand to protein ratios. 120
Figure 3.01 – Common secondary structural types observed in modern proteins. 132
Figure 3.02 – Bioinformatic approach 139
Figure 3.03 – Abiotic proteins will possess a helical propensity based on observed abiotic amino acid abundance. 145
Figure 3.04 – Abiotic-like proteins remain over-enriched in hydrophobic amino acids. 147
Figure 3.05 – Incorporation of cysteine within the search parameters increases the number of functional proteins identified. 150
Figure 3.06 – Proteins identified within the Cys- search are dominated by extended helices and beta solenoids. 151
Figure 3.07 – The Cys+ search reveals two potentially iron binding abiotic-like proteins. 152
Figure 3.08 – SALSA purifies both monomerically and dimerically by SDS-PAGE 155
Figure 3.09 – SALSA purifies as an alpha helical protein. 156
Figure 3.10 – SALSA undergoes cooperative unfolding. 157
Figure 3.11 – SALSA is structurally stable against changes in pH from pH 4-10. 158

Figure 3.12 – SALSA begins to precipitate upon increasing salinity. 159
Figure 3.13 – LysW and Fe-LysW-C purify monomerically. 160
Figure 3.14 – Fe-LysW-C binds to iron 161
Figure 3.15 – Metal binding causing increased structure in Fe-LysW-C. 164
Figure 3.16 – LysW and Fe-LysW-C does not undergo cooperative unfolding with increasing temperature. 165
Figure 3.17 – Metal binding increases pH stability in Fe-LysW-C. 166
Figure 3.18 – Metal binding increases Fe-LysW-C resistance to salt induced protein unfolding. 167
Figure 3.19 – Ferredoxin is purified in the apo-state when expressed in *E.coli*. 170
Figure 3.20 – Ferredoxin adopts a mixed α/β structure upon purification in *E.coli*. 171
Figure 3.21– Ferredoxin does not undergo cooperative unfolding up to 90℃. 172
Figure 3.22 – Ferredoxin is sensitive to acidic pH's. 173
Figure 3.23 – Ferredoxin is insensitive to changes in salinity. 174
Figure 4.01 – Proposed mechanism of thiol-peptide selective peptide ligatio 187
Figure 4.02 – The presence of a free thiol side chain significantly increases peptide yield in the presence of silicate. 204
Figure 4.03 – Cysteine stereo-specifically increases peptide yield in the presence of an achiral oxide mineral. 206
Figure 4.04 – Peptide synthesis with cysteine yields a diverse set of cysteine-containing peptides. 208
Figure 4.05 – Peptide synthesis in the presence of excess methionine produces a low diversity of products. 209
Figure 4.06 – Hydrophobic amino acids are enriched in peptides C18 SPE purified and detected by TIMS-TOF MS. 212
Figure 4.07 – C18 purification prior to injection increases the length of detected peptides by TOF MS 213
Figure 4.08 – Synthesis with L-cysteine increases peptide yield and the length of detected peptides. 214
Figure 4.09 – Abiotic peptide bond formation in the presence of L-cysteine correlates with metal-sulfur bond dissociation enthalpies. 215
Figure 4.10 – Proposed cysteine catalysed mineral peptide synthesis mechanism. 217
Figure 4.11 – D-cysteine and L-cysteine association with silicates is bivalent at alkaline pH. 218
Figure 4.12 – A covalent bond between the cysteine thiolate and silicate is detected by PR-IRRAS. 220
Figure 4.13 – Comparison of the ^{13}C solid-state NMR spectra of silicate-bound L-cysteine and free L-cysteine and free L-cystine. 221
Figure 4.14 – Bivalent surface attachment stereo-specifically orients the amino acid amino group. 224
Figure 4.15 – Probability of CXXC Motif Formation in 10-mer Peptides as a Function of Cysteine Concentration, with and without Adjusting for Reactivity 226
Figure 4.16 – Arc4 spontaneously coordinates Fe^{3+} and Co^{2+} under alkaline conditions. 228
Figure 4.17 – Arc4 is stable at 90℃ and pH 9.5 for up to 36 hours. 231
Figure 4.18 – Arc4 is ligated in the presence of Zn^{2+} and Fe^{2+} under reducing conditions. 232
Figure 4.19 – New ^{1}HN-$^{1}HC\alpha$ and ^{1}HN-$^{1}HC\beta$ resonances following heating in the presence of Zn^{2+} are consistent with peptide ligation. 233
Figure 5.01 – Conceptual approaches for assessing the habitability of an aqueous environment. 262

Figure 6.01 – Theory for abiogenesis through cysteine bearing compounds. 269
Appendix A - Figure 01 – Metal addition at pH 6.2 does not rescue RbdD folding. 272
Appendix A - Figure 02 – RbdDS1 and RbdDS2 aspartates remain deprotonated at pH 4.7. 273
Appendix A - Figure 03 – Aspartate destabilises soluble FeS clusters. 274
Appendix A - Figure 04 – Intact mass of ferredoxin 275
Appendix A - Figure 05 – Purified ferredoxin purifies as an apo-protein 276
Appendix A - Figure 06 – Thiol peptides oxidise under heating at 90°C under non reducing conditions. 277

Conceptual Framework

How life spontaneously began on early Earth and how it may emerge elsewhere is of significant interest to both the public and the research community. To evaluate an environment's capacity to support life, NASA uses the concept of ‘planetary habitability’ (Des Marais *et al.*, 2008). This metric considers several factors, the most relevant of which to the formation of Earth-like life is the compatibility of an environment's aqueous conditions with the synthesis and function, i.e. the catalytic and reactive ability, of organic species (McDonald and Storrie-Lombardi, 2010). The function of these organic compounds is necessary for both the formation and sustainment of life, and their functionality is typically constrained by solution, temperature, salinity, redox potential, and pH (McKay, 2014)(Merino *et al.*, 2019)(Pillot *et al.*, 2023). These physio-chemical parameters are also crucial because they heavily influence other factors essential for biochemical processes, including metal and micronutrient solubility and availability, reaction kinetics, viable energy sources, and ionic strength (Kitadai and Maruyama, 2018).

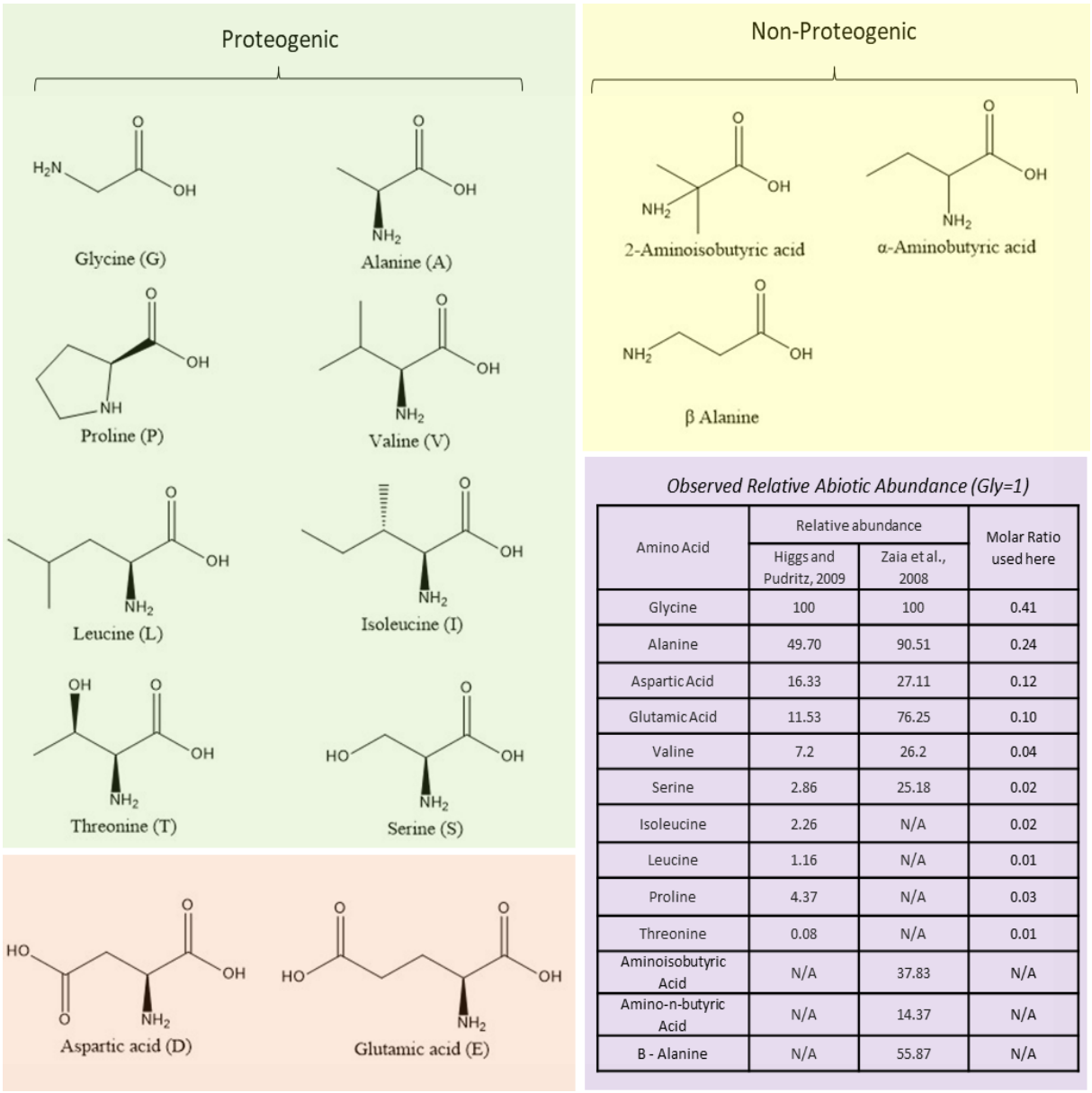
Among the organic species likely present on Earth prior to the emergence of life, amino acids and their peptide polymers are of particular relevance as they provide the basis for nearly all modern biological catalytic activity (Alberts *et al.*, 2002). It is currently predicted that noncoding peptides and proteins were essential enablers of the RNA world, making their functional limits in relation to key solvent parameters such as pH, salinity, temperature and redox potential

relevant to both RNA and RNA/peptide origin-of-life theories (Fried *et al.*, 2022). While significant research into abiotic RNA synthesis and function is also ongoing, defining the stability and functional limitations of abiotic RNA is more complex (Yadav *et al.*, 2020)(Gilbert, 1986). Unlike RNA nucleotides, amino acids are relatively abundant in meteorites and can be synthesised and polymerised into peptides under plausible abiotic conditions.

Abiotic and early RNA coded peptides (fewer than 50 amino acids long) and proteins (greater than 50 amino acids long) would have been restricted to the amino acids that are synthesisable without life, the so-called “abiotic amino acid set,” which is composed of ten proteogenic amino acids (**Fig 1.01**) (Gly, Ala, Asp, Glu, Val, Ile, Leu, Ser, Thr, Pro) (Zaia *et al.*, 2008)(Higgs and Pudritz, 2009). However, definitive confirmation of the contents of the proteogenic amino acid set has only recently been provided by the asteroid sampling missions NASA OSIRIS-REx and JAXA Hayabusa2 (Mojarro et al., 2024)(Potiszil et al., 2023). In addition, cysteine has been proposed as an addition to the abiotic amino acid set in sulfur-rich environments, where cysteine is synthesisable through the conversion of serine to cysteine in the presence of H_2S and ferrocyanide (Foden *et al.*, 2020). Taken together these recent works have set the stage for evaluating whether abiotically formed peptides and proteins can fold, and subsequentially function catalytically, which could have been essential to enable later stages in abiogenesis.

Building on these advances, the subsequent chapters focus on experimentally probing the functional limits of proteins composed solely of abiotic amino acids. By integrating these observations with the solution conditions

proposed on early Earth, Mars and Enceladus, this thesis will aim to describe the habitability of these sites to support abiotic protein function, enabling stronger assessments of planetary habitability to be made of these sites. To do this, my thesis will assess the ability of abiotic proteins to coordinate transition metals, their stability under various solution conditions, and apply my findings on the pivotal role of cysteine within peptide synthesis pathways. Through these efforts, this thesis will aim to clarify which abiotic solvent environments enable abiotic proteins to fold and function as well as identify novel routes for abiogenesis to occur at these sites.



Observed Relative Abiotic Abundance (Gly=1)

Amino Acid	Relative abundance		Molar Ratio used here
	Higgs and Pudritz, 2009	Zaia et al., 2008	
Glycine	100	100	0.41
Alanine	49.70	90.51	0.24
Aspartic Acid	16.33	27.11	0.12
Glutamic Acid	11.53	76.25	0.10
Valine	7.2	26.2	0.04
Serine	2.86	25.18	0.02
Isoleucine	2.26	N/A	0.02
Leucine	1.16	N/A	0.01
Proline	4.37	N/A	0.03
Threonine	0.08	N/A	0.01
Aminoisobutyric Acid	N/A	37.83	N/A
Amino-n-butyric Acid	N/A	14.37	N/A
B - Alanine	N/A	55.87	N/A

Figure 1.01 – The abiotic amino acid set

The currently accepted abiotic amino acid set with proteogenic abiotic amino acids arranged by charge state at neutral pH (uncharged, green), (anionic, amber)) (Higgs and Pudritz, 2009). *Non-proteogenic abiotic amino acids are shown as proposed separately by* (Zaia *et al.*, 2008) *(yellow),with the relative abundances previously reported. The meteoric molar ratio reported previously and used throughout this study are also presented (purple).*

Chapter 1: Introduction

A new lens for assessing planetary habitability.

While the presence of water and organic compounds is widely recognised as the single most important criterion for planetary habitability, relatively little attention has been given to evaluating the functional potential of abiotic peptides and proteins. To help bridge this gap, this introduction will consider the importance, theoretical availability, and observed abundance of amino acids, potential routes to their polymerisation, and the expected environmental conditions of key astrobiological locations. Subsequently, I make predictions of the behaviour of abiotic peptides and proteins, given their decreased size, lack of conservative synthesis, and reduced sequential diversity, which helps to frame the areas explored within later data chapters. In Chapter 5, I will coalesce the solution conditions I have identified as necessary to support abiotic protein function and consider them in the context of the conditions proposed for early Earth, Mars, and Enceladus. This will provide a new lens through which to consider their planetary habitability not just based on their organic content abundance, but also their functional limitations.

Planetary habitability

The earliest attempts to assess planetary habitability focused on the detection of water, driven by the widely held belief that wherever there is liquid water, life is possible. Space agencies such as NASA, ESA, and JAXA allocated significant resources to search for water throughout the solar system. However, these investigations revealed water is in fact abundant in the form of surface ice on Mars (Piqueux *et al.*, 2019), the Moon (Honniball *et al.*, 2021), and in subsurface oceans on moons like Europa (Carr *et al.*, 1998), Enceladus (Waite

et al., 2011), Callisto (Khurana *et al.*, 1998), and Ganymede (Saur *et al.*, 2015) (**Fig 1.02**). While the presence of water at these sites was intriguing, it quickly became clear that water alone is not a biosignature and should instead be seen as one of several criteria in determining the habitability of an environment (McKay, 2014). Other critical variables include temperature, the availability of energy sources, the presence of essential elements like carbon, hydrogen nitrogen, oxygen, phosphorus, and sulfur (CHNOPS), and the stability of the environment's atmosphere, magnetic field, and geology (Schwieterman *et al.*, 2018).

All life on Earth is based on carbon chemistry, organic molecules such as nucleic acids and proteins composed of nitrogenous organic compounds enable self-organisation, genetic transfer, and metabolic processes (Alberts *et al.*, 2002). While it is important not to limit the search for life to an Earth-centric perspective, the broad consensus is that all life, regardless of where it emerges, would require water and be carbon-based due to carbon's unparalleled chemical diversity (Pace, 2001). Carbon's capability to form stable bonds with itself and various other elements to form long-chain polymers facilitates the formation of complex biological macromolecules in a way that is unmatched by any other known element. Although cosmochemically abundant elements such as silicon have been considered, silicon's greater atomic size and lower bond strength limit its ability to form the complex, stable, and versatile molecules required for life's processes (Petkowski *et al.*, 2020). Additionally, silicon's tendency to react with elements such as oxygen, nitrogen, sulfur, and phosphorus to form solid,

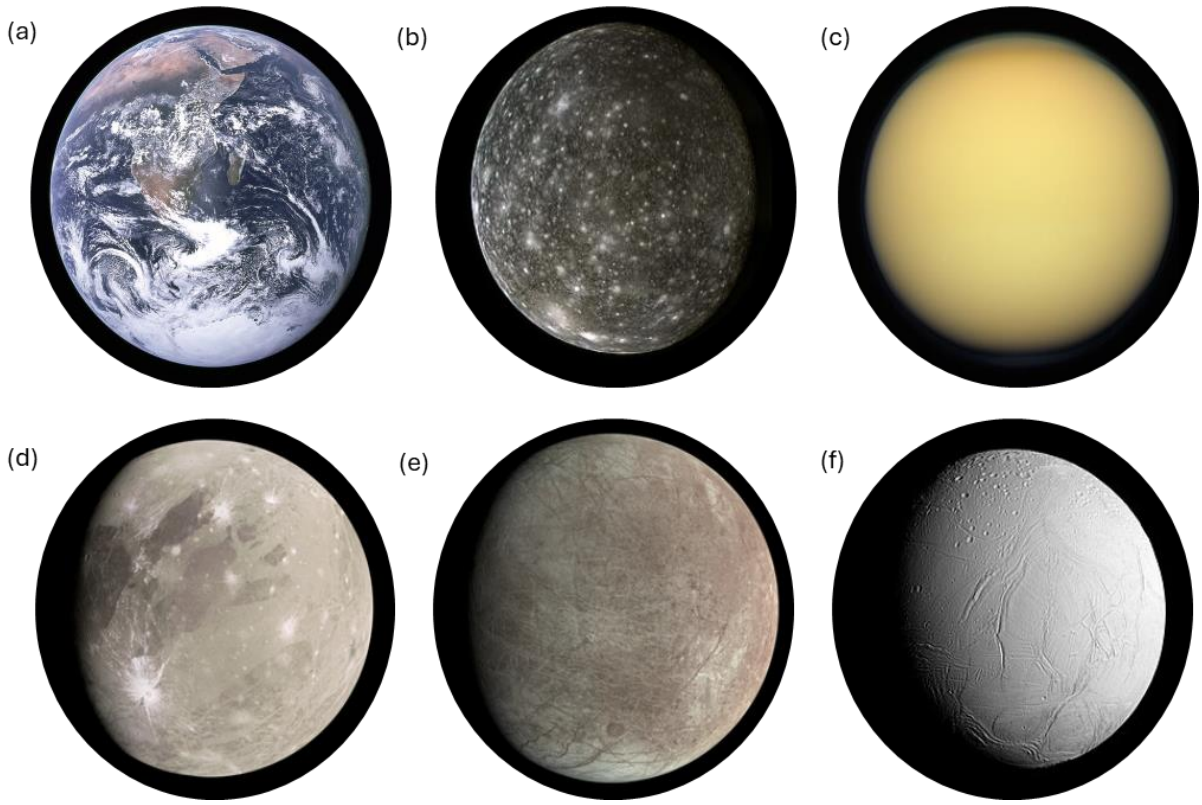


Figure 1.02 – Planets and moons with surface or subsurface liquid water oceans.

*Examples of planetary bodies with confirmed or suspected liquid water oceans. (**a**) Earth image captured during the Apollo 17 mission (NASA). (**b**) Callisto, Jupiter's second-largest moon, as imaged by the NASA Voyager 2 and combined with high-resolution ground images (NASA/JPL-Caltech/Kevin M. Gill). (**c**) Titan, Saturn's largest moon, imaged by the NASA Cassini orbiter (NASA/JPL-Caltech/SSI/Kevin M. Gill); (**d**) Ganymede, Jupiter's largest moon, photographed by the NASA Juno spacecraft in 2021 (NASA/JPL-Caltech/SwRI/MSSS/Kevin M. Gill); (**e**) Europa, Jupiter's 4th largest moon, imaged by the NASA Juno spacecraft in October 2023 (NASA/JPL-Caltech/SwRI/MSSS/Kevin M. Gill); (**f**) Enceladus, Saturn's 6th largest moon, imaged by the NASA Cassini orbiter in October 2015 (NASA/JPL).*

insoluble materials make it highly unlikely that silicon could form the basis of any form of life (National Research Council, 2007).

Similarly, alternatives to water as a solvent for life, such as methane (CH_4) and ammonia (NH_3), have been explored. Methane, which exists in liquid form on the surface of Saturn's moon Titan, is non-polar and lacks the ability to dissolve polar organic molecules, limiting its use as a solvent for life (Pohorille and Pratt, 2012). Though ammonia is a polar solvent and can form hydrogen bonds, its hydrogen-bonding ability is significantly weaker than water and boils at -78 °C, limiting potential reaction rates and viable kinetic pathways (National Research Council, 2007)**.** None of the other solvents considered have water's key advantages, such as its ability to form strong hydrogen bonds, high polarity and dielectric constant, reactivity, high thermal capacity, and abundance, all of which facilitate and directly engage in biochemical processes (Brack, 1993)(Vieira *et al.*, 2020).

As such, exploring evidence for locations with liquid water and carbon-based materials is a logically sound process, but it forms only the first stage of determining habitability. Such investigations must also assess, although rarely have, the potential for that environment to support functions of carbonaceous molecules useful for life. In any environment containing simple organic species, the molecules must remain from planetary formation, be synthesised locally, or be delivered exogenously from carbon-rich sources such as asteroids and comets (Chyba and Sagan, 1992). Numerous research programs have been initiated to (**a**) define viable routes for abiotic organic synthesis, (**b**) determine the chemical inventory of pre-planetary formation materials and asteroids capable of

delivering these compounds to a planet (e.g., direct sampling of carbonaceous asteroids in space), and (**c**) directly sample current or previously water-rich environments themselves. These important efforts have dominated the fields of astrobiology and planetary science for the past several decades and remain the focus of ongoing exploration.

Abiogenesis

Before reviewing previous work on the search and classification of the abiotic organic inventory, which informs which amino acids make up ancient abiotic peptides and proteins, it is worthwhile to consider the methods through which these organics are proposed to enable the emergence of life. Although not the primary focus of this work, which concentrates on the functionality of abiotic proteins that may enable these processes, it is useful to understand why amino acids and their polymers are relevant to all current theories of abiogenesis.

The most prominent theory for the origin of life is the “RNA world” hypothesis, initially formulated by Carl Woese and Leslie Orgel and coined by Walter Gilbert in 1986, it posits that life began with the formation of self-replicating RNA molecules (Gilbert, 1986). The capability of RNA to engage in biological catalysis, as observed in modern ribozymes, and its ability to independently self-replicate is proposed as the key enablers of abiogenesis (Cech, 1990)(Altman, 1990)(Wochner *et al.*, 2011). RNA-based reactions were proposed by Gilbert as the most straightforward route through which catalysis and self-replication can emerge spontaneously. However, RNA’s synthesis complexity and chemical instability are fundamental limitations for the RNA world theory (Oivanen *et al.*, 1998) Furthermore, RNA appears to be particularly unstable at the temperatures

and pH of energetically viable sites within the early Archaean seas (Levy and Miller, 1998)(Yadav *et al.*, 2020)(Joyce, 2002). Given these factors, an increasingly common hypothesis is that the RNA world plays a key role in abiogenesis but emerged later and was preceded by other reactions more likely to proceed in the presence of abiotically available molecules (Le Vay and Mutschler, 2019).

An example of one such pathway is the RNA-peptide world theory that proposes that amino acids and peptides acted as both enablers of the synthesis of RNA and provided stabilisation that enabled RNA to be more resistant to the thermal and hydrolytic stresses in the Archean Ocean (Nainytė *et al.*, 2020)(Fried *et al.*, 2022). Supporting work has demonstrated that peptides can stabilise RNA, although such interactions typically depend on positively charged residues which are not likely to have been abiotically available (Tagami and Li, 2023). However, Giacobelli et al. recently showed that negatively charged abiotic peptides can interact with RNA through Mg^{2+} ion bridging between glutamic acid residues and the phosphate backbone, implying that metals can help stabilise abiotic peptide interactions with RNA (Giacobelli *et al.*, 2022). Importantly, any peptides or proteins that facilitated this pathway would have been uncoded, thus no record of their content or capabilities would have been preserved within the genome, leaving just their amino acid content and potential solvent environments as tools to determine their functional potential (Bernhardt, 2012).

Another significant theory is the “metabolism-first” hypothesis, which proposes that metabolic networks emerged before genetic information systems (Stewart, 2019). This model suggests that simple organic molecules could have

formed self-sustaining, energy-transducing cycles which provided the catalytic pathways necessary for the synthesis and accumulation of biological precursors (Tessera, 2018). These molecules would be capable of spontaneous synthesis and organisation without *a priori* need for nucleic acids, but this theory does not account for the emergence of conservative self-replication needed for complexity to emerge (Pross, 2004).

The “iron-sulfur world” theory emerged under the metabolism first hypothesis and argues that life originated at reducing hydrothermal vents on the ocean floor, where mineral catalysts such as iron and nickel sulfides could have facilitated the synthesis of organic molecules such as amino acids under high-pressure and high-temperature conditions (Wächtershäuser, 1992). These environments would have provided rich chemical gradients and energy sources capable of driving prebiotic chemistry, supporting the emergence of metabolic processes and the molecular synthesis at mineral-solvent interfaces. Follow on theories by William Martin and Michael Russell propose that microenvironments containing iron sulfide minerals within alkaline hydrothermal vents provide nearly all the required conditions for abiogenesis (Martin and Russell, 2003)(Martin and Russell, 2006). Recent data suggest that amino acid synthesis proceeds readily in these high-pressure, high-temperature conditions (Barge *et al.*, 2019).

An alternative theory, the lipid world hypothesis, focuses on the potential role of lipid molecules in the origin of life, proposing that the spontaneous formation of lipid vesicles could have created compartmentalised structures (Segré *et al.*, 2001). In addition to providing a mechanism to protect early organic species from harsh external conditions, lipid-based cell-like compartments could

concentrate organic molecules, enhancing chemical interactions between amino acids and nucleotides and enabling the co-development of metabolic and genetic systems. Much like the RNA world theory, a main criticism of this hypothesis is the synthetic complexity of lipids and the corresponding question of whether they could abiotically localise in sufficient concentrations to allow for compartmentalisation (Santos and Futerman, 2023).

Regardless of the specific pathway through which life emerged (RNA-centric, metabolism-first, or lipid-based), amino acids and peptides are highly likely to have played a key role. Their versatility, chemical robustness, functional potential, and abiotic abundance likely make them an indispensable component of abiogenesis (Orgel, 1998). Understanding the functional limitations of abiotic peptides and proteins is therefore a universally relevant metric for assessing planetary habitability.

Viable routes for organic synthesis

This section will review key research developments over the past decades which have shaped our understanding of the organic inventories of Earth and other celestial bodies. This overview creates the context for which amino acids are used throughout my study of abiotic-like proteins. This will also demonstrate why the presence of amino acids, and their polymers is plausible, and indeed observed, across a swath of abiotic environments, making abiotic protein function a useful and relevant tool for assessing planetary habitability.

Amino acids

We have a very limited ability to directly sample other planetary locations for the presence of nascent life or its biological precursors due to the time and expense required. Astrobiologists and organic chemists instead focus on defining and refining the reaction conditions needed to generate simple organic species such as sugars, nucleobases, and amino acids, which can be polymerised to form more complex structures. Given the wide scope of this subject area and the focus of this thesis, I will concentrate here on viable synthesis routes for amino acids.

Efforts to generate amino acids from simple and cosmochemically abundant feedstocks can be traced back to Adolph Strecker's 'Strecker synthesis' in the 1850s (Strecker, 1850). Strecker synthesis produces alpha-amino acids from the reaction of an aldehyde with hydrogen cyanide (HCN) and NH_3. Increasing interest in the origin of life following explorations of the Moon and Mars since the 1960s saw renewed efforts to identify other mechanisms through which amino acids could be generated (Marshall, 1994)(Bernstein *et al.*, 2002).

The best-known of these studies is the Miller-Urey experiment, which attempted to simulate early Earth conditions by discharging sparks through a mixture of gases thought to represent the primitive atmosphere (Miller, S., 1953). The experiment used a closed system containing a reservoir of heated alkaline water, methane (CH_4), ammonia (NH_3), and hydrogen (H_2) subjected to continuous electrical sparks (Miller and Urey, 1959). This experiment resulted in the formation of a range of organic products later revealed to contain over forty amino acids (Bada, 2013)(Sagan and Khare, 1971). The Miller-Urey experiment proved that amino acids could be synthesised under conditions believed to

resemble those of early Earth. While unappreciated at the time, the critical catalyst in this reaction was the borosilicate glassware surface (Criado-Reyes *et al.*, 2021). As will be discussed at length in Chapter 4, silicate surfaces provide catalytic sites that facilitate amino acid synthesis and peptide ligation due to their silanol moieties, stressed defects, and dangling bonds (Rimola *et al.*, 2016). Silanol moieties, (Si-O-H), are formed at the liquid interface of silicon oxides by protonation of the terminal Si-O group and become more prevalent at high pH, offering an additional pathway for the concentration and reactions of organic compounds (Rimola *et al.*, 2013).

Later investigators criticised the original reaction conditions used by Miller for being overly reducing compared to modern predictions of early Earth's atmospheric composition, which suggest a more neutral or weakly reducing atmosphere dominated by nitrogen (N_2) and carbon dioxide (CO_2)(Kasting, 1993). More recent experiments substituting H_2 with N_2 remained capable of generating significant quantities of amino acids (Cleaves *et al.*, 2008)(Janda *et al.*, 2008), confirming that the reaction pathway remains viable even in less reducing conditions. Further work has suggested that highly reducing conditions could have been locally present on ancient Earth, particularly in volcanic regions or following the impact events which were common during the Eoarchean era (4-3.6Ga), when life is believed to have originated at least 4Ga (Wogan *et al.*, 2023)(Moody *et al.*, 2024).

The synthesis pathways proposed by Strecker are part of a family of reactions now known as aqueous alteration pathways, whereby amino acids are generated within hydrated environments (Kebukawa *et al.*, 2017). It is believed

that aqueous alteration processes occurred within planetesimal bodies before their breakup and ejection, eventually formed asteroids and subsequently delivering the amino acids since found in meteorites on Earth. A range of possible reactions drive these processes, including Fischer-Tropsch-type synthesis, hydrothermal serpentinisation, and impact shock synthesis, all of which provide viable routes to generate amino acids in hydrated environments (Kebukawa *et al.*, 2017)(Cronin *et al.*, 1995).

The various aqueous alteration synthesis routes for amino acid synthesis have been extensively reviewed elsewhere (Kitadai and Maruyama, 2018), but of particular relevance to my study is the role of hydrothermal vents in driving synthesis reactions. Hydrothermal vents provide unique environments where abiotic synthesis of organic molecules can occur under high temperature and pressure (Amend and Shock, 1998)(Martin and Russell, 2006)(Martin *et al.*, 2008)(Hazen and Sverjensky, 2010)(Robinson *et al.*, 2021)(Kitadai, 2015)(Takahagi *et al.*, 2019). The thermal, redox, and pH gradients introduced between the vent fluids and the surrounding seawater create reaction conditions that facilitate the formation of organic compounds (Higgs and Pudritz, 2009)(Harrison *et al.*, 2023). For example, NH_4OH, HCHO, NaCN, and H_2, catalysed by calcium carbide, can react to produce amino acids and aromatic compounds (Marshall, 1994). Production of complex organic molecules in a hydrothermal vent system is well documented in modern vents within the Atlantis massif in the Atlantic Ocean, where non-biogenic aromatics have been detected from core samples within the hydrothermal systems (Jiang *et al.*, 2014).

On Earth, there are two major classes of hydrothermal vents, black smokers, which are typically hot (approximately 350°C), highly acidic (pH 2-5), and enriched in iron sulfides, and off ridge axis white smokers, which are more alkaline (pH 9-11) and cooler (50-90°C) (Martin *et al.*, 2008). While environments on planetesimals and putative hydrothermal systems like those proposed on Enceladus and Europa are likely to vary somewhat from those on Earth, similarities have been proposed (Deamer and Damer, 2017)(Tjoa *et al.*, 2020). In the case of Enceladus, the predicted hydrothermal vents are expected to be more similar to Earth's white smokers due to the silicates observed in the moon's hydrothermal ejecta, which are consistent with cooler (around 90°C), alkaline, reducing hydrothermal systems similar to those in Earth's Lost City hydrothermal field (Hsu *et al.*, 2015). Studies exploring the potential for amino acid synthesis in these sites suggest that the synthesis and ligation of simple amino acids are viable (Steel *et al.*, 2017)(Takahagi *et al.*, 2019).

Ultraviolet (UV) irradiation and cosmic rays can also drive the synthesis of amino acids in interstellar ice analogues. In replicative experiments, UV irradiated ice mixtures containing water, methanol (CH_3OH), acetonitrile (CH_3CN), and ammonia (NH_3) have demonstrated the formation of glycine under low-pressure conditions simulating interstellar conditions (Bernstein *et al.*, 2002). UV circularly polarised light can also induce enantioselective photolysis, preferentially degrading one enantiomer over the other, leading to an excess of L-amino acids (Sharma, 2022). Selective photolysis supports the enantiomeric excess of L-amino acids observed in certain meteorites (Glavin *et al.*, 2020)(Pizzarello *et al.*, 2008), suggesting that reactions in space may have contributed to the

homochirality of biological molecules. However, it has also been suggested these excesses are the result of terrestrial biological contamination (Pizzarello and Shock, 2010). Whether life's stereospecificity was introduced at an initial, extra-terrestrial stage or developed later, during biological evolution, has not yet been determined (see Chapter 4).

Peptide synthesis

Two primary routes provide thermodynamic conditions favouring peptide bond synthesis. The first is uncatalysed high-temperature synthesis via direct nucleophilic attack between neighbouring amino acids, requiring high concentrations, temperatures exceeding 150°C, and dehydrating conditions (Marshall, 1994)(Speight, 2018)(Vieira *et al.*, 2020)(Rodriguez-Garcia *et al.*, 2015). The second involves catalysed bond formation, achieved by condensing agents that enhance amino acids susceptibility to nucleophilic attack or via environmental catalysts that facilitate alternative reaction pathways by increasing local concentrations (**Fig 1.03**) (Sibilska *et al.*, 2017)(Erastova *et al.*, 2017)(Lopes *et al.*, 2009).

In addition to being well-suited to generating amino acids, hydrothermal systems are also capable of promoting peptide bond synthesis between amino acids via surface-catalysed peptide synthesis (Lemke *et al.*, 2009). Surface-catalysed synthesis was first proposed by Bernal and has since been documented under a wide range of reaction conditions and mineral surfaces (Bernal, 1949)(Lambert, 2008)(Kitadai *et al.*, 2017). Mineral surfaces can catalyse peptide synthesis by adsorbing amino acids, typically through electrostatic interactions, thereby significantly increasing local amino acids

concentrations and facilitating condensation reactions (**Fig 1.03**) (Shanker *et al.*, 2012)(Kitadai *et al.*, 2017). The efficiency of this process depends on the residence time of amino acids on the mineral surface, which is influenced by the strength of adsorption and surface properties (Lambert, 2008). Clays, silicates, and metal sulfides are particularly effective due to their high cosmochemical abundance and strong interactions with amino acids, which enhance local amino acid concentrations and thus synthesis rates on the surface (Cleaves *et al.*, 2009)(Lahav and White, 1980)(Hazen and Sverjensky, 2010). Abiotic peptide synthesis is also facilitated by deprotonation of the amino group to form the more nucleophilic NH_2 group, which occurs at alkaline pH (≈9.5) (Rodriguez-Garcia *et al.*, 2015)(Sakata *et al.*, 2010). Moderate heating further facilitates synthesis as peptide bond formation becomes favourable at temperatures above 60°C (Flegmann and Tattersall, 1979). However, excessive heating (above 120°C) can lead to side reactions, such as the formation of diketopiperazines through the dipeptide cyclisation reaction which can terminate the elongation of growing peptide chains (Shock, 1992).

Wet-dry cycles have been proposed as another efficient mechanism for peptide bond formation in prebiotic chemistry (Fleminger *et al.*, 2005). During dry phases, evaporation concentrates amino acids, promoting the condensation polymerisation pathways that form peptides and other biopolymers (Rodriguez-Garcia *et al.*, 2015). Subsequent wet phases allow the peptides to disperse and undergo further reactions, and repeated wet-dry cycles enhance peptide yield and length (Lahav and Chang, 1976)(Ross and Deamer, 2016). The likelihood of these geological niches is discussed further on page 52.

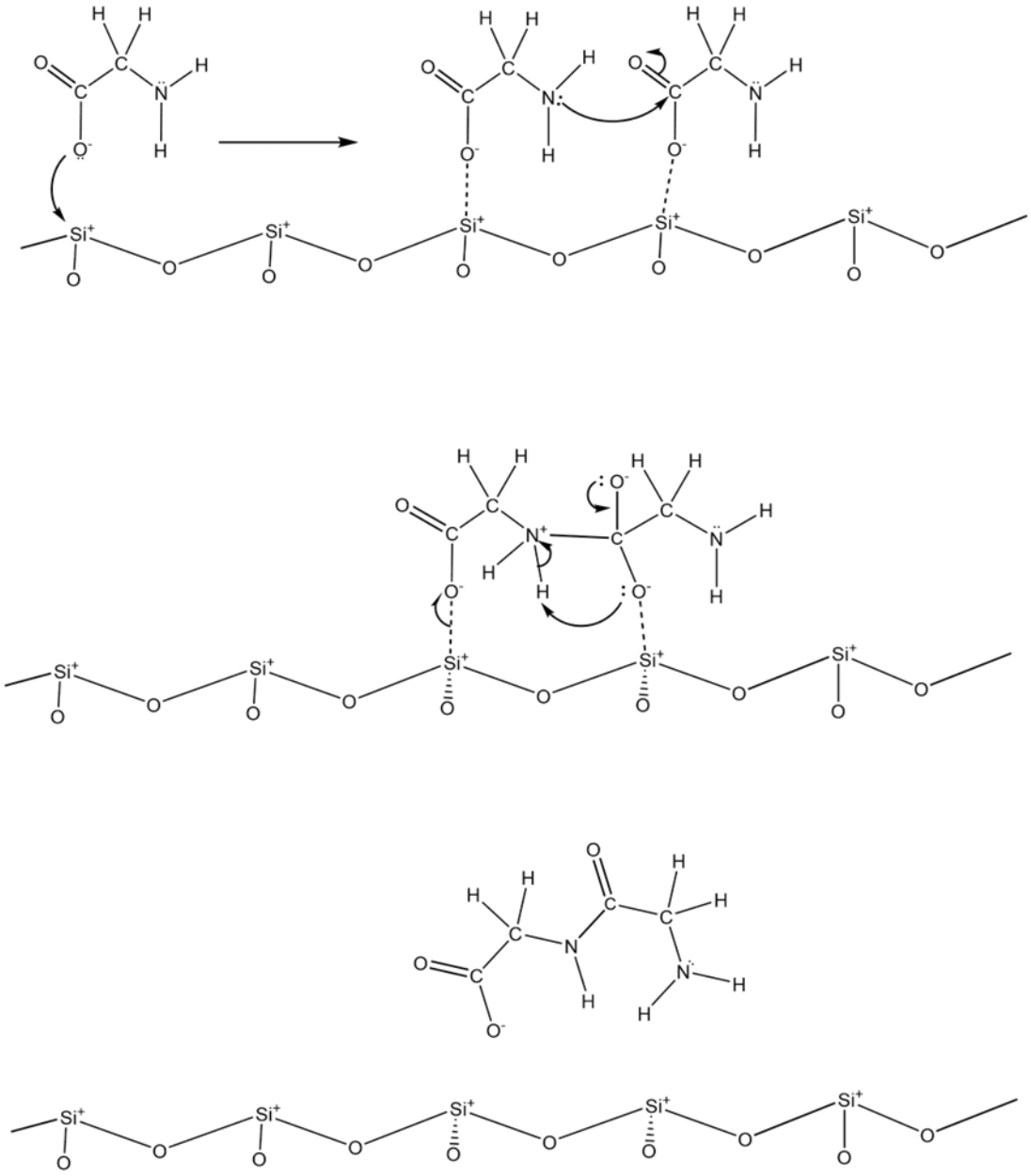


Figure 1.03 – Surface catalysed peptide synthesis

Simplified model of peptide synthesis under alkaline conditions. Adapted from (Takahagi et al., 2019). Silicate materials are shown without surface silanols for ease of understanding. In reality large hydrogen bond networks between surface silanols and water create the charge environment that drives C-O$^-$ interactions with the silicate surface

Another polymerisation technique often cited in abiogenic research is salt-induced peptide formation (SIPF), where high salt concentrations in excess of 3M NaCl or KCl and $CuCl_2$ produce low water activity levels that drive the binding and concentration of amino acids around Cu^{2+} ions at pH 3.0, catalysing peptide bond formation (Fitz *et al.*, 2011)(Rode *et al.*, 1999)(Le Son *et al.*, 1998). A <10% preference for L-amino acids has been observed for metal-mediated ligation in high salt conditions, but the mechanism is not identified (Plankensteiner *et al.*, 2004), While SIPF may have been feasible in certain environments on early Earth, such as evaporating tidal pools or hypersaline lagoons, its requirement for very high salinity and low pH, make it incompatible with biochemically relevant sites such as reducing, alkaline hydrothermal conditions.

While early Earth's exact environmental conditions are still debated, and there is intense debate about the viability of each of the models discussed, the existence of multiple synthesis routes increases the likelihood that amino acids can be ligated under abiotic conditions (Frenkel-Pinter *et al.*, 2020). The next subsection will review the research to confirm viability of the pathways discussed here through direct detection of their products throughout the solar system.

Definition of the solar systems organic inventory

Major sources of organic materials, carbonaceous asteroids, and comets, as well as aqueous environments like Europa and Enceladus have been examined for signs of amino acids or their precursors. While amino acids have been proposed to exist in interstellar space, evidence remains limited (Ioppolo *et al.*, 2021). I focus here on research within our solar system where significant evidence has corroborated the existence and make-up of the abiotic amino acid

set beyond Earth, carbonaceous asteroids, most recently from the NASA OSIRIS-REx and JAXA Hayabusa2 programs (**Fig 1.04**).

Some of the highest concentrations of carbon-based compounds are found on carbonaceous chondrite asteroids and their Earth-impacting meteorites (Kwok, 2016). These asteroids are relics from the earliest phases of planetary accretion and compositionally reflect the makeup of the solar planetary nebula. Carbonaceous chondrites, found in the present-day asteroid belt midway between the orbits of Mars and Jupiter, are a mix of metal and silicate dust grains that coalesced into small bodies which were too small to witness a significant degree of radiogenic heating. While the larger carbonaceous bodies present in the asteroid belt differ in composition, nearly all have a surface covered in carbon rich material. Importantly, the carbonaceous meteorites exhibit varying degrees of aqueous alteration (Pizzarello and Shock, 2010), and therefore possess hydrated minerals such as clays (Trigo-Rodríguez *et al.*, 2019).

The Murchison meteorite is one of the most studied carbonaceous meteorites, with over ninety-six different amino acids detected (Oba *et al.*, 2022), as well as purines and pyrimidines in the form of uracil and xanthine. The amino acid abundances of the Murchison and similar meteorites varies significantly, primarily attributed to sample heterogeneity (Elsila *et al.*, 2016), with detected amino acid concentrations typically ranging from 40 ppb in the Murchison meteorite (Glavin *et al.*, 2021), up to 5400 ppb in the Tagish Lake meteorite (Glavin *et al.*, 2012), and up to 1740 ppm proteogenic amino acids within the LAP 02342 CR2 meteorite (Pizzarello and Shock, 2010). By comparison, nucleotides are significantly less abundant, with the weighted average of purines in the

Murchison and Tagish meteorites ranging from 13 to 24 ppb for purines and 0.3 to 0.4 ppb for pyrimidines (Oba *et al.*, 2022).

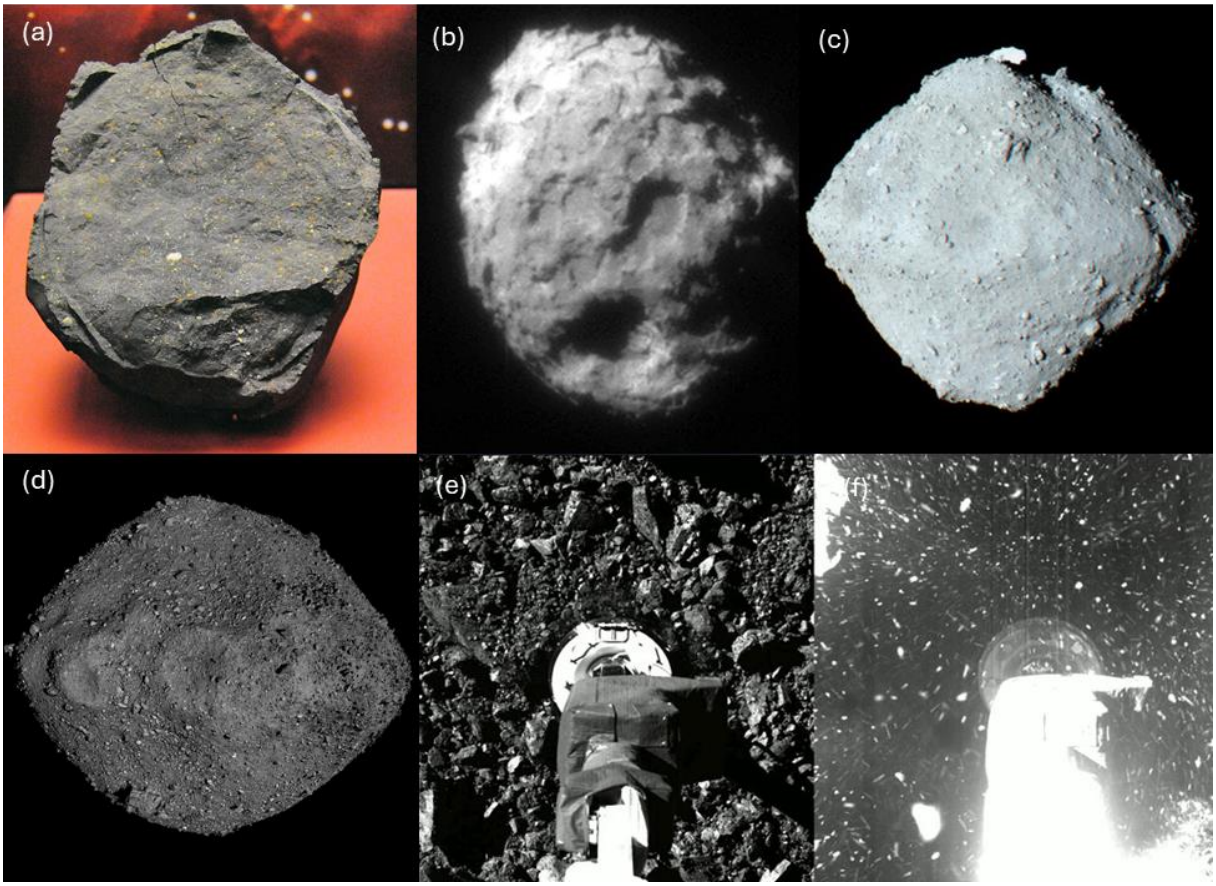


Figure 1.04 – Various missions have defined the solar systems abiotic amino acid inventory.

*Examples of previous research efforts to define the abiotic amino acid inventory. (**a**) Murchison meteorite fragment, housed at The National Museum of Natural History, Washington (**b**) The core of comet 81P/Wild, imaged by NASA Stardust. (**c**) True-colour image of asteroid 162173 Ryugu, captured by Hayabusa2 (JAXA). (**d**) A mosaic of asteroid 101955 Bennu, created using multiple observations by NASA OSIRIS-REx (NASA/Goddard/University of Arizona). (**e**) Pre-contact and (**f**) post-contact images of sample collection by OSIRIS-REx on Bennu on 20 October 2020, at landing site "Nightingale" (NASA/Goddard/University of Arizona).*

The abiotic amino acid set

The organic abundances of meteorites vary due to differences in the conditions of their parent bodies, the meteoritic composition of amino acids is broadly consistent, forming what is known as the abiotic amino acid set (Higgs and Pudritz, 2009). This set includes ten proteogenic amino acids: glycine, alanine, aspartic acid, glutamic acid, proline, serine, threonine, valine, isoleucine, and leucine, as well as three non-proteogenic amino acids: 2-aminoisobutyric acid, α-aminobutyric acid, and β-alanine (**Fig 1.01**) (Zaia *et al.*, 2008)(Higgs and Pudritz, 2009). Unsurprisingly, the abundance of these amino acids is correlated with their complexity, with glycine the most abundant and threonine the least abundant (Higgs and Pudritz, 2009). The composition of the proteogenic set is also supported by Trifonov, who created forty criteria to determine the amino acid codon incorporation chronology, which produced an amino acid genome incorporation order consistent with the abiotic amino set (G/A, V/D, P, S, E/L, T, R, N, K, Q, I, C, H, F, M, Y, W) (Trifonov, 2000).

Non-proteogenic amino acids are commonly excluded from abiotic synthesis studies, primarily due to the difficulty in incorporating them synthetically into proteins (Longo *et al.*, 2020), as well as the fact that none of these three amino acids offer significantly distinct functional groups (Zaia *et al.*, 2008). For these reasons, the non-proteogenic amino acids were not included in my experiments, although I acknowledge that this does introduce a potential limitation of the study.

The abiotic amino acid set is typically restricted to the above amino acids, however recent research has expanded the set to potentially include cysteine, a

sulfur-containing amino acid with significant functional importance in modern proteins. While cysteine is typically disregarded as an abiotic amino acid, Shalayel at al note that this stems from the lack of its detection in amino acid synthesis experiments which themselves do not typically include a source of sulfur, preventing cysteine from being synthesised (Shalayel *et al.*, 2020). They suggest that this has created an 'absence of evidence as evidence for its absence' feedback cycle that keeps cysteine excluded as an abiotic amino acid despite its relevance to proto-metabolic processes and similarity to abiotically available serine. While this would explain its absence in synthesis experiments, this does not explain its absence in meteoric samples, suggesting instead that cysteine requires specific aqueous conditions to support its spontaneous synthesis or conversion from another abiotic amino acid.

Foden *et al*. recently demonstrated a plausible prebiotic pathway for the abiotic synthesis of cysteine under mild conditions (Foden *et al.*, 2020). The key enabler for this reaction system is the acylation of the free amine group, which prevented degradation of cysteine nitrile during Strecker synthesis and enabled synthesis of cysteine from serine in the presence of sulfur and ferrocyanide. Synthesis proceeds with up to 95% yields at alkaline (pH 9.0) or neutral pH (pH 7.0) using NaSH or thioacetic acid as the thiolating agent, respectively. Notably the former thiol donator (NaSH), which is more likely available abiotically, is unstable below pH 7 and at temperatures in excess of its boiling point at 115°C. This suggests that this pathway is more consistent with conditions locally around an alkaline hydrothermal system than its black smoker acidic homologues where

sulfur containing small molecules typically react spontaneously to create H_2S and dissipate away.

The inclusion of cysteine in the abiotic amino acid set has significant implications for any environment compatible with its formation (Deng *et al.*, 2024). Cysteine's thiol side chain is significantly more nucleophilic than the carboxylate side chain of aspartate or glutamate, which is the only other anionic side chain class in the abiotic amino acids set (Brotzel and Mayr, 2007). To illustrate the scale of this difference, if the nucleophilicity of the aspartate side chain is compared to the kinetic impact of a small arms weapon, the nucleophilicity of the thiol side chain, which is $10^{9.61}$ more nucleophilic, is proportional to the kinetic explosive force of a tactical nuclear warhead. This increased nucleophilicity significantly enhances cysteine's ability to participate in catalytic processes and coordinate metal ions (Jordan *et al.*, 2021), greatly influencing protein folding, stability, and catalytic activity (Lukács *et al.*, 2021). Given that cysteine's availability as an abiotic amino acid is still a relatively recent proposal, in this thesis I will consider the abiotic-like peptides and proteins both with and without cysteine included, to monitor its impact on the potential functionality of these species.

Sampling missions of organic rich asteroids and comets

Concerns about contamination from Earth's microorganism in Earth impacting meteorites studied previously, necessitated direct sampling missions of pristine asteroids and comets to confidently define the organic species available for abiotic synthesis (Pizzarello and Shock, 2010). Comets and carbonaceous asteroids, proposed as significant sources of Earth's water and

organics, were the primary target for small body sample return missions. The first of these experiments was the NASA Stardust mission, which targeted Comet Wild 2 in January 2004 (Brownlee, 2014). Passing through the comet tail of Wild 2, it collected particulate matter in an aerogel which was subsequently returned to Earth. While the sample possessed a relatively low abundance of organic compounds, it did contain glycine, and was hailed as the first detection of a biological compound and tentative proof that the building blocks of life were synthesisable and likely available on early Earth prior to biological emergence (De Gregorio *et al.*, 2010)(Elsila *et al.*, 2009).

The most recent sample-seeking missions were the JAXA Hayabusa2 and the NASA OSIRIS-REx missions, which directly sampled asteroids 162173 Ryugu and 101955 Bennu, collecting pristine samples which were returned to Earth for analysis in December 2020 and September 2023, respectively. Subsequent analysis revealed an amino acid abundance between 8.3 ± 0.1 and 27.3 ± 1.9 ppm for Ryugu (Potiszil *et al.*, 2023) and 73 ppm for Bennu (Mojarro *et al.*, 2024), with 9/10 and 10/10 abiotic proteogenic amino acids identified. The lower-than-expected amino acid content for Ryugu was attributed to differences in Ryugu's parent body characteristics, which may have resulted in reduced amino acid synthesis rates (Parker *et al.*, 2023). The results from these pristine samples confirm beyond any reasonable doubt that the abiotic amino acid set is a valid subset of amino acids present in extraterrestrial materials and support the predicted mechanisms of spontaneous amino acid formation in abiotic niches. The proteogenic amino acid set is therefore a generalisable set of constraints for studies like ours. While it remains plausible that additional amino acids could form

in unique environments through modification of this set, the strong observational constraints on the ten abiotic amino acids make them the foundational set used in this thesis (Koga and Naraoka, 2017)(Foden *et al.*, 2020).

Folding nucleation of abiotic proteins and peptides by metal ions

Efforts to better understand the earliest stages of protein evolution have involved engineering modern proteins to resemble their hypothesised abiotic precursors. Substituting modern residues with those thought to be abiotically available produces “abiotic-like” proteins (Longo and Blaber, 2014). Studies on these simplified proteins revealed that these proteins often struggle to fold into stable, functional structures (Farías-Rico and Mourra-Díaz, 2022). This is consistent with observations from naturally compositionally depleted proteins, such as those with shortened sequences and reduced aromatic or thiol-containing residues, where the loss of stabilising elements leads to increased disorder (Romero *et al.*, 1999)(Murphy *et al.*, 2000)(Fan and Wang, 2003). The limited amount of previous research in this field demonstrates the need to identify and study naturally occurring peptide/protein systems that more closely mimic truly abiotic like sequences.

Taking a step back from these engineered systems, it is worth considering what the abiotic amino acid set alone implies *a priori* for early protein function. A core principle in biochemistry is the requirement for a protein to fold into a single or small number of unique conformations, even transiently, to perform its function (Anfinsen *et al.*, 1961). Terrestrial proteins, especially enzymes, have a 'sequence-structure-function' relationship, in which a defined amino acid sequence folds into a specific three-dimensional structure which enables protein

function. Thus, it is reasonable to expect that abiotically generated proteins must also fold to be functional within protometabolic processes, although they may do so in a less well-defined or permanent manner than is observed in modern proteins.

Typically, proteins fold into three-dimensional structures due to hydrophobic interactions that drive hydrophobic residues to pack together into the core of the protein, exposing hydrophilic amino acids to the solvent (Pace *et al.*, 2011). The inability to conserve a sequence-structure relationship abiotically is a common criticism of protein-based studies in this field, however there is evidence that some non-coded proteins and peptides can catalyse abiotic reactions (Tagami and Li, 2023).

Although the abiotic amino acid set includes hydrophobic amino acids including valine, leucine, and isoleucine, they are among the least abundant. According to ratios reported by Higgs and Pudritz, for every molecule of leucine detected, there are approximately eighty-six molecules of glycine (Higgs and Pudritz, 2009). In addition, hydrophobic folding mechanisms depend heavily on the precise placement and conservation of hydrophobic residues within the amino acid sequence. This requirement for carefully arranged, sequence-specific hydrophobic patterns make it less likely that early proteins could achieve stable folds through hydrophobic effects.

In modern proteins, metal ion binding sites are found in approximately one-third of known proteins, including many which bind nucleic acids and have enzyme active sites (Berg, 1990). Metal ion binding increases protein stability,

with metal ion binding motifs found amongst the oldest conserved protein fragments (Pozdnyakova and Wittung-Stafshede, 2001)(Alva *et al.*, 2015). Even small peptides composed of simple amino acids can chelate metal ions (Kozłowski *et al.*, 1999). The redox properties of certain transition metals such iron, nickel, and cobalt, also offer catalytic properties that may prove essential to early metabolic activity (Bromberg *et al.*, 2022).

Coordination to a metal ion or ligand is also a commonly observed alternative nucleating factor for modern protein folding. Metal-driven folding motifs, such as the CXXC zinc and iron binding sequences, are more sequence-agnostic than the sequences used to create hydrophobic cores, as evidenced by the high conservation of hydrophobic amino acids in modern proteins. In contract, as long as cysteines residues are appropriately spaced, cysteine can bind metals and promote folding with less stringent constraints on the surrounding amino acids (Lukács *et al.*, 2021). Placing this in the abiotic context, this would suggest that metal ion binding could have offered a more accessible and flexible route to stabilise early protein conformations into semi-structured domains capable of enabling catalytic function than hydrophobic core formation alone, assuming metal coordinating amino acids were available (**Fig 1.05**) (Pozdnyakova and Wittung-Stafshede, 2001)(Morleo *et al.*, 2010)(Dokmanić *et al.*, 2008). Further if we consider the high metal availability, in particular, Fe^{2+}, in the Archean oceans, this points to this pathway as a particularly relevant mechanism of ancient protein. folding that could pre-date hydrophobic based protein folding that is observed today. A requirement for metal binding in abiotic peptide folding would also suggest that soluble metal availability is a fundamental constraint for abiogenesis.

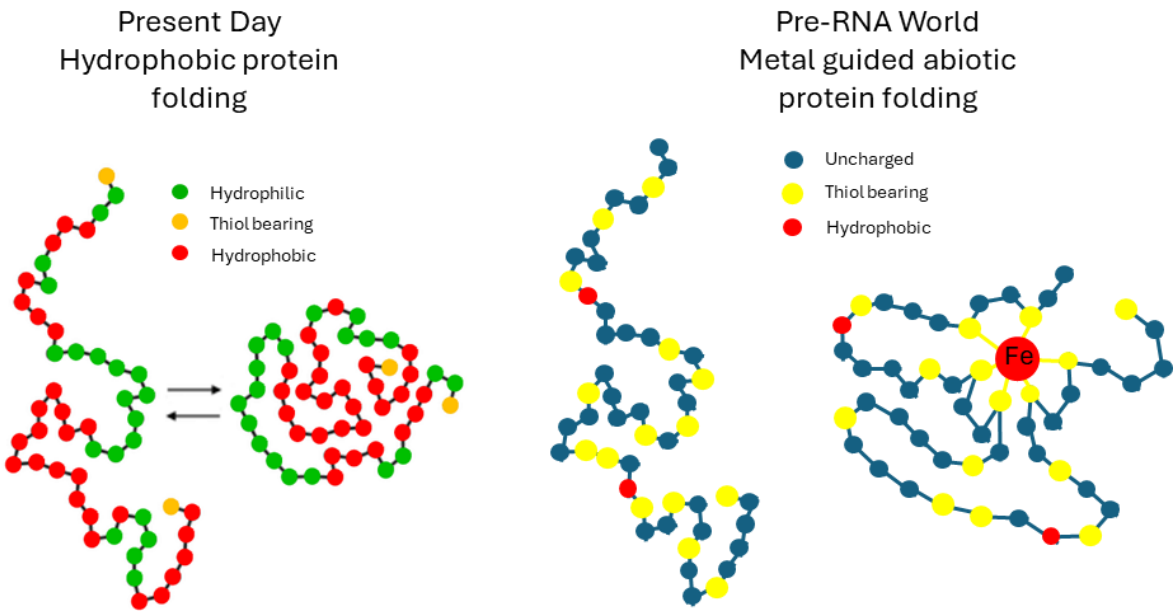


Figure 1.05 – Metal guided abiotic protein folding provides an alternate route for early proteins to fold around soluble metal ions.

Transition Metal Coordination: Influence of Ionic Size, Charge, and Crystal Field Effects

If metal coordination were a significant driver of ancient protein folding, it is worthwhile briefly considering why certain metals preferentially adopt specific coordination shell configurations and exhibit differing affinities to ligands. The biological ligands considered here are all fairly weak field ligands, so crystal field effects should be considered alongside electrostatic attraction and steric effects when understanding why particular metals favour specific coordination numbers and geometries. As proton number increases across the first-row transition series, increasing nuclear charge draws electrons closer to the nucleus. This leads to a consequent decrease in ionic radius and an increase in charge density. The result is an increased electrostatic attraction of negatively charged ligands to the metal.

First row transition metal cations with higher charge states, such as Fe^{3+}, display contracted ionic radii and hence are more charge-dense with greater electrostatic attraction to ligands. This favours the formation of octahedral complexes, where the steric cost of additional ligands is offset by the increased number of ligands contributing to favourable bond energies. In contrast, metals in lower oxidation states, such as Fe^{2+}, or those earlier in the transition series, have larger ionic radii and are less charge-dense. For Fe^{2+} there will be a small ligand field preference for octahedral, but for bulky moderate field strength ligands like thiolates, tetrahedral coordination is common. For the group 2 metals, Mg^{2+} and Ca^{2+}, there are no ligand field effects, and bonding is dominated by

electrostatic attraction, anionic carboxylate ligands are effective at stabilising these metals, and octahedral coordination is typical.

From an abiotic perspective, this chemistry imposes some notable potential constraints on plausible early metalloproteins. The abiotic abundance of aspartate suggests a tendency toward magnesium and calcium high-coordinate environments in early proteins, which are readily coordinated by carboxylate side chains if sufficient ligands are available. The likelihood of long peptide/peptide assemblages capable of possessing full octahedral shells suited to the coordination of highly charge dense metals like Mg^{2+} is likely quite low in non-biological settings unless there is a significant environmental excess of carboxylate side chain bearing amino acids. However, the prevalence of reduced transition metal ions under abiotic conditions, such as Fe^{2+}, could also be seen to favour lower coordination geometries if suitably soft, bulky ligands are available (i.e. thiol bearing amino acids). It is for this reason that I predict that in environments where short peptides predominate, such as those expected for abiotic synthesis, there will be a tendency to support transition metal complexes with lower coordination number requirements. Here, cysteine is therefore essential for stabilising transition metals in the earliest and shortest peptides. In contrast, the incorporation of transition elements into the earliest, shortest peptides may be hindered in aspartate based abiotic peptides, where geometric preferences for octahedral coordination requires longer proteins, and thus higher complexity to effectively position each ligand for coordination.

Abiotic solvent environments

It is essential to understand why the solution conditions I consider throughout this thesis matter to abiotic early peptide and protein function. Today, pH, salinity, temperature, and redox potential critically influence the stability, folding, and functionality of proteins, and these parameters are also likely to influence the partial folding of abiotic peptides and proteins (McDonald and Storrie-Lombardi, 2010)(Kocher and Dill, 2024). Here I consider each of these parameters in turn to justify their consideration throughout this work.

Solution pH determines the ionisation states of amino acid functional groups, both in solution and when incorporated into peptides, thereby influencing protein folding and catalytic activity (Talley and Alexov, 2010). For instance, the side chains of acidic residues like aspartic acid and glutamic acid have pK_a values of approximately 3.9 and 4.1, respectively, becoming negatively charged above these pH values. The thiol group of cysteine has a pK_a of about 8.3. Thus, at pH values above 8.3, it is deprotonated, forming the thiolate anion that is highly nucleophilic (Brotzel and Mayr, 2007). Notably, basic amino acids are absent abiotically, and cysteine has only recently been proposed as abiotically viable, as previously discussed. Changes in solution pH thus can drive changes in the ionisation states of these side chains and can, in the case of salt bridges between opposite charges, either enable or disrupt protein structure if the charges on amino acids change in response to changes in solution pH.

Salinity, defined by the concentration of dissolved salts, influences the ionic strength of a solvent. High salinity can lead to the 'salting out' of proteins, causing them to precipitate or aggregate due to decreased solubility (Matsue *et*

al., 2001). For example, ammonium sulfate at concentrations above 1M is commonly used to precipitate proteins from solution. Conversely, low salinity conditions may not provide sufficient ionic shielding to stabilise charged interactions within proteins, leading to disruption of these interactions and destabilisation of the global structures (Duan and Wang, 2024). The optimal ionic strength for protein stability can vary but typically reflects the moderate salt concentrations found in biological organisms around 0.1-0.2 M NaCl. Notably halophilic proteins produced by halophiles adapted to hypersaline conditions can be stable at up to 5M NaCl. These proteins are often dominated by the acidic amino acid's aspartate and glutamate (DasSarma and DasSarma, 2015). The abundance of these amino acids in abiotic conditions has led previous authors to propose that early proteins were halophilic (Longo *et al.*, 2013).

Temperature impacts both the kinetic energy of molecules and the thermodynamics of protein folding. Elevated temperatures can increase catalytic reaction rates, however, higher temperatures above the mesophilic range (40°C) often denature proteins due to loss of stabilising interactions such as hydrogen bond networks and electrostatic bridges (Miyawaki *et al.*, 2016). Some proteins, from thermophilic organisms are adapted to remain stable at high temperatures, but these proteins often require significant conserved sequence adaptations (Sterner and Liebl, 2001)(Reed *et al.*, 2013).

While redox potential has relatively little impact on the stability and function of amino acids, baring cysteine which readily oxidises to cystine above -220mV, redox potential has a significant impact on the availability of metal ions. In reducing environments, metals like Fe^{2+} and Mn^{2+} remain soluble and available

for incorporation into metalloproteins, providing an avenue for assisting protein folding and enabling catalytic function through the fixation of metal ions within protein structures. Conversely, in modern proteins oxidising conditions can promote the formation of covalent disulfide bonds between cysteine residues that can stabilise protein structures, but oxidising conditions also leads to the precipitation of transition metal ions as insoluble oxides or hydroxides, thereby limiting their availability. Therefore, if the abiotic amino acid set does not contain cysteine, the main impact of redox potential on abiotic peptides and proteins will be for metal ion availability, however if cysteine is a member of the abiotic amino acid set, this introduces a new consideration for the stability of abiotic peptides whose function will be influenced by the potential oxidation of the cysteine sulfur group.

In summary, the pH, salinity, temperature, and redox potential defines the major solvent parameters that define chemical landscape in which abiotic peptides and proteins would function. Examining the impact of these parameters on abiotic peptides and protein is expected to help predict the functional limits of proteins in astrobiological environments. The conditions discussed here are summarised in **Table 1.01**.

Environment	Temperature (°C)	Salinity (M NaCl eq.)	pH Range	Redox Potential (mV)
Early Earth Alkaline Hydrothermal Vents	70-90[1,2,3]	0.5[4,5]	9-11[1,3,5]	-400 to -300[2,6,10]
Early Earth Bulk Ocean	40-70[7]	0.5[4]	6-7[8,9]	-330[10]
Early Earth Surface Pools	50-80[11,12]	0.5-3[13,14]	6-8[8,9,15]	-100 to 0[16]
Early Mars Transient Water Bodies	-20-20[17]	0.5-3[18]	5.0-6.5[19,20]	0 to +300[21,22]
Enceladus Alkaline Hydrothermal Vents	90[23,24,25,31]	0.1[26,27]	8.5-10.5[24,27,28]	-350[3,31]
Enceladus Bulk Ocean	0-5[28,29]	0.1[28,29]	8.5-10.5[26,27,28]	-200 to 0[29,30]

Table 1.01 – Summary of abiotic solvent conditions described.
[1]Martin and Russell (2006), [2]Martin et al. (2008), [3]Barge et al. (2019), [4]Knauth (2005), [5]Kitadai and Maruyama (2018), [6]Martin and Russell (2003), [7]Kasting and Howard (2006), [8]Grotzinger and Kasting (1993), [9]Krissansen-Totton et al. (2022), [10]Catling and Zahnle (2020), [11]Ross and Deamer (2016), [12]Lahav and Chang (1976), [13]Fitz et al. (2011), [14]Rode et al. (1999), [15]Bernhardt (2012), [16]Russell (2021), [17]Wordsworth (2016), [18]Tosca et al. (2008), [19]Fairén et al. (2004), [20]Hurowitz and McLennan (2007), [21]Wade et al. (2021), [22]Hecht et al. (2009), [23]Hsu et al. (2015), [24]Choblet et al. (2017), [25]Sekine et al. (2015), [26]Postberg et al. (2018), [27]Glein and Waite (2020), [28]Waite et al. (2017), [29]Liu, C. et al. (2024), [30]Glein et al. (2015), [31]Kang et al. (2022).

AlphaFold in Protein Structure Prediction

Before considering the abiotic solvent conditions, it is worth briefly introducing Alphafold, a tool which became first available at the beginning of this work, and whose output I exploited to explore viable structural configurations for abiotic like proteins. The introduction of AlphaFold2 marked a significant advance in protein structure prediction, achieving high accuracy protein structure predictions based on a user provided amino acid sequence for a wide range of naturally occurring protein folds. Its approach relies on deep learning algorithms trained on large databases of experimentally determined protein structures from the PDB and beyond, enabling prediction of backbone and sidechain conformations (Jumper *et al.*, 2021). However, AlphaFold2 has recognised limitations: it does not explicitly model the presence of bound cofactors, metal centres, or non-canonical amino acids. Instead, predictions are optimised for the polypeptide chain alone, with no direct modelling of ligands unless their binding is tightly conserved and implicit in the training set. As a result, AlphaFold2 often fails to accurately reproduce binding site architecture or sidechain positions for metalloproteins. This limitation is exacerbated when predicting the structure of engineered or highly divergent sequences not well represented in existing databases.

AlphaFold3, released during the latter stages of this project, introduces a broader framework for multi-modal prediction, aiming to model protein-ligand, protein-nucleic acid, and protein-protein interactions, including metal coordination sites (Abramson *et al.*, 2024). While AlphaFold3 improves the capacity to predict structures with ligands, it still relies heavily on precedent within its training data.

Novel metal binding sites, especially those with unusual geometry or non-biological sequence contexts, remain a significant challenge. For engineered or consensus protein sequences, such as those constructed in this work, both AlphaFold2 and AlphaFold3 can produce models that appear to be fully folded but are not experimentally valid. This remains the case for Alphafold3 structures, where metal ions interactions are not explicitly modelled through molecular dynamics during structure prediction, making their predictions idealised representations rather than accurate renditions.

Proposed Archean Ocean composition

Having established the importance of pH, salinity, temperature, and redox potential to abiotic peptide and protein function, I now turn to the specific solvent environments of early planetary systems. These provide the context within which the functional limits of abiotic proteins must be evaluated.

The earliest potential evidence for life on Earth dates to 3.7 Ga from ^{13}C depleted biogenically derived graphite (Ohtomo *et al.*, 2014). The earliest fossil evidence from stromatolites dated to 3.5 billion years old from the Dresser Formation in Western Australia (Van Kranendonk *et al.*, 2008) and microfossils from the Strelley Pool Chert, which are estimated at 3.4 billion years old (Allwood *et al.*, 2007). The origin of life is thought to have occurred sometime between 4.2 and 3.8 Ga (Knoll and Nowak, 2017), during an overlapping period between the Hadean (4.6-4.0 Ga) and the Archean (4-2.5 Ga) eras. Thus, life was likely first appearing and acquiring the biological frameworks of modern life during the Eoarchean era, 4.0-3.6 billion years ago (Knoll and Nowak, 2017).

Geochemical analyses suggest that the Archean oceans and atmosphere were markedly different from those of today, with lower pH levels, reducing conditions, moderate salinity, and high temperature variance (Catling and Zahnle, 2020). Analysis of ancient marine sediments and carbon cycle modelling indicates that the Archean oceans were mildly acidic to neutral, with pH estimates ranging from 6.0 to 7.0 (Grotzinger and Kasting, 1993)(Krissansen-Totton Joshua *et al.*, 2022) compared to the modern ocean's pH of 8.1 (average). Elevated levels of CO_2 would have increased the dissolution of carbonic acid into ocean waters, acidifying the bulk ocean pH.

The anoxic and chemically reducing nature of the Archean oceans reflect their equilibrium with an atmosphere comprised of volcanic gases emanating from Earth's mantle (Hao et al., 2019). Negligible partial pressure of oxygen (O_2) and mildly acidic pH promoted the solubility of transition metals (Tostevin and Ahmed, 2023). The bulk ocean redox potential predicted from concentrations of CH_4 and H_2 is ≈ -330mV (Catling and Zahnle, 2020). Iron was significantly more abundant in its ferrous (Fe^{2+}) than ferric (Fe^{3+}) state (>105 times present Fe^{2+} concentration), leading to high concentrations of dissolved iron in the oceans, which likely made the early oceans appear green (Holland, 2006)(Moore *et al.*, 2017).

Although harder to determine from sparse geological data, salinity levels in the Archean oceans are thought to have been comparable to those of modern oceans (Knauth, 2005). However, sub-ocean brine pools could have created extremely hypersaline niches with up to 3M NaCl that favoured pathways such as SIPF to enable early di- and tri-peptide synthesis (Fitz *et al.*, 2011)

Temperature estimates for the Archean oceans vary. It is generally accepted that they were warmer than today's oceans but the evidence to support these estimates is relatively weak (Knauth, 2005)(Tostevin and Ahmed, 2023). Studies based on atmospheric modelling typically report ranges from 40°C to 70°C for the bulk ocean, where temperature was driven up by a denser atmosphere and increased volcanic activity despite a fainter young sun (Kasting and Howard, 2006). These elevated temperatures could increase reaction rates but also pose challenges for organic stability due to increased hydrolysis rates (Kitadai and Maruyama, 2018).

Hydrothermal vent systems were likely widespread along mid-ocean ridges. These vents emitted hot, mineral-rich fluids into the ocean, creating localised environments with steep chemical and thermal gradients (Tosca and Tutolo, 2023)(Martin *et al.*, 2008). Alkaline hydrothermal vents, such as those analogous to the modern Lost City Hydrothermal Field on the Atlantis Massif, provide environments with temperatures of 70°C to 90°C, high pH (9-11), and abundant hydrogen (H_2) and methane (CH_4), potentially facilitating prebiotic chemical reactions like those proposed by Martin *et al* for the synthesis of amino acids and aromatics (Martin and Russell, 2006). While transition metal availability within alkaline conditions such as these is typically poor, on early Earth this alkaline hydrothermal fluid flowed into an acidic ocean. This would create interfaces at the ocean/mineral boundaries at hydrothermal vent where organics synthesised under alkaline conditions flowed into an ocean where transition metal availability was much higher.

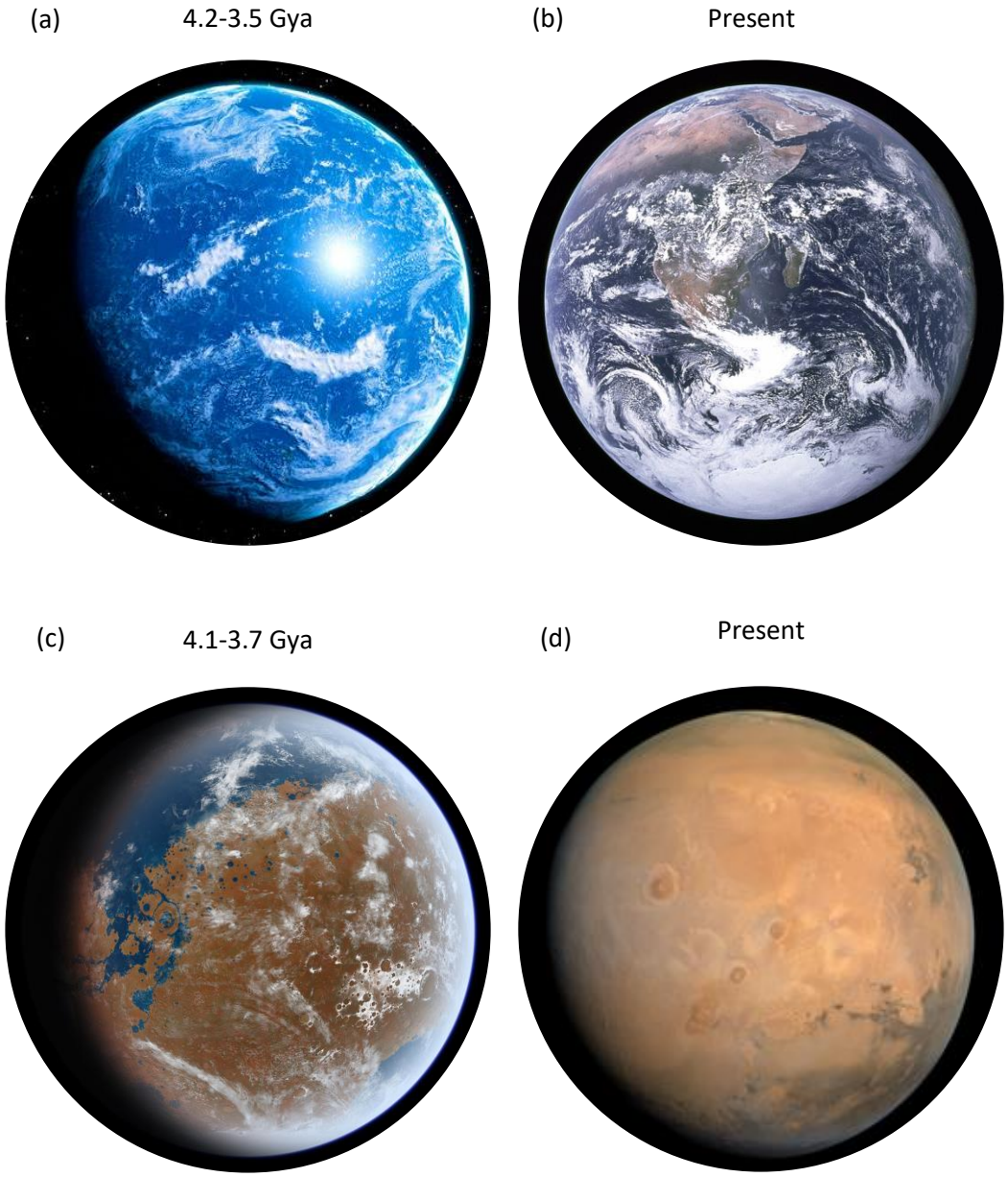


Figure 1.06 – Liquid water was present on Earth and Mars up to 4.1 billion years ago.

*Liquid water was available on Mars and Earth during the Hadean/Eoarchean periods. (**a**) an artist's impression of early Earth as an ocean world (Mark Garlick, Science Photo Library). (**b**) Earth, as observed during the Apollo 17 mission (NASA). (**c**) an artist's impression of early Mars, based on northern global ocean theories (Ittiz, Wikimedia). (**d**) a true-colour image of Mars captured by the Hope Orbiter (UAESA)*

The availability of dry land during the Archean era remains a subject of debate in large part due to the increasing sparsity of the geological record (**Fig 1.06**). Evidence from zircons dated to about 4.4 billion years ago suggests the presence of continental crust (Wilde *et al.*, 2001)(Gamaleldien *et al.*, 2024). However, other research, which notes that zircons are not representative samples for determining conditions on early Earth indicate that emergent land would have been limited due to the inability of the significantly softer mantle to support land structures (Monteux *et al.*, 2020)(Flament *et al.*, 2008)(Lahav and Chang, 1976)(Kamber, 2015). Limited landmass would reduce the likelihood of terrestrial environments for facilitating dehydration synthesis necessary for polymer formation, necessitating that all abiotic pathways proceed in marine environments. I note that while this may appear to be quite definitive evidence against the existence of dry land and the wet/drying cycles they enable, this remains an actively contested area of research (Bernhardt, 2012)(Russell, 2021).

Proposed noachian Mars Ocean compositions

Mars, while lacking surface liquid water today, exhibits extensive geological evidence suggesting that it hosted significant amounts of liquid water during the Noachian period, approximately 4.1 to 3.7 billion years ago. The quantity and accessibility of this water are under debate, some authors have proposed that an ocean covered up to one-third of the planet's surface, while separate analysis suggests that any bodies of water were more transient flows from thawing surface ice (Clifford and Parker, 2001)(Carr and Head, 2003)(Di Achille and Hynek, 2010)(Hynek *et al.*, 2010).

Early Martian oceans are estimated to have been more acidic than those of early Earth, with pH ranging from 5.0 to 6.5 (Fairén *et al.*, 2004)(Hurowitz and McLennan, 2007). This acidity is attributed to the abundance of sulfur dioxide (SO_2) and carbon dioxide in the atmosphere (Kerber *et al.*, 2015). The prevalence of sulfate minerals detected on the Martian surface, including jarosite and gypsum identified by the Opportunity rover, and the absence of carbonate deposits is proposed as key supporting evidence for an acidic ocean (Squyres *et al.*, 2004).

Conditions on early Mars were initially more reducing than early Earth, but Mars underwent a significant oxidation event early in its planetary evolution that drove processes such as iron oxidation (Liu, J. *et al.*, 2024)(Tuff *et al.*, 2013). These oxidation events were primarily driven by water-rock interactions and indicate that metals such as iron likely existed in the insoluble Fe^{3+} state, rather than its more soluble and bioavailable Fe^{2+} ion (Wade *et al.*, 2021)(Hecht *et al.*, 2009)

Salinity levels in early Martian oceans were likely high, driven by evaporation (Tosca *et al.*, 2010). Evaporite minerals such as sulfates and chlorides have been identified by orbital spectrometers and rovers such as Opportunity and Curiosity (Squyres *et al.*, 2004). These suggest that as the Martian climate evolved and water bodies evaporated, dissolved salts became increasingly concentrated, which would have forced any life present to become increasing halophilic (Tosca *et al.*, 2008).

Temperature estimates for early Mars indicate a relatively cold climate, with average surface temperatures possibly ranging from -20°C to 20°C

(Wordsworth, 2016). However, transient warming events caused by volcanic activity or large impacts may have created localised regions where liquid water was present (Segura *et al.*, 2008), which is supported by the presence of valley networks suggesting that liquid water flowed on the surface at least episodically (Di Achille and Hynek, 2010).

Organic molecule detection has been a focal point in assessing Mars' habitability. The Sample Analysis at Mars (SAM) instrument on Curiosity has identified chlorobenzene and dichloroalkanes indicating the preservation of organic carbon in ancient sediments (Freissinet *et al.*, 2015). Additionally, the delivery of organics via carbonaceous chondrites and interplanetary dust particles is estimated to contribute up to 2.4×10^5 kg of organic carbon annually (Flynn, 1996). However, more recent analysis suggests that Mars' lack of a significant magnetic field would result in the ultraviolet photolysis of any surface organic species within a 20-million-year window (Pavlov *et al.*, 2022).

Considering the solvent conditions, acidic pH, progressively increasing oxidising redox conditions, high salinity, and generally low temperatures with possible transient warming, early Martian oceans present challenges for the stability and formation of biomolecules. Acidic and oxidising environments often degrade organic molecules (Benner *et al.*, 2000), while low temperatures slow down the kinetics of chemical reactions (Pavlov *et al.*, 2022). In this overall inhospitable Martian landscape, localised hydrothermal environments or subsurface aquifers might have offered more favourable conditions for prebiotic chemistry (Sauterey *et al.*, 2022), although their contents remain mostly speculative.

Proposed Enceladian ocean compositions

While similar environments may exist on other moons within the solar system, such as Europa, Enceladus remains the best characterised extraterrestrial ocean world (**Fig 1.07**) (Glein and Waite, 2020). The sampling of the ocean ejecta from the Enceladian south pole geysers with the Cassini Ion and Neutral Mass Spectrometer (INMS) and Cosmic Dust Analyzer (CDA) from 2008 to 2015 enabled cursory characterisation, revealing an alkaline (pH 8.5-10.5), reducing (1-5% H_2), and carbonate-rich ocean (Postberg *et al.*, 2023)(Postberg *et al.*, 2018)(Waite *et al.*, 2017)(Glein *et al.*, 2015)(Glein and Waite, 2020). Beneath this ocean lies a proposedly rocky differentiated core estimated to be 185-205 km wide, composed primarily of silicate-based olivine and orthopyroxene, with a significant degree of permeability (Neumann and Kruse, 2019).

Analysis of the SiO_2 content in Saturn's E ring, which is sourced from the Enceladus ejecta, and direct sampling of plume content indicate hydrothermal activity generating significant amounts of H_2 (Hsu *et al.*, 2015)(Choblet *et al.*, 2017). The redox potential is predicted to be approximately -350mV based on the CO_2 + H_2 content within the ejecta (Liu, C. *et al.*, 2024). Enceladian hydrothermal systems are thought to be consistent with Earth's white smokers, with local temperatures estimated up to 90°C feeding into an ocean with a salinity range (0.1M NaCl) slightly lower than modern Earth (Sekine *et al.*, 2015)(Kang *et al.*, 2022). Given the alkaline, reducing conditions, the availability of biologically relevant metals such as iron, nickel, and zinc, is relatively low due to formation of insoluble precipitates under these conditions (Kitadai, 2015)(Hao *et al.*, 2022).

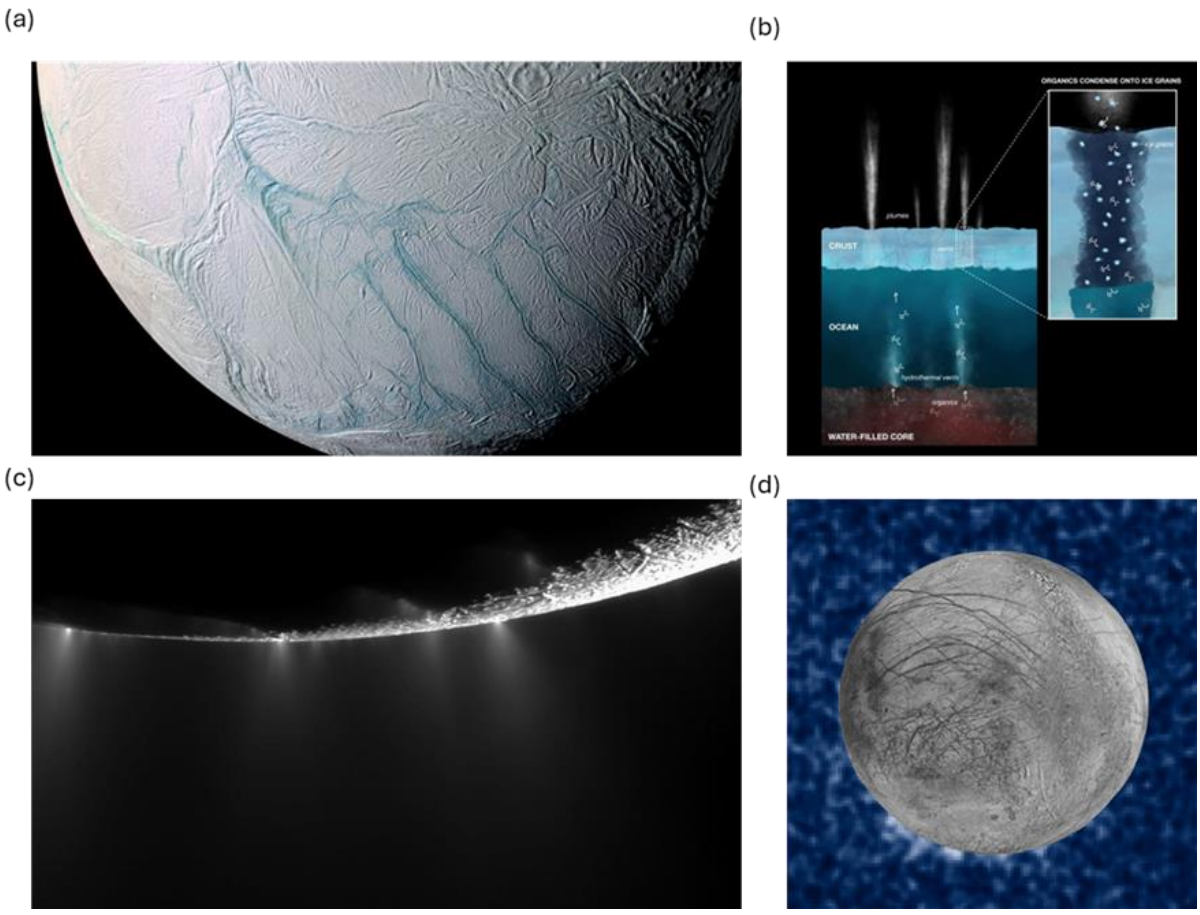


Figure 1.07 – Enceladus and Europa contain subsurface liquid oceans that are actively ejecting into space.

*Europa and Enceladus show strong evidence for the presence of a subsurface liquid ocean. (**a**) false-colour image of Enceladus's south pole region (NASA/ESA/JPL/SSI/Cassini). (**b**) a model illustrating how hydrothermal activity drives Enceladus's plume activity (NASA/JPL-Caltech). (**c**) a panorama of Enceladus' south pole taken by Cassini on November 21, 2009, show water plumes erupting into space (NASA/JPL-Caltech/SSI/CICLOPS/Kevin M. Gill). (**d**) a composite image showing suspected water vapor plumes erupting from Europa observed by the NASA Hubble Space Telescope. (NASA/ESA/W. Sparks [STScI]/USGS Astrogeology Science Center).*

Unlike on Earth, where alkaline hydrothermal fluid flowed into an acidic ocean, the high pH. of Enceladus' Ocean suggests that transition metal availability is low both in the hydrothermal vent fluid and the bulk ocean it flows into.

A fundamental issue in biogenesis common to all subsurface ocean worlds is that condensation polymerisation, needed to link precursor molecules such as amino acids and nucleotides, is kinetically and thermodynamically unfavourable in bulk oceans due to monomer dilution and high-water activity. However, condensation polymerisation does proceed readily at the water/mineral interfaces near thermally active hydrothermal systems (Takahagi *et al.*, 2019)(Cleaves *et al.*, 2009)(Kitadai and Maruyama, 2018)(Deamer and Damer, 2017). As such tidal friction and serpentinisation-powered hydrothermal systems at these sites may be the most chemically productive sites on Enceladus (Damer and Deamer, 2020). Reactions within these environments are likely responsible for the chemical species contained within the plume observed from space, including potential nitrogen and oxygen bearing macromolecules (Khawaja *et al.*, 2019)(Postberg *et al.*, 2018)(Stüeken *et al.*, 2013).

Predictions of protein function with abiotic solvents

The preceding sections have explored several proposed early planetary environments, from alkaline hydrothermal vents on early Earth feeding into a mildly acidic metal rich ocean, to the acidic, oxidising waters of early Mars and the alkaline, metal-poor ocean of Enceladus. I have also considered the abiotic amino acid set and its implications for peptide synthesis and potential early protein function. The diverse solvent conditions across Archean Earth, Mars, and Enceladus provide a range of scenarios for testing the functional limits of abiotic

peptides and proteins. Building on these, I now make a series of predictions regarding the likely function, stability, and folding behaviours of abiotically derived and early RNA coded peptides and small proteins to create testable predictions that will be investigated throughout this thesis. Importantly, these hypotheses should be noted here as predictions only until tested experimentally later in this work.

Hypothesis 1: Cysteine is essential to enable stable transition metal coordination in abiotic peptides and proteins.

In the absence of basic and aromatic residues, alternative folding nucleators for early peptides will likely become crucial. Metal ions, particularly transition metals, could serve this role if available in solution. The abiotic amino acids aspartate and glutamate contain carboxylates, which have the capacity to facilitate peptide binding to divalent cations such as magnesium or calcium similarly to some structures observed in modern life but are less able to coordinate the transition metals needed to enable the redox chemistry to drive the emergence of protometabolism such as iron and copper. Cysteine meanwhile provides a thiolate donor group, which stabilises low coordination number, redox-active metal centres in simple tetrahedral geometries. Carboxylate ligands, in contrast, are weak-field donors and typically stabilise high-spin, octahedral complexes best suited to hard cations like Mg^{2+} or Ca^{2+}.

Although cysteine-based coordination is the central focus of this study and forms a central hypothesis for testing here, it is important to briefly recognise that not all metal-binding sites in biology involve cysteine. Proteins such as calmodulin bind calcium exclusively through aspartate- and glutamate-rich motifs, while

siderophores employ carboxylate to coordinate iron. However, these modes of ligation are not generally suited to the formation of redox-active transition metals coordinating sites required for early metabolic processes.

Cysteine may therefore have provided a more suitable scaffold for abiotic metal binding, and a plausible scheme for the conversion of serine to cysteine has been proposed (Foden *et al.*, 2020). Such thiol-metal interactions could have stabilised early peptides, offering folding routes driven by the strength of those interactions rather than the hydrophobic effect. It is my prediction then that it is highly unlikely that stable, redox active transition metal coordination sites would form in the absence of cysteine, another sulfur-bearing amino acid, or another more chemically complex amino acid, like histidine, which can present a free nitrogen lone pair.

Hypothesis 2: Acidic pH at or near pH 4 to 5 will cause protonation of acidic residues and induce unfolding in abiotic peptides and proteins.

The abiotic amino acid set contains carboxylate sidechains with a pK_a ~4 in the monomeric state and potentially higher when incorporated into peptides and proteins. Thus, these are likely to be highly sensitive to pH in those ranges. In this range, the protonation of the carboxyl side chains of aspartate and glutamate may result in the disruption of any stabilising electrostatic interactions either in the protein or with water molecules in the protein hydration shell, resulting in protein destabilisation and/or decreased solubility, resulting in aggregation.

Hypothesis 3: Abiotic peptides and proteins adopt molten globule states rather than fully folded tertiary structures.

Modern proteins rely on hydrophobic residues to achieve stable tertiary folds. In contrast, abiotic proteins produced from the limited amino acid set and lacking a conserved sequence are unlikely to form stable hydrophobic cores While it has been reported that abiotic peptides and proteins could fold under high salt conditions, this result did not come from a protein composed exclusively of the consensus abiotic amino acid set and was stabilised by biotic positively charged sidechains. Even if coded for within early RNA catalysed peptide synthesis pathway, there is a low likelihood of the formation of stable core formation in early proteins due to the depletion of aromatic and the reduced side chain topologies that can optimise packing geometries. Similarly, Mauro *et al*, also predicted that abiotic proteins may only achieve states reminiscent of a "molten globule" (Di Mauro *et al.*, 2012). These structures are a partially folded conformation that retains some secondary structure but lacks the tightly packed tertiary interactions characteristic of fully folded proteins. Such molten globule states are flexible and offer limited specificity or catalytic capability without additional stabilising factors, such as metal ions, to impose some level of local structural order.

Hypothesis 4: Increasing salinity does not significantly alter abiotic peptide and protein conformation due to an absence of tertiary folding interactions.

High salinity environments can disrupt or stabilise charge-based interactions. Previous studies suggest that proteins composed predominantly of acidic residues can tolerate elevated salt conditions (Longo *et al.*, 2013).

However, without a robust hydrophobic core or complementary basic residues, these abiotic peptides may remain only marginally stable. As such, the limited structural rigidity predicted for these proteins makes it unlikely that high or low salinity conditions will significantly disrupt the structure given that there is not enough tertiary structure to disrupt. As such, although salinity shifts may have some effect on their secondary structure, the net effect is likely to be minimal compared to charge dependent effects predicted for pH.

Hypothesis 5: Reducing conditions are required for abiotic peptides and proteins to incorporate transition metals.

While having a relatively limited influence on modern protein folding, local redox potential determines the solubility and availability of transition metals in solution. If early peptide folding was metal assisted, then reducing environments would facilitate metal incorporation and enable peptide folding. Alternatively, under more oxidising regimes, such as those that developed on Mars, metals would have been less available in soluble forms, restricting metal-assisted folding. However, if thiol-containing amino acids were present, oxidation could foster disulfide-like linkages, to enable disulfide driven protein folding. This may stabilise peptide structures to create stable structures, however the type of folded generated would be extremely heterogeneous with a stabilising guided folding pathway like hydrophobic or metal induced protein folding (Anfinsen *et al.*, 1961).

Predictions on an abiotic peptide folding window:

Taken together these predictions suggest that white smoker like conditions feeding into an acidic ocean, with moderate salinity, alkaline pH, reducing redox states and accessible transition metals, would offer the best environment for

producing peptides that achieve at least transient and partially folded states. While transition metals typically utilised by modern life for redox chemistry (e.g. iron, manganese and copper) are usually not highly abundant under such conditions due to the formation of highly insoluble hydroxide precipitates, the presence of metal-coordinating cysteine peptides could enable spontaneous coordination and solubilisation. Thus, warm, reducing, alkaline and moderately saline conditions provide suitable conditions for abiotic peptide synthesis, folding and function.

Aim and Scope of the Thesis

This thesis aims to define the solvent conditions necessary to support the function of proteins composed exclusively of abiotic amino acids. To achieve this and test the predictions made, I will model abiotic-like proteins derived from them and use this to propose the minimal amino acid set and the solvent parameters required in terms of pH, temperature, redox potential, and salinity to support their function. This will create an abiotic protein functional window which I can compare current conditions on Archean Earth, Mars, and Enceladus to measure their compatibility with abiotic protein function and thus further define the range of conditions most likely to yield planetary habitability. While my primary experimental focus was on abiotic protein stability and function, the potential role of cysteine, a sulfur-containing amino acid, is heavily considered throughout.

Overarching Aims

1. Define the solvent environments conducive to abiotic protein folding and function. (Chapter 2-4)

2. Model abiotic-like proteins using exclusively abiotic amino acids, assessing their stability and folding behaviour. (Chapter 3)
3. Propose the minimal amino acid set required for protein folding and activity in abiotic conditions. (Chapter 2,4)
4. Consider the impact of including or excluding cysteine in abiotic protein systems, given the recent discovery of a plausible abiotic synthesis. (Chapter 2-4)

To achieve these aims, the thesis is structured as three interconnected studies, each addressing specific aspects of the hypothesis generated through a combination of experimental and computational approaches. In Chapter 2 I investigate the capability of an ancient metalloprotein to coordinate transition metals with and without cysteine (Aim 3-4, Hypothesis 1). In Chapter 3 I identify modern abiotic-like protein and investigate their folding dynamics in solution (Aim 2, Hypothesis 2-4). In Chapter 4 I investigate the capability of cysteine to catalyse peptide synthesis under alkaline reducing conditions as predicted (Aim 1, Hypothesis 5). Finally in Chapter 5 I integrate my findings and compare them against conditions observed reported for a range of potentially habitably niches as discussed throughout Chapter 1 (Aim 1,3). In Chapter 6 I then discuss future lines of investigation based on the results presented in this thesis.

Hypothesis Statement

Abiotic peptides and proteins composed solely of the predicted abiotic amino acid set can achieve partially folded, rudimentary functional states under moderate salinity, elevated temperature, alkaline, and reducing conditions. The inclusion of cysteine in this set enables more effective metal coordination and

catalysis, providing a molecular bridge toward increasingly complex early biochemistry.

Chapter 2: Elucidating the basis of abiotic metal ion coordination using an ancient metalloprotein model system.

Introduction

Study rationale

Building on the concepts discussed in Chapter 1, I begin by investigating my first hypothesis: *Cysteine is essential to enable stable transition metal coordination in abiotic proteins*. Previously, I reasoned that early and abiotic proteins would have struggled to fold spontaneously into globular structures due to the limited diversity and abundance of hydrophobic amino acids within the abiotic set, even if coded for by primitive RNA based systems. Among the most plausible solutions to this would be protein folding via metal ion coordination. This would create a route whereby disordered sequences formed three dimensional structures around metal ions that enabled subsequent function independently of hydrophobic interactions and with a less stringent requirement for amino acid sequence conservation (**Fig 1.05**).

To achieve this metal-assisted stability however, abiotic proteins must have been able to coordinate transition metals using only the anionic amino acids available in the abiotic set. Yet, in modern proteins, simple metal binding motifs typically rely on cysteine and its abiotic presence has only recently been proposed (Dokmanić *et al.*, 2008)(Foden *et al.*, 2020). This makes its presence far less certain than the currently accepted abiotic anionic amino acids, aspartate, and glutamate. In addition, the late incorporation of cysteine into the genetic code raises a fundamental question for the entire abiogenesis process: Could early proteins, composed solely of the abiotic amino acid set, coordinate transition metals without cysteine? If not, the emergence of protein folding via metal ion coordination and functional metalloproteins, would have been hindered unless

cysteine was available locally or until it eventually became available through biogenesis. This would also suggest that cysteine is a required amino acid for the early stages of abiogenesis, constraining the conditions within which life could emerge to sulfur rich, alkaline conditions given that thiolates are typically highly unstable under acidic conditions.

To test this hypothesis, I examine here whether the metal-binding roles performed by cysteine in modern metalloproteins could be plausibly replicated by abiotic amino acids like aspartate and serine. To achieve this, I require a model protein system in which the cysteine of a simple metal binding site could be systematically replaced with anionic abiotic amino acids in an otherwise stable folded protein that provided suitable binding geometry for metal coordination even in the apo state. To enable selection of an appropriate model protein, I first consider the plausible metal binding architectures of ancient proteins and the relevant literature on ancient metalloprotein folds.

Modern metalloproteins and metal-binding architectures

Metalloproteins are abundant in modern biology (Berg, 1990). These proteins participate in a wide array of processes, including electron transfer, catalysis, and gene expression. For example, iron and copper-containing proteins are heavily involved in metabolic respiration and photosynthesis. Meanwhile, zinc is used in over 300 enzymes and transcription factors, such as zinc finger proteins (Berg, 1990).

The evolution of metal-binding sites likely progressed from simple coordination sites to the complex structures observed today such as the heme-

copper centre in cytochrome c oxidase (Harel *et al.*, 2014). This progression in binding site complexity, and thus the need for larger proteins to correctly position coordinating residues, mirrors the increasing hydrophobicity and size of proteins over evolutionary history (Edwards *et al.*, 2013)(Siddiq *et al.*, 2017). Conversely, this trend suggests that ancient metalloproteins were smaller and used simple coordination structures to bind metals (Alva *et al.*, 2015)(Edwards *et al.*, 2013).

Tetrahedral coordination is one of the simplest geometries in which transition metals bind within proteins (**Fig 2.01**)(Permyakov, 2021). In these complexes, metal ions are coordinated by four ligands, such as the side chains of cysteine, histidine, and aspartate. The preference of amino acid side chains for binding certain metals can be understood through (Pearson, 1963). As explored in Chapter 1, the coordination geometry preferences of these metals are strongly dictated by their electron configuration and crystal field stabilisation. Cu^{+}, with its complete d^{10} configuration, has no crystal field stabilisation energy and hence little energetic preference for either octahedral or tetrahedral coordination environments, and is typically observed in linear or tetrahedral complexes. In contrast, Cu^{2+} is d^{9}, and experiences significant crystal field effects upon coordination that often favours a significantly distorted octahedral or tetrahedral structure.

Fe^{2+}, with a d^{6} configuration, gains appreciable stabilisation upon coordination, particularly in octahedral environments with stronger field ligands, but not so much to drive cysteine-coordinated Fe^{2+} to an octahedral configuration. This can promote the formation of redox active centres for catalysis. Fe^{3+} (d^{5}), in

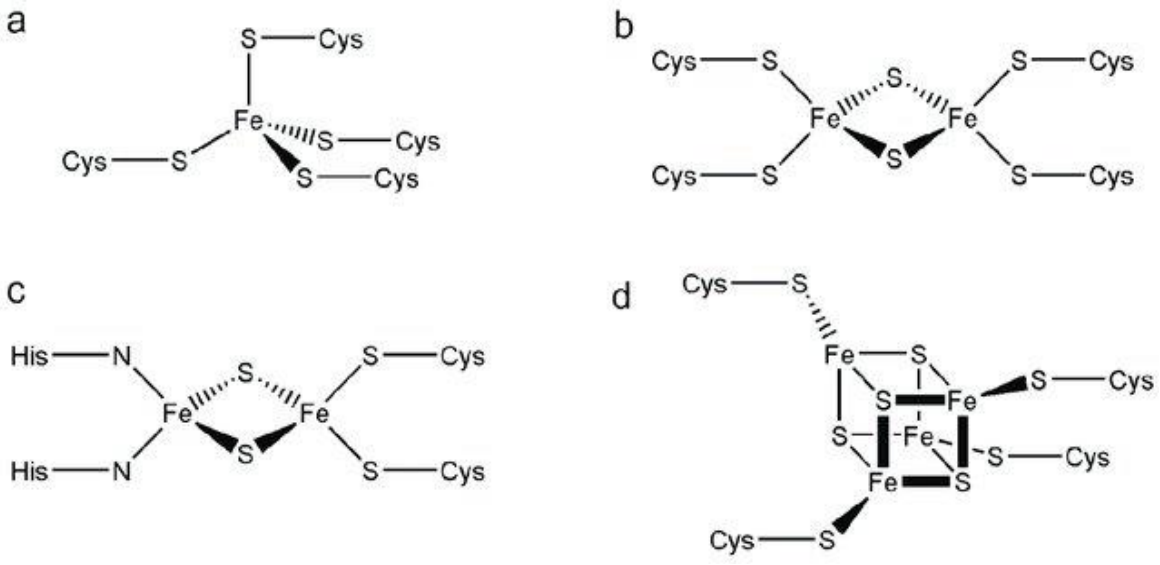


Figure 2.01 – Rubredoxin possesses the simplest Fe/S cluster structure observed in modern biology.

*(**a**) Rubredoxin, (**b**) [Fe_2S_2]-ferredoxin, (**c**) Rieske-centre, (**d**) [Fe_4S_4]-ferredoxin cluster. Adapted from* (Hwang and Han, 2015)

high spin complexes tends to favour a great number of electrostatic interactions with its ligands, hence favouring octahedral geometries. Iron binding sites thus display coordination geometry preferences that reflect the oxidation state of the metal and the field strength of the coordinating groups.

Selectivity of amino acid side chains for transition metal coordination

Cysteine, with its highly nucleophilic thiolate group, is particularly effective at coordinating transition metals due to its 'softness' and ability to coordinate metals in multiple oxidation states. Histidine, containing an imidazole nitrogen, is also a common metal-binding residue but is extremely difficult to synthesise abiotically and requires ordered geometric arrangements to coordinate metals effectively. This is because nitrogen lone pair in histidine's side chain is directionally constrained by the ring system it is bonded to, necessitating a more complex protein scaffold to position it correctly for metal binding. In no modern protein is zinc, cobalt, or nickel bound tetrahedrally exclusively by histidine and only in five non-homologous sequences is iron coordinated in this way. Far more common, however, are cysteine-exclusive and mixed tetrahedral cysteine/histidine binding sites (Dokmanić *et al.*, 2008).

Although aspartate and glutamate readily coordinate certain metals such as calcium and magnesium in modern proteins (e.g., within EF hand domains via the DxDxDG motifs), these structures generally do not provide the redox-active metal interactions likely crucial for early metabolic activity (Rigden and Galperin, 2004). When aspartate and glutamate are observed coordinating transition metals, it is always in the presence of more nucleophilic amino acids to form stable complexes (Dokmanić *et al.*, 2008). While modern proteins do not exhibit

aspartate-only tetrahedral coordination of iron, zinc, copper, cobalt or nickel, this absence may reflect evolutionary replacement rather than an inherent impossibility. As soon as cysteine became available through biosynthesis, its superior metal-binding properties likely replaced any early carboxylate-based sites. Therefore, investigating whether aspartate can coordinate transition metals in the absence of cysteine is needed to understand the capabilities of early proteins but will require dedicated mutagenesis of a modern tetrahedrally coordinating metalloprotein.

Criteria for selecting a model metalloprotein.

Reviewing the literature on ancient and conserved protein folds, I identified rubredoxin, a small iron-binding protein found in anaerobic bacteria and archaea involved in electron transfer, as a suitable candidate (Alva *et al.*, 2015). Rubredoxin is a 54-amino acid protein that contains a single $Fe^{2/3+}$ ion coordinated tetrahedrally by the thiolate groups of four cysteine residues within two CXXC sequence motifs near the N and C termini (Bertini *et al.*, 1998). This minimalistic metal-binding motif represents one of the simplest known folded structures and was identified by Alva *et al* to contain one of the earliest known protein structural fragments (Alva *et al.*, 2015). In addition, its small size, ease of purification, thermostability up to 120 °C, and lack of complex tertiary interactions, make it readily suitable for screening experiments (Maiti *et al.*, 2017). On a similar basis, Nanda *et al* and *Chino et al* have successfully used the beta-loop-beta architecture of rubredoxin to design rubredoxin mimics at both the protein and peptide scale, demonstrating its utility within mutagenesis studies (Nanda *et al.*, 2005)(Chino *et al.*, 2023).

Research has been conducted on rubredoxin since the 1970s, where it was used as one of the earliest models for metalloprotein purification and structural resolution (Watenpaugh, K. D. *et al.*, 1972). High-resolution crystal and NMR structures of rubredoxin in its iron-bound and various transition metal-bound states, including zinc, nickel, and cobalt, have provided strong constraints on the purification processes and folding topology of rubredoxin (Watenpaugh, K. *et al.*, 1980)(Maher *et al.*, 2004). Despite this, the structure and stability of the apo form (metal-free) is comparatively less understood. Previous NMR studies suggest that apo rubredoxin is purified predominantly in an unstructured state, while more recent studies using synchrotron radiation circular dichroism indicate that the apo form retains the limited secondary structure content observed in native rubredoxin (Gondim *et al.*, 2019)(Almeida *et al.*, 2021). This gap presents a minor challenge for this research, necessitating that I purify and assign the apo state to ensure I understand what the apo to holo transition looks like in the native protein before screening a range of mutant structures against different metals. While this is a challenge, the compelling data on its ancient origins and extreme ease to rubredoxins purification and analysis by NMR justified its choice as a model metalloprotein here (Xiao *et al.*, 1998).

Approach

In this study I aimed to identify if aspartate or serine can substitute for cysteine within tetrahedral metal binding sites in a stable folded protein. To achieve this, I designed and purified four rubredoxin variants. A consensus sequence (RbdC) as a positive control and three variants (RbdD, RbdDS1, RbdDS2) featuring substitutions of the iron-binding cysteines with aspartate

and/or serine. Using NMR spectroscopy, UV-visible spectroscopy, circular dichroism (CD), and other biophysical assays, I assessed the rubredoxin apo state folding capability and the capacities of the variants to bind iron, nickel, zinc, cobalt, and magnesium.

These transition metals were chosen because they represent a set of soft metals that were likely bioavailable in the Archean Ocean (iron, nickel cobalt) or play critical roles in modern metalloproteins (zinc) (Berg, 1990)(Hazen and Sverjensky, 2010). Magnesium was included to assess how well the mutant proteins can coordinate non-transition metals.

If these cysteine-free variants fail to bind metals within an otherwise folded protein, then it would suggest that life, requiring a means to fix metal ions for catalytic function, must have been localised to an environment where other metal coordination mechanisms were provided from the environment directly or abiotic cysteine was available. Alternatively, if metal binding proceeds readily without cysteine, it would suggest that the emergence of cysteine, while advantageous, may not have been strictly necessary.

In this study, I used a ‘consensus’ sequence rubredoxin, generated by aligning rubredoxin sequences from several thousand species collected via BLAST and selecting the most common residue at each position using JALVIEW to create a generalisable representative of the family. My rationale for this approach was twofold: first, I wanted to produce a sequence that captured features conserved across the rubredoxin class, rather than relying on a single organism’s version, such as that from Clostridium. Second, I opted for the

consensus strategy because I did not have the technical expertise or available time to perform a full ancestral sequence reconstruction, which is a far more involved computational process. I recognise this was a relatively high-risk strategy compared to using a naturally evolved or reconstructed ancestral sequence, since it could in principle generate a protein with reduced stability or loss of function. In hindsight, selecting a well-characterised extant sequence might have provided a safer starting point. Nonetheless, the consensus design ultimately yielded a functional rubredoxin that allowed exploration of the broader features of iron-sulfur coordination in this protein family.

Methods

Purification

Consensus and mutant selection and purification in E.coli

A consensus sequence for rubredoxin was generated using the European Bioinformatics Institute (EBI) BLAST tool, with rubredoxin from Clostridium pasteurianum (PDB ID: 1BFY) serving as the input sequence. The BLAST search was set to 5000 entries, which were exported using the aligned export tool and processed with the Jalview software to produce a consensus sequence (RbdC) and used to generate a sequence logo via the WebLogo tool.

RbdC

MKKYVCTVCGYIYDPEVGDPDNGIAPGTAFEDLPDDWVCPLCGVGKDDFEPVEE

Mutant sequences were generated by substituting cysteine residues at positions 6, 9, 39, and 42 with aspartic acid or mixtures of aspartic acid and serine. Structural predictions for each sequence were generated using AlphaFold3 (Abramson *et al.*, 2024).

RbdD

MKKYVDTVDGYIYDPEVGDPDNGIAPGTAFEDLPDDWVDPLDGVGKDDFEPVEE

RbdDS1

MKKYVDTVSGYIYDPEVGDPDNGIAPGTAFEDLPDDWVSPLDGVGKDDFEPVEE

RbdDS2

MKKYVSTVDGYIYDPEVGDPDNGIAPGTAFEDLPDDWVDPLSGVGKDDFEPVEE

The sequences were codon-optimised for expression in *Escherichia coli* and cloned into the pET-11a vector by GenScript. The resulting plasmids (4µg) were resuspended in 20µl of Milli-Q water and stored at -20 °C.

Transformation into competent *E. coli* BL21(DE3) cells was performed using the heat shock method. Competent cells were incubated with 0.5µl of 200 ng/µl plasmid at 45 °C for 45s, followed by a 5 min recovery on ice. Transformed cells were plated on LB agar containing 100 µg/ml ampicillin and incubated overnight at 37°C. Single colonies were selected and grown in LB broth (100 µg/ml ampicillin) for 4h until an optical density at 600nm (OD_{600}) of 1.0 was reached. A 20ml aliquot of this starter culture was used to inoculate 1l of LB high-salt medium (LB HS) with 100 µg/ml ampicillin, grown at 37 °C until an OD_{600} of 0.6-0.8 was achieved.

For unlabelled protein expression, cultures were induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) and iron chloride ($FeCl_3$) to final concentrations of 1mM and 20µM, respectively, which drives expression of the RbdC protein under control of the pET 11a lac promoter sequence and to provide sufficient iron for metal acquisition during overexpression. An additional 500µl of 100 mg/ml ampicillin was added. Cultures were incubated for 16-20h at 18 °C. Cells were harvested by centrifugation at 7000x*g* for 15 min, resuspended in 40ml of lysis buffer (20mM PIPES, 150mM NaCl, 1mM EDTA, pH 5.0), and stored at -80 °C until purification.

^{13}C-^{15}N Isotope labelling for NMR Studies

For isotopically labelled grows, M9 minimal medium was prepared under sterile conditions using an open flame. 2x M9 was prepared from 200ml of 10x M9 salts for NMR (432mM Na_2HPO_4, 222mM KH_2PO_4, 86mM NaCl) and 800 ml of autoclaved Milli-Q water were combined. Magnesium sulfate ($MgSO_4$) was added to a final concentration of 1mM. Thiamine (1 mg/ml stock), ampicillin (100 mg/ml stock) and 0.5ml of 2000x trace metals solution was also added. Finally, 2 g/L of ^{13}C-glucose and ^{15}N ammonium chloride (NH_4Cl) was added at 1 g/L.

Cells were grown in LB medium until an OD600 of 0.5-0.6 was reached, then harvested by centrifugation at 4000g for 15 min at room temperature. Pellets were gently resuspended into 1L 1x M9 medium and centrifuged as before to remove residual LB components. Cells were resuspended in 2x M9 (with ^{13}C glucose and ^{15}N ammonium chloride) and incubated for 15 min at room temperature. IPTG was added to a final concentration of 2mM, and cultures were incubated for 16-18hrs at 18 °C with shaking at 180 rpm. Cultures were harvested as described previously.

Rubredoxin purification

Thawed cell pellets were sonicated on ice using a probe sonicator set at 45 amplitudes with 5s on/25 s off cycles. The lysate was heated to 80°C (for RbdC) or 50°C (for Rbd variants) for 30 min in a water bath to denature *E. coli* host proteins, then cooled on ice for 5 min. The lysate was centrifuged at 30,000g for 45 min at 4°C to remove precipitated material.

The supernatant, which exhibited a deep red colour for RbdC due to its iron content, was subjected to ammonium sulfate precipitation by sequentially adding solid ammonium sulfate to achieve 40%, 60%, 80%, and 90% saturation. Centrifugation at 15,000g for 15 min at 4°C was performed between each addition with Rubredoxin variants precipitating between 80% and 90% ammonium sulfate saturation.

The pellet obtained at 90% saturation was resuspended in 20mM Tris-HCl buffer, pH 7.4, and dialysed overnight against 4l of the same buffer to remove ammonium sulfate. The dialysed solution was loaded onto a 5ml HiTrap Q FF anion exchange column (Cytiva) equilibrated with buffer A (20mM Tris-HCl, pH 7.4, containing 10mM EDTA) on an ÄKTA pure FPLC (Cytiva). Proteins were eluted with a linear gradient from 0 to 500mM NaCl in buffer B (20mM Tris-HCl, pH 7.4, containing 500mM NaCl).

Fractions containing rubredoxin, identified by SDS-PAGE and by eye (based on colour), were pooled and concentrated to 25 mg/ml using a Vivaspin 20 centrifugal concentrator with a 5 kDa molecular weight cutoff with concentration monitored using a Benchtop Nanodrop (Thermo NanoDrop 1000 UV-Vis Spectrophotometer). During concentration, the solution was buffer exchange into 0.1M phosphate buffer, pH 6.0 by 2x successive dilution and concentration steps.

Size-exclusion chromatography (SEC) was conducted using a Superdex 75 Increase 10/300 GL column (Cytiva) equilibrated with 0.1M phosphate buffer, pH 6.0 on an ÄKTA pure FPLC (Cytiva). A 1ml aliquot of 25 mg/ml rubredoxin

was injected onto the column and eluted at a flow rate of 0.75 ml/min over 20 minutes. Eluted fractions were pooled and concentrated as before. Purification success was confirmed by SDS-PAGE, mass spectrometry and SEC-MALS.

SDS-PAGE analysis

Protein samples (30µL) were mixed with 10µl of 4x NuPAGE LDS Sample Buffer (Thermo Fisher Scientific). 10µl aliquots were loaded onto NuPAGE 4-12% Bis-TRIS protein gradient gels (Thermo Fisher Scientific) assembled in an XCell SureLock Mini-Cell electrophoresis system with NuPAGE MES SDS running buffer. Electrophoresis was performed at a constant voltage of 200V, 110mA for 35 minutes. Gels were stained using a Coomassie Blue staining solution and destained. Protein bands were visualised using a GelDoc imaging system.

Generation of apo-rubredoxin from Fe-holo-rubredoxin

Purified Fe-Holo-RbdC (500µl at 25 mg/ml in 0.1M Phosphate Buffer pH 6.0) was treated with 73µl of 80% (w/v) trichloroacetic acid (TCA) to achieve a final concentration of 10% TCA alongside 20mM EDTA, which acidified the solution causing complete protein unfolding and release of free iron into solution which is scavenged by EDTA as insoluble precipitates. The mixture was incubated on ice for 10 min and then centrifuged at 7000g for 15 min at 4°C. Pellets were retained at each step and the supernatant was carried over into a new tube. The TCA precipitation step was repeated three times on the supernatant to ensure complete removal of metal ions. The protein pellet was washed twice with 50ul MilliQ water and finally with 20ul of 0.5M NaOH to aid re-solubilisation.

The pellets were resuspended in 500µl of 8M urea containing 0.3M β-mercaptoethanol and a twofold molar excess of the desired metal chloride ($ZnCl_2$, $CoCl_2$, $NiCl_2$). When generating apo-RbdC, 0.3M β-mercaptoethanol was substituted for 0.3M tris(2-carboxyethyl)phosphine (TCEP) and no metal chloride was added. The solution was incubated on ice for at least 1h or until the protein was fully solubilised (which could be observed through the loss of precipitation on the tube wall). The solution was then diluted twofold with Milli-Q water and dialysed overnight against 4L of unbuffered Milli-Q water at pH 4.5 (and 25mM TCEP for Apo-RbdC), with at least two buffer changes to remove urea and excess reagents.

Protein solution NMR spectroscopy

^{13}C-^{15}N labelled samples of Apo-RbdC, Holo-RbdC (metal-bound forms Zn^{2+}, Fe^{2+}), RbdD, RbdDS1, RbdDS2 were prepared at up to 1.8mM concentrations in 0.1M Phosphate buffer or in unbuffered MilliQ and made up to 90% H_2O/10% D_2O, at pH 4.5, adjusted using HCl within 5mm Bruker NMR tubes. NMR data were recorded on spectrometers operating at 14.1, 17.6 or 22.3 Tesla (Oxford Instruments magnets) equipped with high sensitivity 5mm TCI cryoprobes and Avance III HD consoles (Bruker). 2D ^{1}H-^{15}N heteronuclear single quantum correlation (^{1}H-^{15}N HSQC) spectra were acquired with 2 scans (***Sc***) for HoloRbdC and up to 36 ***Sc*** for less concentrated samples, and the total points (***Td***) and sweep widths (***Sw***) were 2048 and 128, and 16.02ppm and 34.00ppm, for the direct and indirect dimensions, respectively. 3D triple-resonance spectra were collected as follows: HNCA, ([12 ***Sc***], [1022, 50, 110 ***Td***], [14.03, 34.00, 30.00ppm ***Sw***]), HNCACB ([16 ***Sc***], [1022, 50, 110 ***Td***], [14.03, 34.00, 70.00ppm

Sw]), HNCO ([4 ***Sc***], [1022, 50, 110 ***Td***], [14.03, 34.00, 15.00ppm ***Sw***]), where ***Td*** and ***Sw*** values are listed for ^{1}H, ^{15}N, and ^{13}C dimensions, respectively. All experiments were conducted at 20°C at pH 4.5 unless otherwise indicated. Resulting spectra were processed using NMRpipe and analysed using CARA (Delaglio *et al.*, 1995)(Keller, 2004).

The RbdD variant was also assigned at pH 7.0 for comparative purposes and titrations of RbdD were conducted by addition of incremental volumes of 0.1M NaOH to 1.0mg/ml RbdD from pH 4.0-8.0. The effect of salinity of the titration curve was monitored by repeating this analysis and making the solution up to 0.1M phosphate buffer, and further addition of anhydrous NaCl to a final concentration of 0.1M NaCl.

Metal binding screening

Purified ^{13}C, ^{15}N-labeled Zn-Holo-RbdC, RbdD, RbdDS1, and RbdDS2 were prepared at 80, 635 1004, 1812µM respectively and equilibrated in 100% N_2 overnight. Samples were screened against metal chlorides (Fe^{2+}, Ni^{2+}, Co^{2+}, Zn^{2+}, Mg^{2+}) at metal-to-protein molar ratios of 2:1 and 10:1. Samples were placed in 3mm NMR tubes and analysed on a 22.3 Tesla spectrometer (Oxford Instruments magnets) equipped with a high sensitivity 5mm TCI cryoprobe and Avance III HD consoles (Bruker) with a SampleJet autosampler at 20°C at pH 4.7 ±0.15. 2D ^{1}H-^{15}N HSQC spectra were acquired ([32 ***Sc***], [2048, 256 ***Td***], [15.96, 32.07 ***Sw***]) to monitor metal binding and assess changes in protein structure. Resulting spectra were processed using NMRpipe and exported into CARA where peak positions were adjusted. Updated chemical shifts were exported and

chemical shift perturbations for each assigned residue vs its assigned structure at pH 4.5 were calculated (Williamson, 2013).

Heteronuclear NOE experiments

Two-dimensional ^{1}H-^{15}N heteronuclear NOE (hetNOE) experiments were conducted on Apo-RbdC and Zn-Holo-RbdC to investigate structural dynamics. Samples were prepared at 100µM in 90% H_2O/10% D_2O, pH 4.5, and loaded into 5mm Bruker NMR tubes. Experiments were performed at 25 °C on a 17.6 Tesla Spectrometer using the ***hsqcnoef3gpsi*** Bruker pulse sequence. ([128 ***Sc***], [2048, 120 ***Td***], [14.00, 16.00 ***Sw***]). The resulting spectra was processed using NMR pipe and resulting split peak intensities exported within the CARA software, with the hetNOE ratio reported.

Biophysical Analysis

UV-visible spectroscopy

Purified Fe-Holo RbdC and RbdD were diluted to 1 mg/ml in 0.1M phosphate buffer, pH 6.0. 200µl samples were placed in Hellma Analytics high-precision quartz cuvettes with a 10mm path length. UV-visible spectra were recorded from 200nm to 800nm using a Cary 50 Bio spectrophotometer (Agilent Technologies) at a scan rate of 4800 nm/min with 1nm data intervals. Baseline correction was performed using the buffer as a reference.

Differential scanning fluorimetry (NanoDSF)

Samples of Apo-RbdC, Holo-RbdC (Fe^{3+}, Zn^{2+}, Co^{2+}, Ni^{2+} forms), and RbdD were prepared at 5 mg/ml in 0.1M phosphate buffer at pH 4.5 or pH 6.2. 10µl aliquots were loaded into standard 10µl microcapillaries and analysed using

a Prometheus NT.48 instrument (NanoTemper Technologies). Thermal unfolding was monitored by increasing the temperature from 20°C to 95°C at 1°C/min while measuring intrinsic fluorescence at 330nm and 350nm. First derivative of fluorescence at 330nm was analysed to determine melting temperatures (Tm).

Circular dichroism spectroscopy

Note: The following CD experiment was conducted by Dr David Staunton within the Biophysics Facility (Department of Biochemistry, University of Oxford)

Spectral analysis was conducted by the author

Samples of purified Fe-Holo-RbdC and RbdD were diluted to 0.1 mg/ml in 0.1M phosphate buffer at various pH values. Spectra were recorded using a Jasco J-815 CD spectropolarimeter (JASCO International) at 20°C. Samples were placed in a Starna Type 21 quartz cuvette with a 1 mm path length. Spectra were acquired from 250nm to 200nm, averaging nine accumulations per sample. Data with voltage readings exceeding 700 V were excluded due to poor ellipticity signal-to-noise ratios.

Size-Exclusion Chromatography-Multi-Angle Light Scattering (SEC-MALS)

Note: The following SEC-MALs experiment was conducted by Dr David Staunton within the Biophysics Facility (Department of Biochemistry, University of Oxford).

Purified Fe-Holo-RbdC was prepared at concentrations of 25 mg/ml, 12.5 mg/ml, and 6.25 mg/ml by serial dilution in 0.1M phosphate buffer, pH 6.0. Aliquots (100µL) of each sample were injected onto a Superdex 75 Increase 10/300 GL column (Cytiva), which had been equilibrated with the same buffer.

Protein elution was monitored simultaneously using a DAWN HELEOS 8+ multi-angle light scattering detector (Wyatt Technology), an Optilab T-rEX refractive index detector (Wyatt Technology), and a UV detector (SPD-20A, Shimadzu). Prior to SEC-MALS analysis, proteins had been purified by SEC to ensure sample purity. Chromatographic and light-scattering data were analysed using the ASTRA v6 software package (Wyatt Technology) to determine the molecular weight and assess sample homogeneity.

Proteomics

In-solution digestion of protein samples

50µl of 2.5mg/ml protein samples were denatured by adding 8M urea in 100mM ammonium bicarbonate in a 1:1 volume with the sample. Samples were incubated at room temperature with shaking at 650 rpm for 10 minutes. Reduction was performed by adding TCEP to a final concentration of 5mM and incubating for 30 minutes at room temperature without shaking. Alkylation was conducted by adding chloroacetamide to a final concentration of 10mM and incubating in the dark for 30 minutes at room temperature.

Predigestion with Lys-C was performed at an enzyme-to-substrate ratio of 1:100 (w/w) for 2h at 37 °C with shaking at 800 rpm. The urea concentration was then diluted to 2M with 100mM ammonium bicarbonate. Calcium chloride was added to a final concentration of 1mM. Chymotrypsin digestion was carried out at an enzyme-to-substrate ratio of 1:40 (w/w) overnight at 37 °C with shaking at 800 rpm. Digestion was quenched by adding formic acid to a final concentration of 1% (v/v). Samples were stored at -20 °C until desalting.

C18 Tip desalting

Samples were centrifuged at 14,000 rpm for 30 minutes at 4°C to remove insoluble material. C18 tips were prepared by inserting a 2mm disk of C18 extraction resin (Empore SPE Disks, 3 M) into a 10µl pipette tip using a blunt needle.

The C18 tips were activated with two washes of 60µl 100% acetonitrile, centrifuged at 4000 rpm at room temperature. Tips were equilibrated with 60µl of 0.1% trifluoroacetic acid (TFA) twice, centrifuging between washes. Peptide samples were loaded onto the tips in 60µl aliquots, centrifuging at 4000 rpm until the liquid passed through before adding the next aliquot.

After sample loading, tips were washed twice with 60µl of 0.1% TFA, centrifuging each time. Peptides were eluted into fresh low-binding microcentrifuge tubes using two 60µl aliquots of 50% acetonitrile/0.1% TFA, centrifuging at 4000 rpm. Eluted peptides were dried in a vacuum concentrator and stored at -20 °C prior to mass spectrometry analysis.

Mass spectrometry analysis

Note: *The following analysis was conducted by Dr Majorie Fournie using samples prepared by the author at the Proteomics Facility (Department of Biochemistry, University of Oxford).*

Peptides were reconstituted in 0.1% formic acid (FA) in water. Nano-liquid chromatography was performed using an Ultimate 3000 RSLC system (Thermo Fisher Scientific) coupled to a Q Exactive mass spectrometer with an EASY-Spray source. Peptides were trapped on a C18 PepMap100 precolumn

(300µm × 5mm, 100Å) and separated on an EASY-Spray RSLC C18 analytical column (75µm × 50cm, 2 µm particle size) using a linear gradient from 15% to 35% solvent B (0.1% FA in acetonitrile) over 15 min at a flow rate of 200 nl/min.

The mass spectrometer operated in data-dependent acquisition mode. Full MS spectra were acquired in the Orbitrap (m/z 350-1500) at a resolution of 70,000, AGC target of 3×10^6, and maximum injection time of 50ms. The top five most intense ions were selected for higher-energy collisional dissociation (HCD) fragmentation at a normalised collision energy of 30%. MS/MS spectra were acquired at a resolution of 17,500, AGC target of 5×10^4, and maximum injection time of 120ms. Dynamic exclusion was set to 5s. Data analysis was performed with a 1% false discovery rate (FDR) applied for protein identification.

Results

Conservation of rubredoxin structure and design of variants

The consensus sequence generated from a multiple sequence alignment of 5000 rubredoxin sequences exhibits very high levels conservation throughout, particularly at structurally or functionally important positions such as the metal coordination site (residues 6,9,39 and 42) and positions within the π-stacked hydrophobic core (residues 13, 30, 37) (**Fig. 2.02**). To investigate the role of cysteine and the potential for metal coordination by abiotic amino acids, several rubredoxin variants were designed (**Fig 2.03a-3e**). AlphaFold3 produced high confidence structural models for all sequences when submitted with Fe^{3+} as a ligand with ipTM and pTM scores of (0.91/0.86, 0.89/0.85 0.87/0.84 and 0.76/0.81), where ipTM indicates confidence for the protein/Fe complex. This suggests that metal binding may be possible in all tested structures.

Purification and characterisation of rubredoxin variants

Consensus rubredoxin (RbdC)

RbdC was successfully expressed and purified in *E. coli*. Anion exchange chromatography showed a prominent peak corresponding to RbdC (**Fig 2.04a**), and SEC displayed a single nearly symmetric peak, indicating a homogenous sample (**Fig 2.04b**). CD spectroscopy spectra are consistent with previous reports for Rubredoxin (**Fig 2.04c**) (Zartler *et al.*, 2001)(Greenfield, 2006). UV-visible spectroscopy exhibited absorption bands were observed at approximately 350, 490, 590 and 660nm, consistent with previously reported Fe-S clusters in rubredoxin (**Fig 2.04d**) (Sushko *et al.*, 2021). SDS-PAGE analysis revealed a

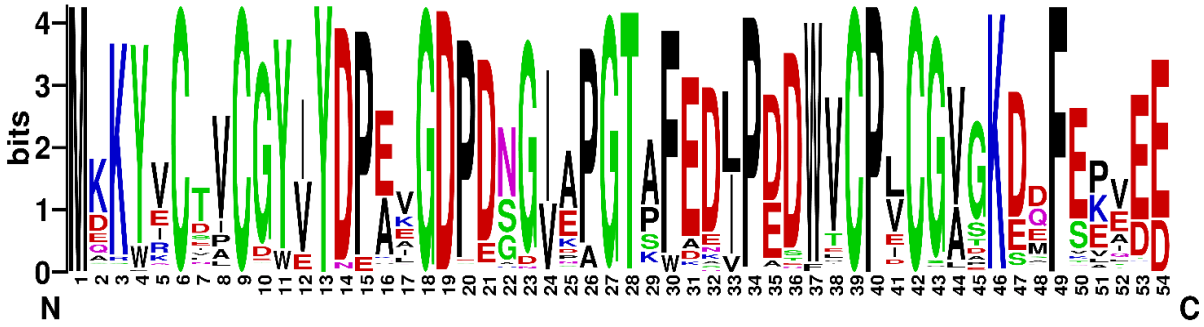


Figure 2.02 – Rubredoxin shows complete conservation of its melting binding site and hydrophobic core.

A sequence logo for Rubredoxin was generated from a BLAST multiple sequence alignment of 5,000 rubredoxin sequences using the sequence from PDB structure 1BFY as the input. Visualization was performed with WebLogo (Crooks *et al.*, 2004)

Name: Rubredoxin (RbdC)
Amino Acids: 54
Mass: 5.95kDa
Isoelectric Point: 3.73
Thermostability: Up to 95C
CXXC Iron Binding Motif

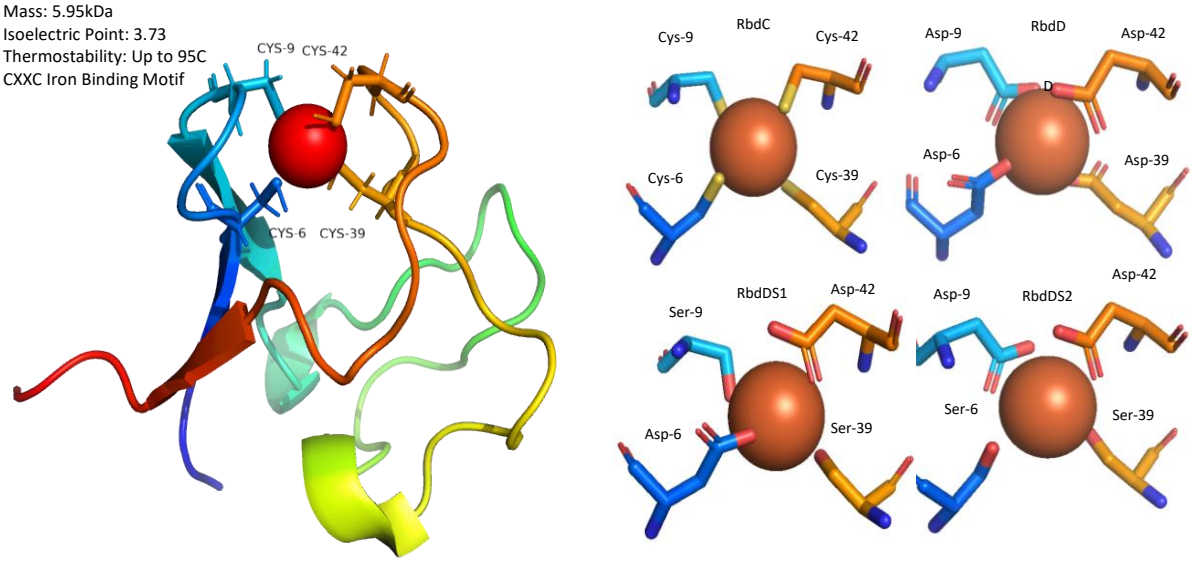


Figure 2.03 – AlphaFold3 predictions of consensus Rubredoxin and its variants

RbdC, D, DS1 and DS2 sequences were submitted to Alphafold3 alongside a Fe^{3+} ligand producing ipTM scores of 0.91, 0.89 0.87 and 0.76 respectively, consistent with a strong prediction of protein/ligand complex formation for all structures baring DS2. Predicted structures were visualised and images exported from Pymol (educational licence)

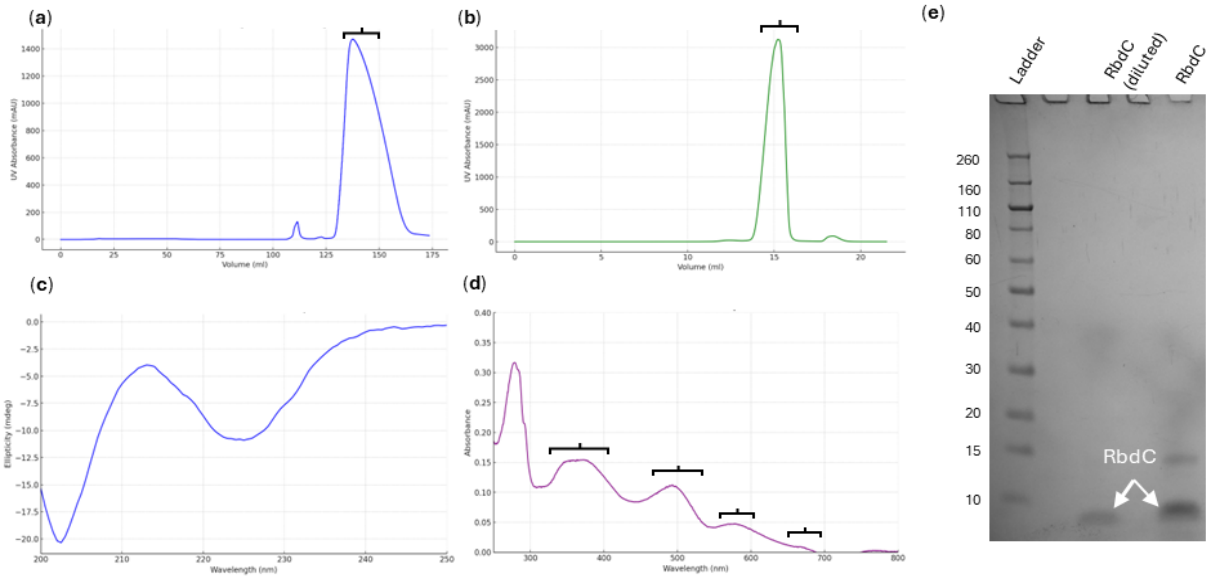


Figure 2.04 – Rubredoxin purifies readily without requiring affinity labelling based purification.

*(**a**) Fe-Holo-RbdC was eluted using a 5ml HiTrap Q FF column equilibrated with 20mM Tris-HCl, pH 7.4, 10mM EDTA with a linear gradient from 0 to 500 mM NaCl. A large peak between 130 and 160ml was attributed to Fe-Holo-RbdC based on visual colour of the fraction. (**b**) Pooled fractions were concentrated using a 5kda Vivaspin concentrated and 1ml was injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1M phosphate buffer, pH 6.0. A single asymmetric large peak was observed from 14-16mins. (**c**) CD spectra of the eluted protein concentrated to 2.5mg/ml was recorded using a Jasco J-815 CD spectropolarimeter at 20 °C. The spectrum, obtained in 0.1M phosphate buffer (pH 6.0), is, consistent with literature CD signals for rubredoxin.(**d**) UV-visible spectroscopy of RbdC was recorded with a Cary 50 Bio spectrophotometer, showing absorption bands centred at approximately 350, 490, 590, and 660 nm, consistent with iron-sulfur sites in rubredoxin. (**e**) SDS-PAGE analysis of purified Fe-Holo-RbdC, displaying a single band just below 10 kDa and a minor band at 17kda, consistent with the expected molecular weight of rubredoxin (5.9 kDa) and an oligomeric product.*

single band below 10 kDa, consistent with the theoretical molecular weight of rubredoxin (5.9kDa) (**Fig 2.04e**).

Aspartate mutant rubredoxin (RbdD)

RbdD was also successfully purified. Anion exchange chromatography showed a broad peak (185-230ml), with broadening likely due to high sample concentrations resulting in column overloading, and SEC showed a clean large peak at ≈15ml similar to RbdC (**Fig 2.05a-b**). Unlike RbdC, the UV-visible spectrum of RbdD lacked the characteristic Fe-S absorption bands, confirming that spontaneous iron binding at the point of expression has been lost (**Fig 2.05c**). SDS-PAGE following SEC showed the presence of a protein band between 5-10kDa, consistent with the theoretical mass of rubredoxin (**Fig 2.05d**).

SEC coupled with multi-angle light scattering (SEC-MALS) estimated the molecular weight of purified RbdC to be approximately 4.3 kDa, which while 1 kDa less than the calculated mass, is consistent with the reduced signal to noise expected from SEC-MALS for a protein of this size. The SEC-MALS confirmed that the protein was in a monomeric state in solution (**Fig 2.06a**). Mass spectrometry proteomics confirmed the identity of RbdD and RbdC with 81.48% and 83.02% sequence coverage, respectively. While other contaminants were observed, the comparatively low sequence coverage indicated that these were minor contaminants that were unlikely to significantly impact further analysis (**Fig 2.06b**).

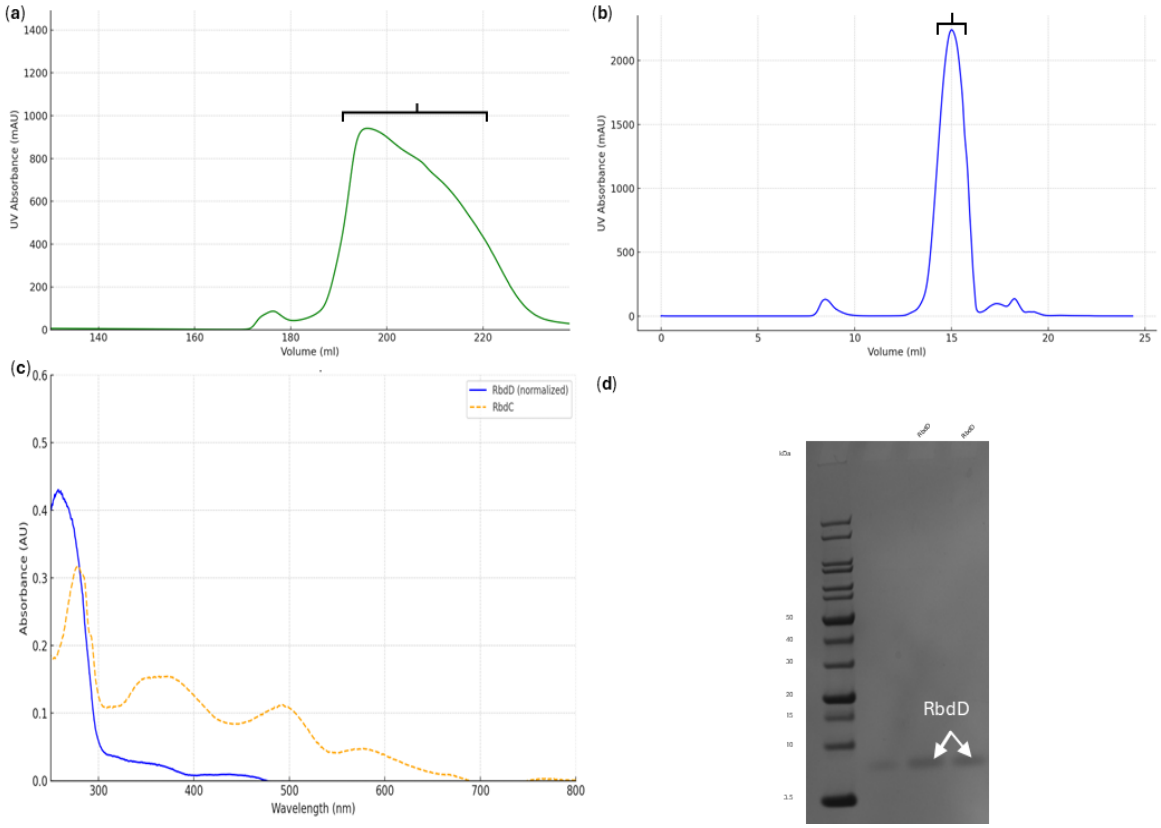


Figure 2.05 – RbdD purifies similarly to RbdC.

*(**a**) RbdD was eluted using a 5ml HiTrap Q FF column equilibrated with 20 mM Tris-HCl, pH 7.4, 10mM EDTA with a linear gradient from 0 to 500mM NaCl. A large peak between 185 and 230ml was observed. (**b**) Peak fractions were concentrated using a 5kda Vivaspin concentrated and injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1M phosphate buffer, pH 6.0. A single asymmetric large peak was observed from 14-16mins.(**c**) UV-visible spectroscopy of RbdD (blue) was recorded with a Cary 50 Bio spectrophotometer, showing loss of iron sulfur absorption bands (orange) consistent with the loss of an iron-sulfur cluster in this variant. (**d**) SDS-PAGE analysis of purified RbdD, displayed a single band just below 10 kDa consistent with the expected mass of 5.9 kDa for RbdD.*

(**a**)

Concentration	Mw (kDa)	Uncertainty	Calculated mass (µg)	Mass fraction (%)
25mg/ml	3.9	24.20%	810.78	100
12.5mg/ml	3.9	14.00%	395.95	100
6.25mg/ml	4.6	11.60%	194.48	100
Average	4.1		467.07	100
Standard deviation	0.3		374.35	0
% Standard deviation	8.4		80.15	0
Minimum	3.9		194.48	100
Maximum	4.6		810.78	100

(**b**)

ESI Identified Proteins									
Accession	Description	Score A2	Coverage A2	# Peptides A2	# PSM A2	Score C2	Coverage C2	# Peptides C2	# PSM C2
Rubredoxin	DEF_USER|RbdC_Rubredoxin					7.98	83.02	1	2
Rubredoxin mutant (RbdD)	DEF_USER|RbdD_Rubredoxin_mutant	18.64	81.48	2	5				
P03051	Regulatory protein rop OS=Escherichia coli					1.87	53.97	2	2
P62593	Beta-lactamase TEM OS=Escherichia coli					0.00	3.50	1	1
P25240	Modification methylase Eco57IB OS=Escherichia coli					0.00	7.90	1	1
Q9XCW3	dTDP-4-amino-4,6-dideoxy-D-glucose acyltransferase OS=Escherichia coli	0.00	6.25	1	1				
P03008	Transposase for transposon Tn3 OS=Escherichia coli	0.00	0.79	1	1	0.00	0.79	1	1

Figure 2.06 – RbdD and RbdC are monomeric and nearly pure in solution following purification.

*(**a**) SEC-MALS analysis was conducted on previously purified Fe-Holo-RbdC. Samples were prepared at concentrations of 25 mg/mL, 12.5 mg/mL, and 6.25 mg/mL in 0.1M phosphate buffer, pH 6.0 and injected on to a Superdex 75 Increase 10/300 GL column and eluted into a DAWN HELEOS Multi-Angle static Light Scattering (MALS) detector. RbdC created a single large peak with a molecular weight prediction of approximately 4.1 kDa, consistent with a monomeric state of the protein in solution. The observed molecular weight was 1.8 kDa less than the predicted mass, which is attributed to the reduced signal-to-noise ratio expected for proteins of this size. (**b**) Mass spectrometry proteomics of in solution LysC, and chymotrypsin digested purified RbdD and Fe-Holo-RbdC resulted in identified peptide mass hits consistent with 81.48% and 83.02% sequence coverage, respectively. Minor contaminants were detected but were judged to be unlikely to significantly affect downstream analysis due to their comparatively low sequence coverage.*

Additional variants (RbdDS1, RbdDS2)

Purification of RbdDS1, and RbdDS2 yielded results consistent with RbdD and RbdC. RbdDS1 and RbdDS2, AEX (**Fig 2.07a-b**) and SEC (**Fig 2.07c-d**) chromatograms indicated two intense elution peaks which were both retained and pooled. Subsequent analysis by SDS page for RbdDS1 (**Fig 2.07e**) indicated that both peaks had the same apparent mass, and NMR confirmed that they had identical ^{1}H-^{15}N HSQC resonances, suggesting that the two-peak elution originates from an interaction of the rubredoxin variant with the columns as opposed to contamination by another protein.

Structural analysis using NMR spectroscopy.

Assignment of holo and apo Rubredoxin

The holo and apo forms of RbdC were assigned through triple resonance experiments. The holo form exhibited fewer resonance signals due to signal suppression from the paramagnetic iron, which induces fast nuclear spin relaxation, broadening peaks and reducing signal intensities (**Fig 2.08a**). In contrast, the apo form showed well-dispersed signals, indicating a well-folded protein without the metal ion (**Fig 2.08b**). The assigned ^{1}H-^{15}N HSQC of zinc-bound RbdC displayed a spectrum visually similar to the apo form (**Fig 2.09a**). However significant chemical shift perturbations (CSP) were observed around the metal-binding site confirming that the apo state was indeed metal free (**Fig 2.09b**). Visual mapping of these CSPs onto the rubredoxin structure confirmed that zinc binding affects the immediate area around the metal-binding site without disrupting the hydrophobic core of the protein (residues 13-37) (**Fig 2.09c**).

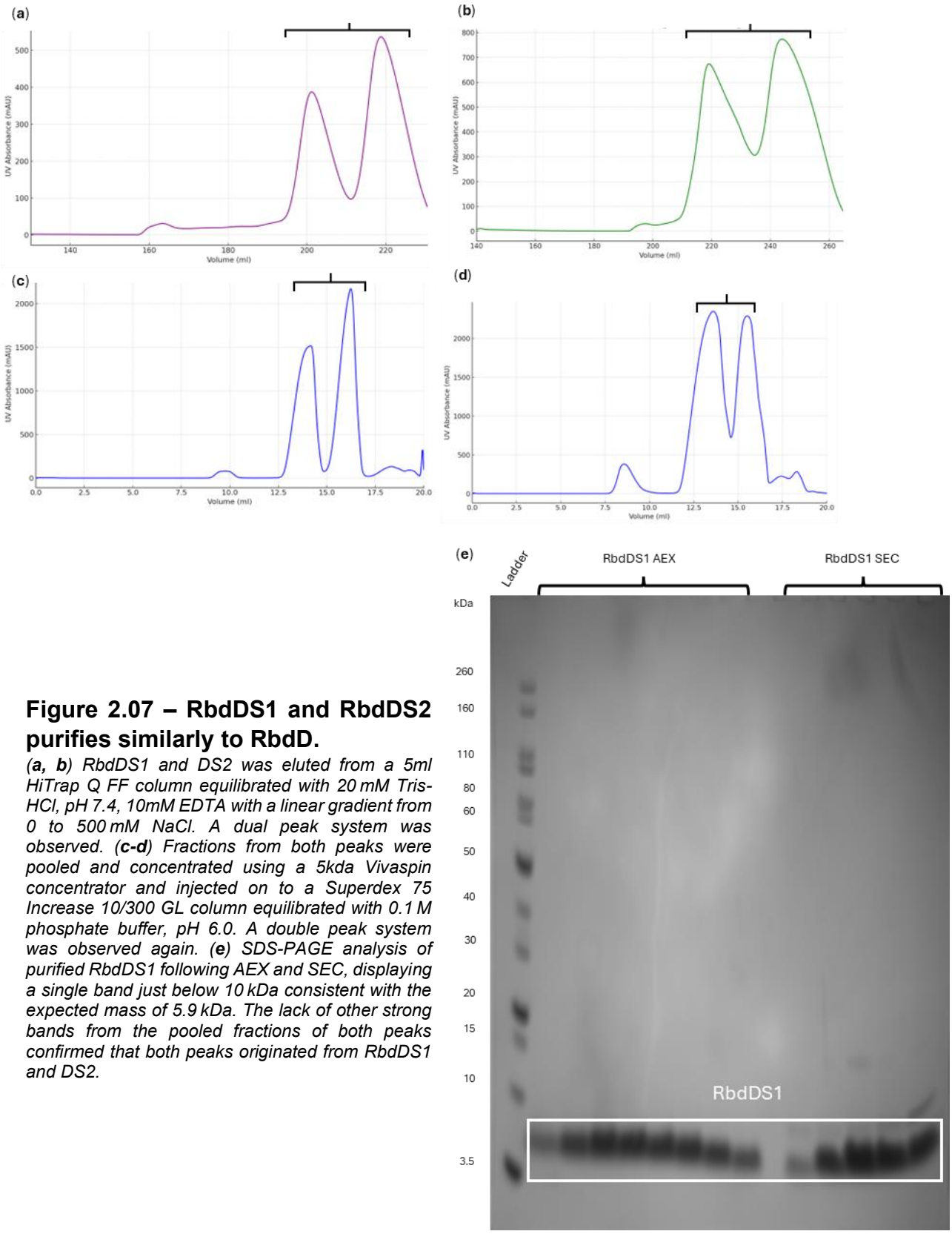


Figure 2.07 – RbdDS1 and RbdDS2 purifies similarly to RbdD.

*(**a, b**) RbdDS1 and DS2 was eluted from a 5ml HiTrap Q FF column equilibrated with 20 mM Tris-HCl, pH 7.4, 10mM EDTA with a linear gradient from 0 to 500 mM NaCl. A dual peak system was observed. (**c-d**) Fractions from both peaks were pooled and concentrated using a 5kda Vivaspin concentrator and injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1 M phosphate buffer, pH 6.0. A double peak system was observed again. (**e**) SDS-PAGE analysis of purified RbdDS1 following AEX and SEC, displaying a single band just below 10 kDa consistent with the expected mass of 5.9 kDa. The lack of other strong bands from the pooled fractions of both peaks confirmed that both peaks originated from RbdDS1 and DS2.*

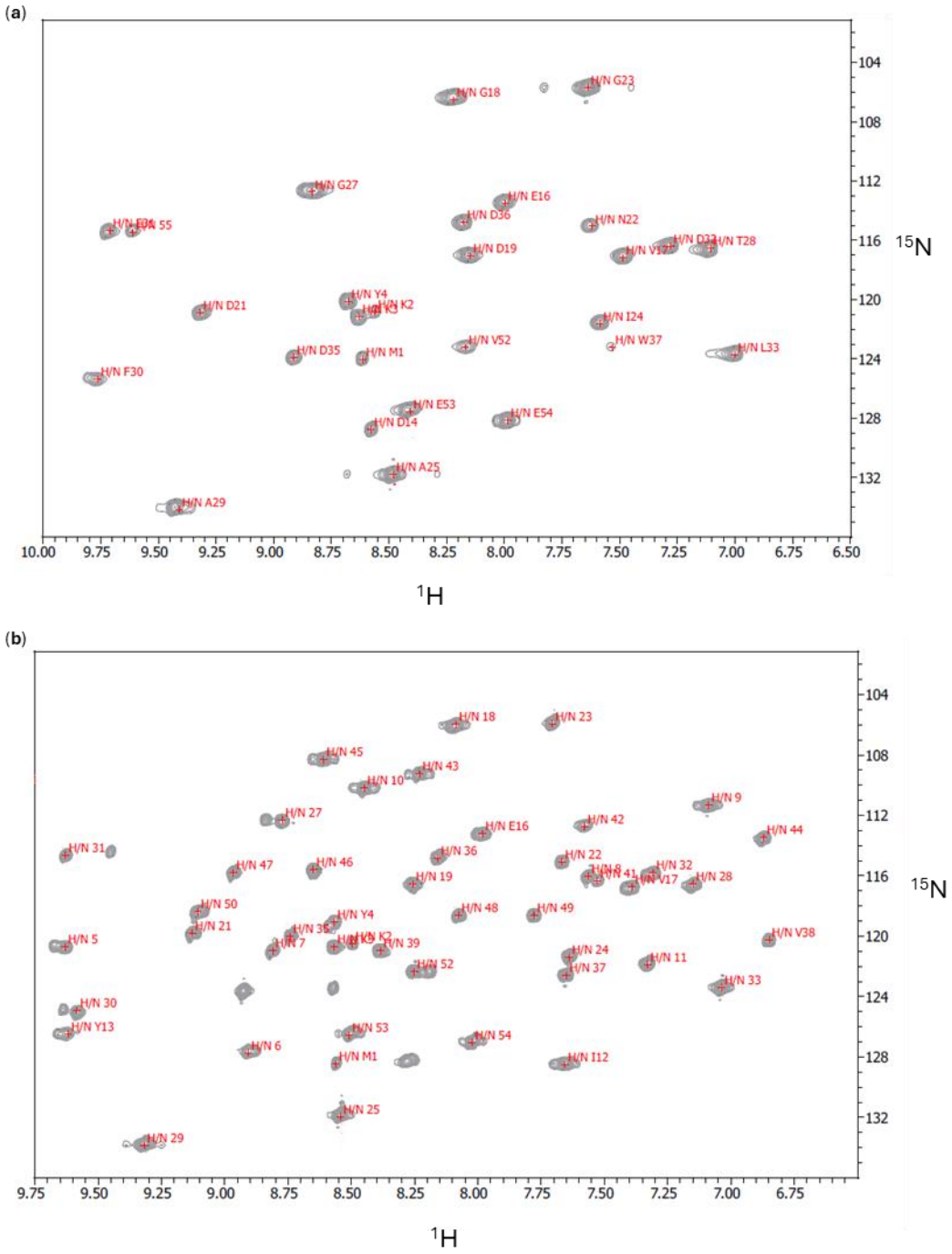


Figure 2.08 – Apo-RbdC is structured in solution.

*(**a**) A 2D projection of the assigned HNCA spectrum for Fe-Holo-RbdC is shown. Spectra were recorded at 20°C and pH 4.5 in unbuffered water with 12 scans per increment, 1022 total points (t_1), and 14.03ppm sweep width in the direct (^{1}H) dimension and 64 increments with a sweep width of 34.00ppm in the indirect (^{15}N) dimension. The reduced number of resonance signals reflects the paramagnetic broadening caused by the Fe^{3+}ion, consistent with previous observations by UV-vis of iron-sulfur cluster formation. (**b**) A 2D projection of the assigned HNCA spectrum for Apo-RbdC at 20°C in pH 4.5 unbuffered water. Apo-RbdC was prepared by removing Fe^{3+} from Fe-Holo-RbdC using three sequential trichloroacetic acid (TCA) precipitation steps. The protein pellet was washed with Milli-Q and 0.5M NaOH and resuspended in 8M urea containing 0.3M TCEP as a reducing agent. Following incubation on ice, the solution was dialysed overnight against unbuffered Milli-Q water at pH 4.5 (with 25mM TCEP). Data was acquired with 40 scans per increment, and 1276 t_1, and 12.02ppm in the ^{1}H dimension and 64 increments and 30.00ppm in the ^{15}N dimension. The spectrum reveals well-dispersed resonances, consistent with a folded protein. The greater number of peaks compared to Fe-Holo-RbdC confirms the absence of Fe^{3+} in the apo-protein.*

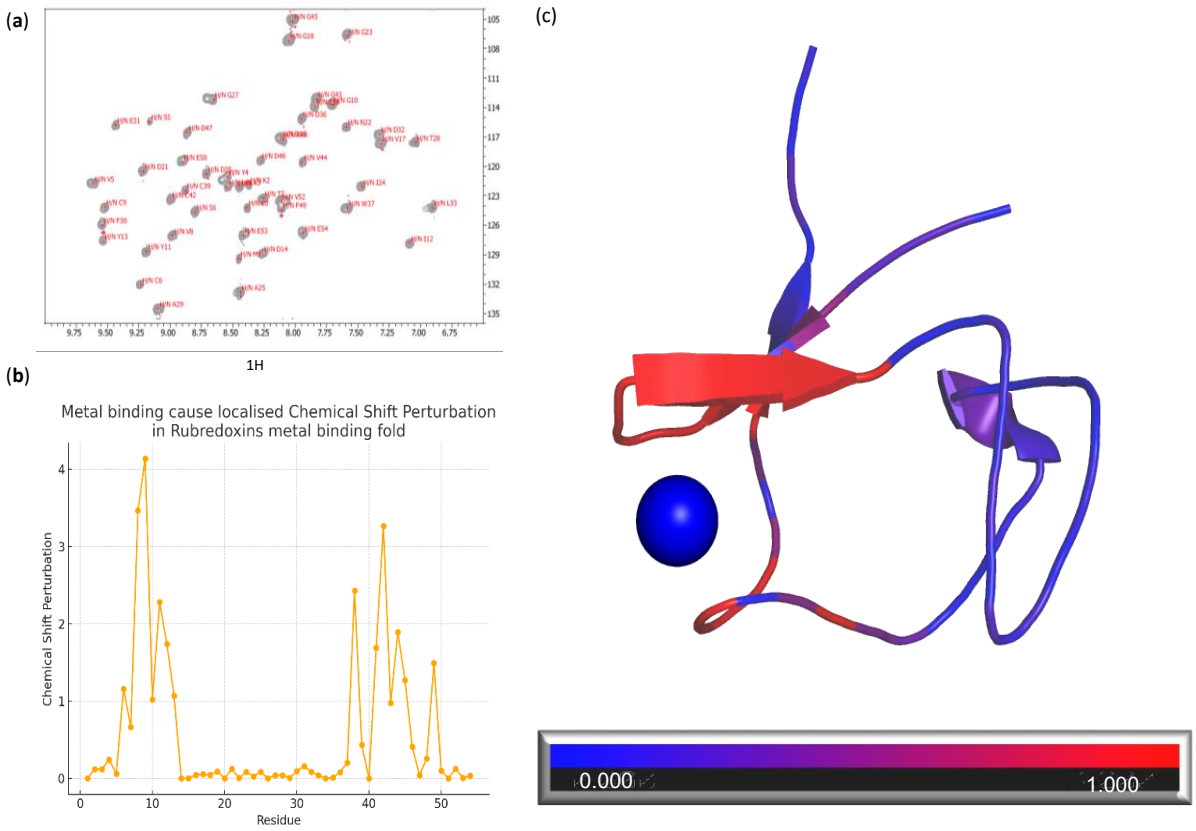


Figure 2.09 – Metal binding causes localised chemical shift perturbations in apo-RbdC.

*(**a**) 2D projection of the HNCA spectrum for Zn-Holo-RbdC at 20°C in pH 4.5 unbuffered water. Zn-Holo-RbdC was prepared by removing Fe^{3+} from Fe-Holo-RbdC using three sequential trichloroacetic acid (TCA) precipitation steps. The protein pellet was washed with Milli-Q and 0.5M NaOH and resuspended in 8M urea containing 0.3M β-Merceptaethanol and 20mM $ZnCl_2$. Following incubation on ice, the solution was dialysed overnight against unbuffered Milli-Q water at pH 4.5. Data was acquired with 12 scans per increment, 1022 total points (t_1), and 14.03ppm sweep width in the direct (^1H) dimension and 34.00ppm in the indirect (^{15}N) dimension. Chemical shifts were assigned in the CARA package. (**b**) Chemical shift perturbations were calculated from the formula proposed by* (Williamson, 2013) *of the Zn-Holo-RbdC compared to Apo-RbdC. (**c**) Chemical shift perturbations were mapped on the Alphafold3 predicted structure of Fe-Holo-RbdC using a custom script (**Appendix C.1**).*

Metal binding enhances rubredoxins thermal stability.

Differential scanning fluorimetry (DSC) revealed that apo rubredoxin unfolds cooperatively with a T_m of 66.6°C (**Fig 2.10**), which was increased by 9.6°C or 10.6°C upon binding of nickel or cobalt ions, respectively. The shape of the melting curves for zinc-bound and iron-bound rubredoxin suggest that the T_m values were at least 30˚C higher than for the apo protein. These results indicate that coordination to zinc or iron has a much greater effect on the stability of rubredoxin than nickel or cobalt, consistent with previous reports (Maher *et al.*, 2004).

Rubredoxin is well-folded in the apo-state.

NMR experiments showed that non-terminal regions of both apo and holo forms of RbdC had hetNOE values approaching 0.8 (**Fig 2.11a-b**), consistent with a rigid backbone structure independent of metal ion binding. While the higher average signal to noise in the holo dataset could indicate more flexibility in this state, given that some values exceeded 1, it is assessed that this was more likely caused by saturation artifacts. This data is consistent with apo-rubredoxin being structurally rigid with the metal-coordinating amino acids positioned close in space even in the absence metal binding and validates the rubredoxin as a useful system for testing whether cysteine-free variants are able to bind transition metals.

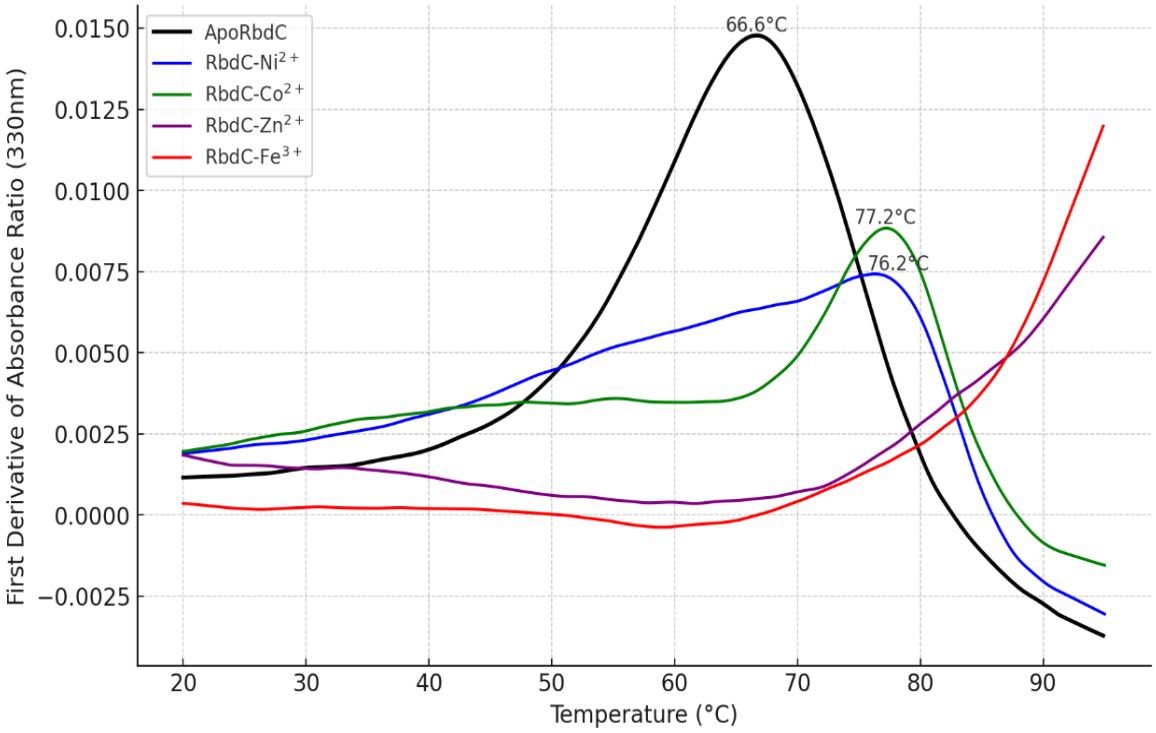


Figure 2.10 – Melting binding increases rubredoxin melting temperature.

The first derivative of fluorescence at 330nm vs temperature is shown for Apo-Rubredoxin (black) and Holo-Rubredoxin bound to Ni^{2+} (blue), Co^{2+} (green), Zn^{2+} (purple), and Fe^{3+} (red). Samples were prepared at 5 mg/mL in 0.1M phosphate buffer at pH 6.2 and subjected to thermal unfolding from 20°C to 95°C at 1°C/min. Holo-bound versions were generated by reconstitution with the respective metal ions (Ni^{2+}, Co^{2+} , Zn^{2+}) using purified Apo-Rubredoxin, except for the Fe^{3+} bound form, which was used as purified from E.coli expression. Apo-Rubredoxin exhibited a melting temperature (Tm) of 66.6°C and Ni^{2+} and Co^{2+} binding increased the Tm to 76.2°C and 77.2°C, respectively. Zn^{2+} and Fe^{3+} bound rubredoxin showed no clear melting transition within the tested temperature range, although the increasing signal up to and beyond 95°C is consistent with very high thermal stability.

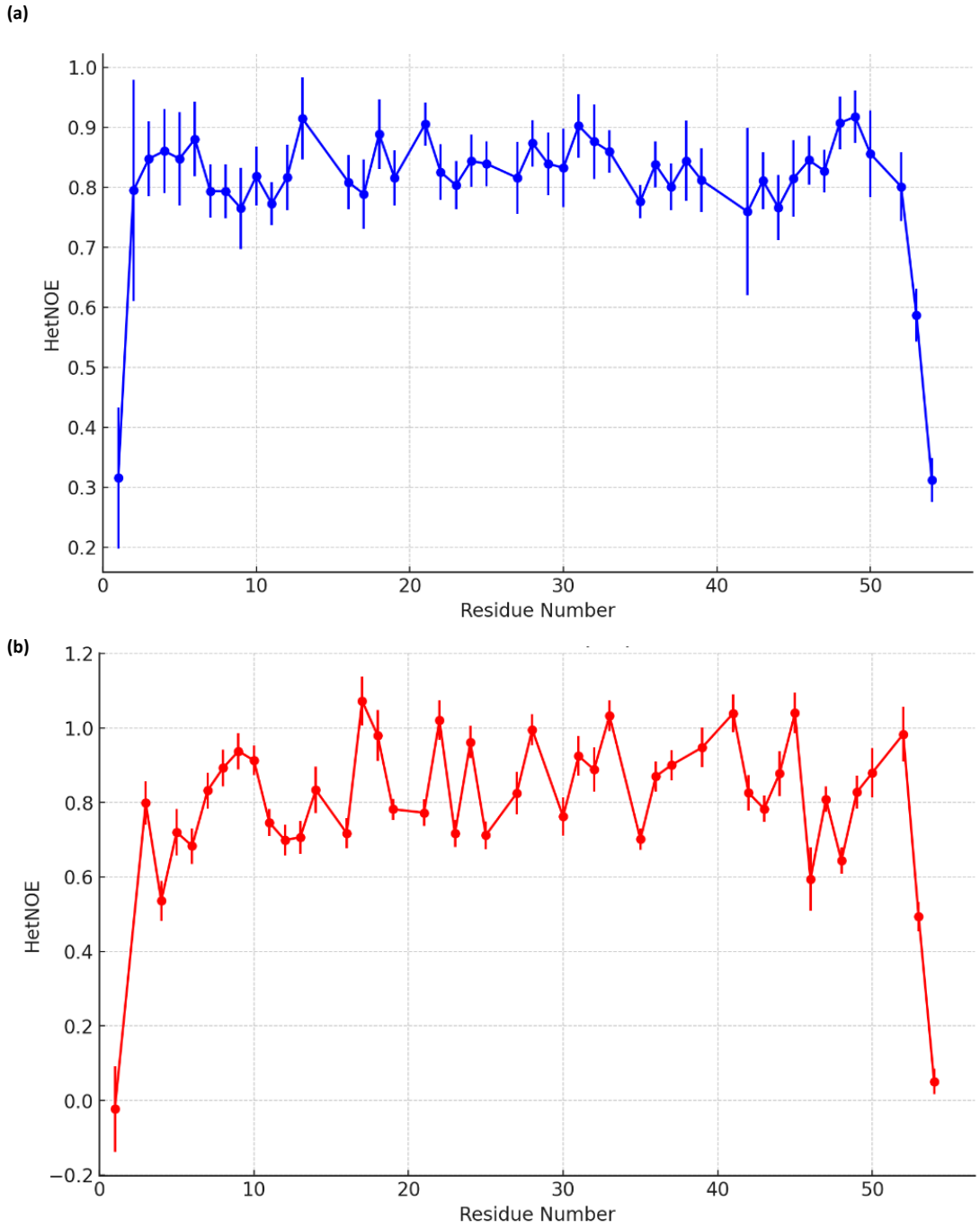


Figure 2.11 – Rubredoxin has limited flexibility in the apo and holo state.

*Heteronuclear NOE (hetNOE) values for 100μM Apo-RbdC (**a**) and 100μM Zn-Holo-RbdC (**b**) are shown across all assigned residues. Errors reflect calculated spectral noise scaled to unsaturated and saturated peak intensity. Measurements were conducted at 100μM in 90% H_2O/10% D_2O, pH 4.5, at 25°C using the hsqcnoef3gpsi pulse sequence with 2048 total points (t_1), and 14.00ppm sweep width in the direct (^1H) dimension and 120 increments with a sweep width of 16.00ppm in the indirect (^{15}N) dimension. Both apo and holo forms showed hetNOE values above 0.8 for most residues, consistent with a rigid structure and limited backbone flexibility. Some Zn-Holo-RbdC hetNOE values exceeded 1, which is attributable to saturation artifacts rather than increased flexibility.*

pH-dependent stability of RbdD

Unlike Apo-RbdC and RbdDS1 and RbdDS2, RbdD displayed a pH dependent folding behaviour. Further investigation of this was needed to ensure that it was folded for metal binding screens.

At pH 4.5, RbdD displayed a well-dispersed ^{1}H-^{15}N HSQC spectrum consistent with a well-folded protein (**Fig 2.12a**). By contrast, at pH 7.0, the spectrum exhibited significant peak collapse, indicative of unfolding (**Fig 2.12b**). Comparison of NMR assignments at pH 4.5 and pH 7.0 revealed large CSPs across the entire protein (**Fig 2.12c**). Mapping these CSPs onto the rubredoxin structure demonstrated that the unfolding event affected all residues (**Fig 2.12d**), strongly indicating that RbdD undergoes a pH-dependent global unfolding.

The pH-dependent unfolding was explored further by CD. CD spectra of RbdD recorded from pH 3.5 to pH 7.0 indicated loss of β-sheet content and unfolding of the protein at pH values above 6.0, which was not observed in Fe-Holo-RbdC at the same pH (**Fig 2.12e**). Thermal unfolding experiments showed that increasing the pH from 4.5 to 6.2 in RbdD resulted in a decrease in Tm by 26°C (**Fig 2.12f**). This substantial reduction in thermal stability correlated with the loss of structural integrity observed in the NMR spectra.

Further investigation of RbdD by ^{1}H-^{15}N HSQC spectra revealed that the mutant at intermediate pH exhibits peaks from both the folded and unfolded states (**Fig 2.13a**). The presence of both states is particularly clear for the ^{1}H-^{15}N resonance of Ala_{25}, which is a surface residue and one of the residues furthest from the metal binding site. The Ala_{25} crosspeak shows a relatively modest but

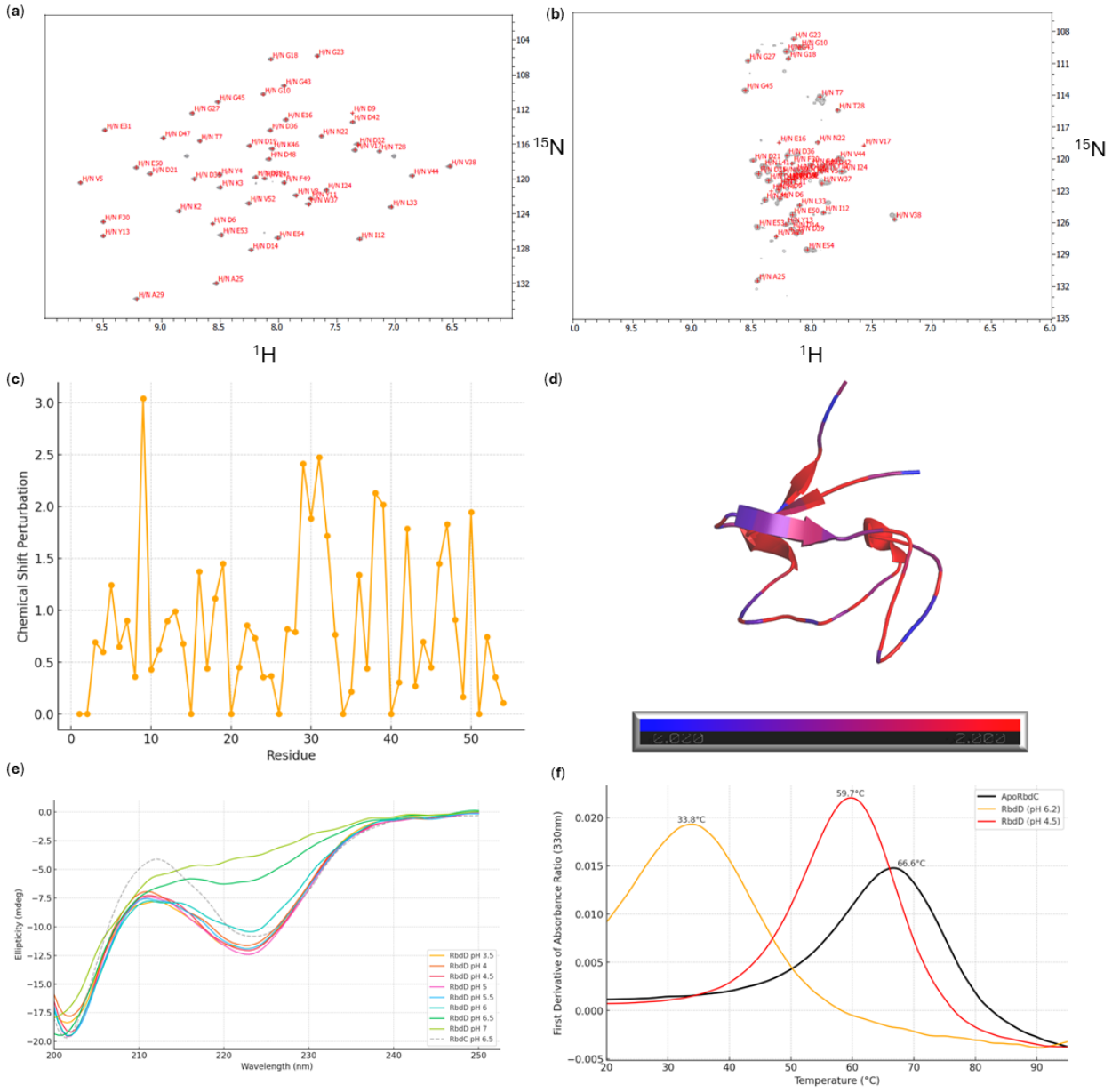


Figure 2.12 – RbdD undergoes global unfolding at neutral pH.

*(**a**) A 2D projection of the HNCA spectrum for 1mM RbdD, recorded at 20°C and pH 4.5 in unbuffered water. Data was acquired with 12 scans per increment, 1022 total points (t_1), and 14.03ppm sweep width in the direct (^{1}H) dimension and 64 increments with a sweep width of 34.00ppm in the indirect (^{15}N) dimension. (**b**) 2D projection of the HNCA spectrum for RbdD recorded at 20°C and pH 7 as before. Chemical shifts were assigned in the CARA package. (**c**) Chemical shift perturbations were calculated from the formula proposed by* (Williamson, 2013) *of RbdD at pH 7 compared to RbdD at pH 4.5. (**d**) Chemical shift perturbations were mapped on the Alphafold3 predicted structure of RbdD using a custom script (**Appendix C.1**) (**e**) CD spectra of the eluted protein concentrated to 2.5mg/ml at pH 4-7 in 0.5pH increments was recorded using a Jasco J-815 CD spectropolarimeter at 20°C. The spectrum shows a loss of ellipticity signal at 222nm as a function of pH from pH 6-pH 7, consistent with the loss of alpha helical structural elements (**f**) The first derivative of fluorescence at 330 nm vs temperature is shown for Apo-Rubredoxin (black) and RbdD at pH 4.5 (red) and pH 7.0 (orange). Samples were prepared at 5 mg/mL in 0.1M phosphate buffer at pH 6.2 and subjected to thermal unfolding from 20°C to 95°C at 1°C/min.*

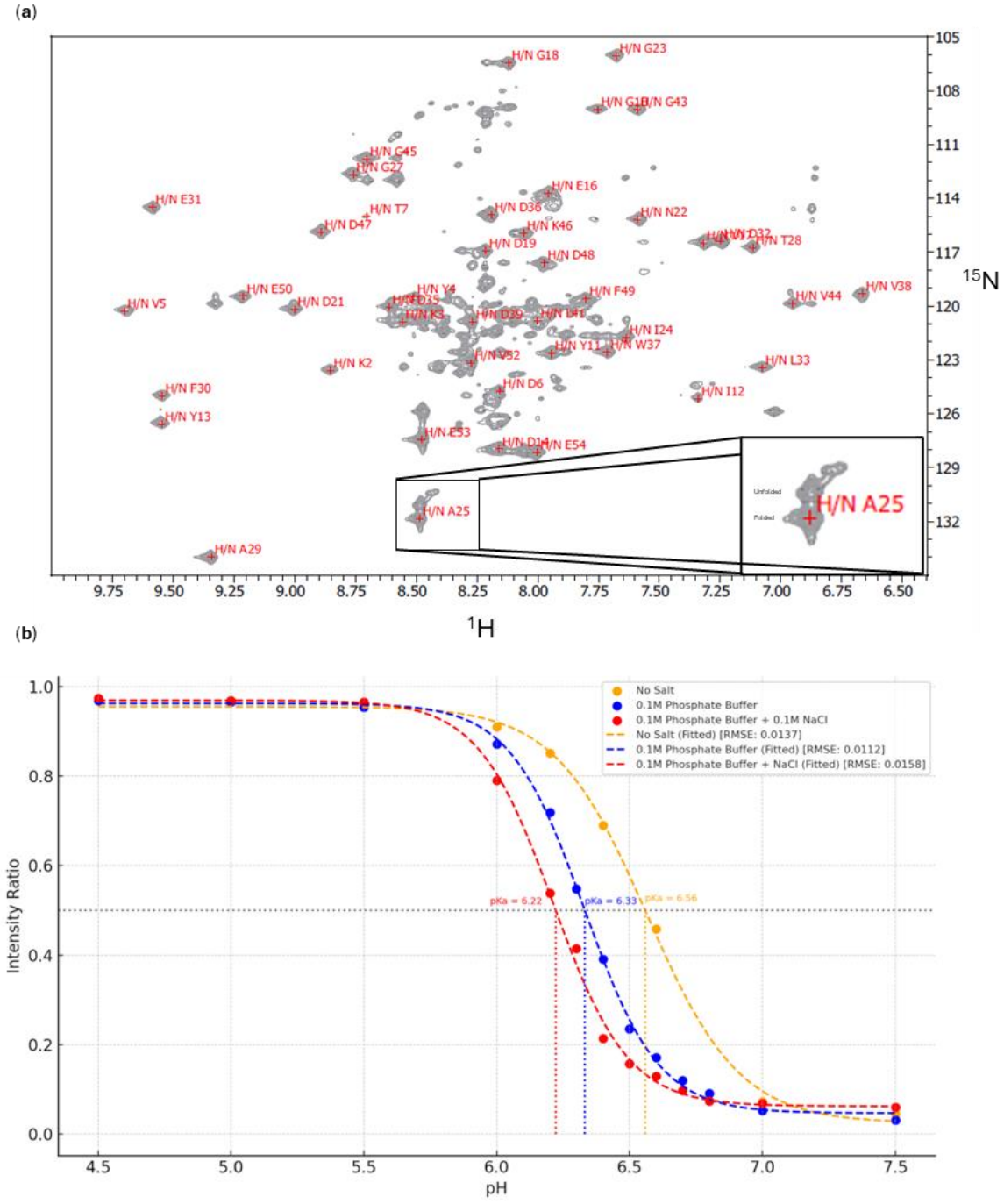


Figure 2.13 – RbdD shows a salt dependent titration curve.

*(**a**) 2D ^{1}H-^{15}N HSQC spectrum of 1mM RbdD, recorded at 20°C and pH 6.4 in unbuffered water. Data was acquired with 8 scans per increment, 2048 total points (t_1), and 16.02ppm sweep width in the direct (^{1}H) dimension and 128 increments with a sweep width of 34.00ppm in the indirect (^{15}N) dimension. Assignments of RbdD at pH 4.5 are shown for reference (**b**) The peak height intensity of the folded Ala_{25} and unfolded Ala_{25} across identically ran ^{1}H-^{15}N HSQC from a pH range of 4.5-7.5 were divided to create an intensity ratio of the folded state over the unfolded state across the range. The resulting ratios were recorded at 1.0mg/ml RbdD samples equilibrated in unbuffered water (yellow), 0.1M phosphate buffer (blue) and 0.1M phosphate buffer, 0.1M NaCl (red). The intensity ratios were plotted using Python and a curve was fitted with the pH calculated at 0.5 intensity ratio on the fitted curve reported as the pKu. The shift in pKu with increasing salinity suggests that aspartate protonation occurs more readily as salinity increases.*

resolvable CSP upon unfolding, and the unfolding transition midpoint, whereby 50% of the protein was folded, and 50% unfolded, on the pH scale (pK_u) could be determined from the ratio of intensities of the folded and unfolded Ala_{25} crosspeaks as a function of pH.

Since apo-rubredoxin is stably folded with the coordinating cysteines close in space, I hypothesised that charge repulsion between the ionised aspartate residues destabilises the RbdD fold at neutral pH. The pK_U was observed to decrease with increasing ionic strength (**Fig 2.13b**), consistent with the presence of unfavourable electrostatic interactions in folded RbdD at higher pH.

While Ala_{25} is useful to determine unfolding of the protein hydrophobic core, which Ala_{25} is positioned next to, chemical shift analysis of Asp_9 in RbdD was better suited to probe the protonation state of the aspartate side chains within the metal binding site itself. A significant CSP in the Asp_9 crosspeak was observed between pH 6.2 and pH 6.4 (**Fig 2.14**), which is attributable to deprotonation of the carboxylate sidechain and fully consistent with the pK_U of 6.33 determined from Ala_{25}. This correlation between the deprotonation event of Asp_9 and the observed structural collapse indicates that sequential deprotonation of the aspartate residues in the metal-binding site directly causes charge-charge repulsion that likely drives the global protein unfolding that was observed. This also suggests that when the protein is folded, the aspartate residues are also protonated, making them far less likely to be capable of metal coordination. In the context of the thesis objective to define abiotic protein function, the observed structural collapse of RbdD at near-neutral pH demonstrates a potential critical limitation of aspartate as a substitute for cysteine. This limitation may potentially be

generalisable to any abiotic protein where aspartates are positioned in this proximity within a transition metal binding compatible pore. Conversely however, these results do not preclude the formation of a metal binding site between short aspartate enriched peptides, which inherently have more positional flexibility to achieve more stable binding topologies than in preformed metal binding sites within a protein.

Structural Characterisation of RbdDS1 and RbdDS2

RbdDS1 and RbdDS2 were assigned and displayed well-dispersed ^{1}H-^{15}N HSQC spectra at pH 6.2, consistent with folded proteins (**Fig 2.15a-b**). The presence of sharp, well-resolved dispersed peaks suggested that these variants maintained structural integrity unlike RbdD at this pH. CSP analysis comparing RbdDS1 and RbdDS2 at pH 6.2 to RbdD at pH 4.5, showed significant perturbations around the metal-binding site but minimal changes in the hydrophobic core region (residues 13-37) (**Fig 2.15c**). This was consistent with the substitutions affecting the local environments near the mutations, likely due to the changes in charge states of the ligands themselves between variants, without significant disruption to the overall fold of the protein and ensuring the coordinating residues are close in space within the metal binding loop. Similar CSP calculations with RbdDS1 and DS2 at pH 4.7 revealed similar perturbations, consistent with the aspartates of DS1 and DS2 being deprotonated to at least pH 4.7 and above, similar to aspartate's typical pKa_3 (**Appendix A - Figure 02**).

Metal binding screening of rubredoxin variants

Zn-holo-RbdC and its variants, RbdD, RbdDS1, and RbdDS2 were screened for metal binding using ^{15}N-^{1}H HSQC spectroscopy upon addition of

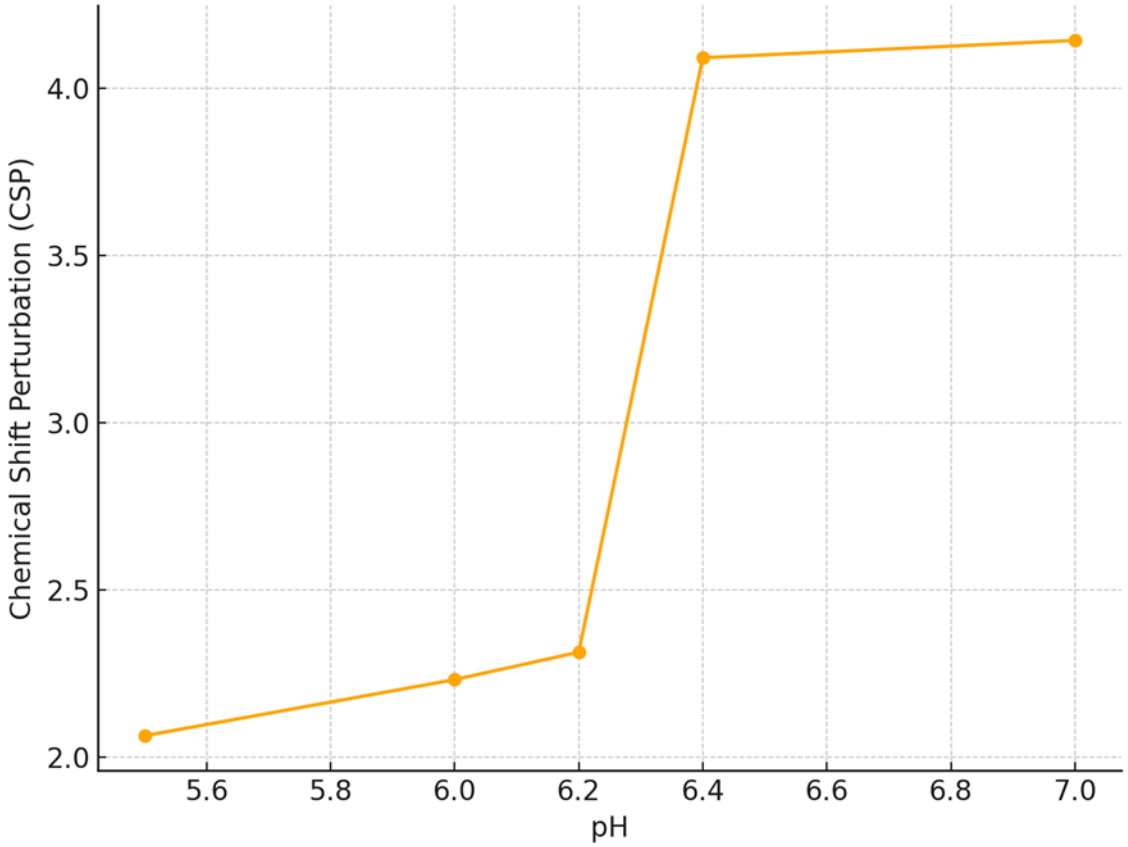


Figure 2.14 – RbdD aspartate deprotonation drives protein unfolding.

Several 2D ^{1}H-^{15}N HSQC spectrums of 1mM RbdD, was collected at 20°C in 0.1M phosphate buffer at varying from pH 4-8 with 8 scans per increment, 2048 total points (t_1), and 16.02ppm sweep width in the direct (^{1}H) dimension and 128 increments with a sweep width of 34.00ppm in the indirect (^{15}N) dimension. Chemical shift perturbations were calculated for Asp_9 per the formula proposed by (Williamson, 2013)*. A large chemical shift perturbation is observed for Asp_9 between pH 6.2 and 6.4 with deprotonation of the Asp_9 side chain consistent with the reported pKa by Ala_{25} for 6.33 at this salinity range (**Fig 2.13**). Deprotonation of the aspartates within the metal binding site causes protein unfolding.*

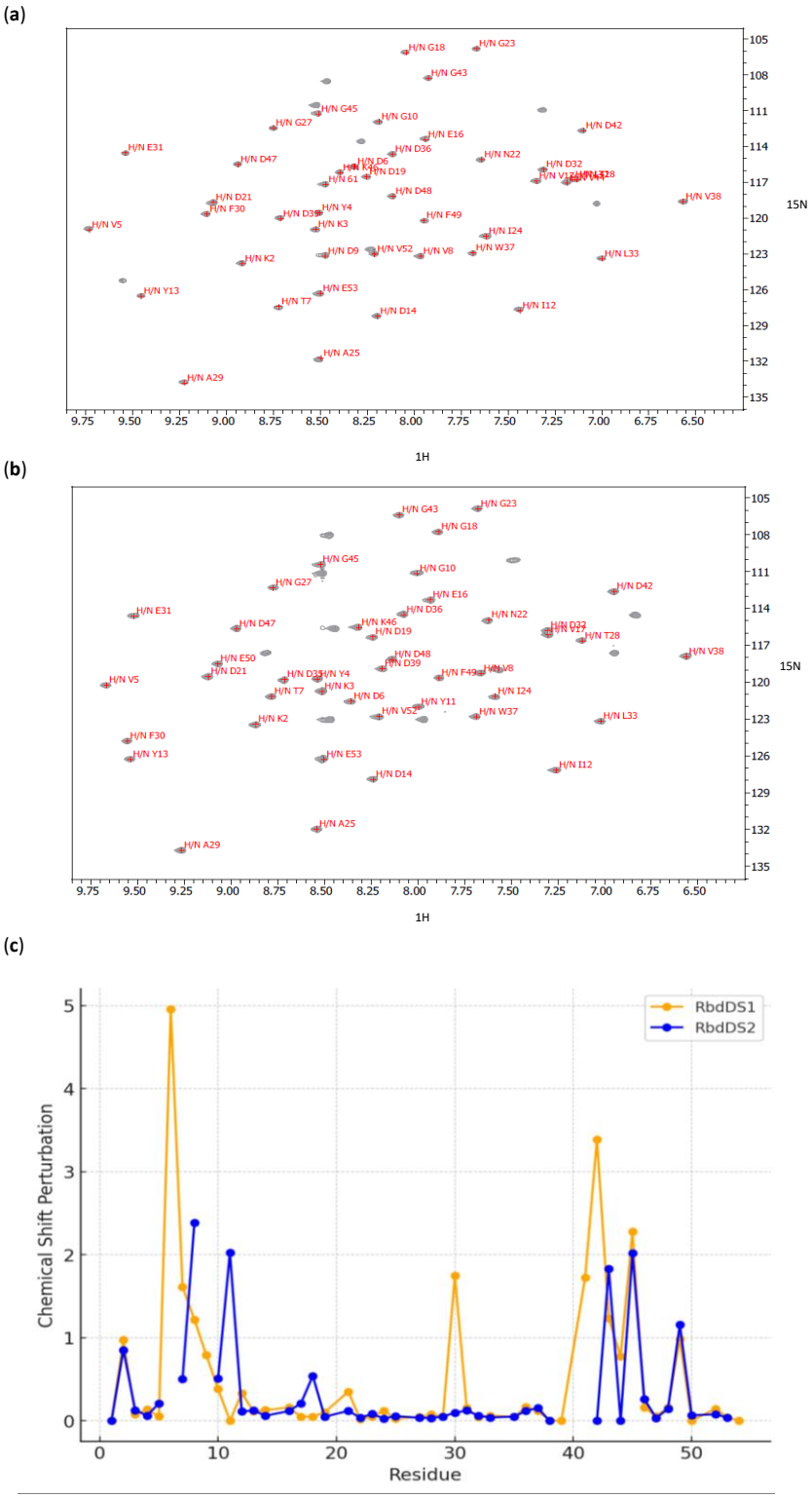


Figure 2.15 – RbdDS1 and DS2 remain folded at pH 6.2.

*(**a**) 2D projection of the HNCA spectrum for 1mM RbdDS1 was recorded at 20°C and pH 6.2 in unbuffered water. Data was acquired with 40 scans per increment, 1276 total points (t_1), and 12.00ppm sweep width in the direct (^{1}H) dimension and 64 increments with a sweep width of 30.00ppm in the indirect (^{15}N) dimension. (**b**) 2D projection of the HNCA spectrum for 1mM RbdDS2 recorded at 20°C and pH 6.2 as before. Chemical shifts were assigned in the CARA package. Observed spectra were consistent with a folded protein given the uniform distribution of peaks (**c**) Chemical shift perturbations were calculated from the formula proposed by* (Williamson, 2013)*, of RbdDS1 (yellow) and RbdDS2 (blue) compared to RbdD at pH 4.5.*

metal ions (Zn^{2+}, Mg^{2+}, Co^{2+}, Fe^{2+}, Ni^{2+}) at a 2:1 metal-to-protein molar ratio at pH 4.7, one pH unit above the pKa_3 of aspartate in solution (pH 3.5-4.0). No significant CSPs at the metal binding site were observed for the variants that were consistent with apo to holo folding of the protein (**Fig 2.09b**). This was, especially so when compared against the apo-holo CSP's observed upon successful melting binding of RbdC, which resulted in localised CSPs of 0.5-4.1 (**Fig 2.16a-d**). This was consistent with the absence of any new strong metal coordination in the rubredoxin variants. At an increased metal-to-protein ratio of 10:1, some CSPs were observed, however, these perturbations were not localised to the metal-binding site but rather to surface-exposed residues (**Fig 2.17a-d**). Additional experiments with RbdD and zinc at a higher pH (pH 6.2) did not yield significant CSPs (**Appendix A - Figure 01**). Mapping the CSPs onto the rubredoxin structure confirmed that the affected residues were distinct from the canonical metal-binding site (**Fig 2.17e**). These findings indicated that the observed CSPs were likely due to nonspecific interactions rather than specific metal binding. The lack of specific metal binding by all variants suggests that the aspartate and serine sidechains cannot replace the metal binding function of the cysteine thiol even when those residues are close in space.

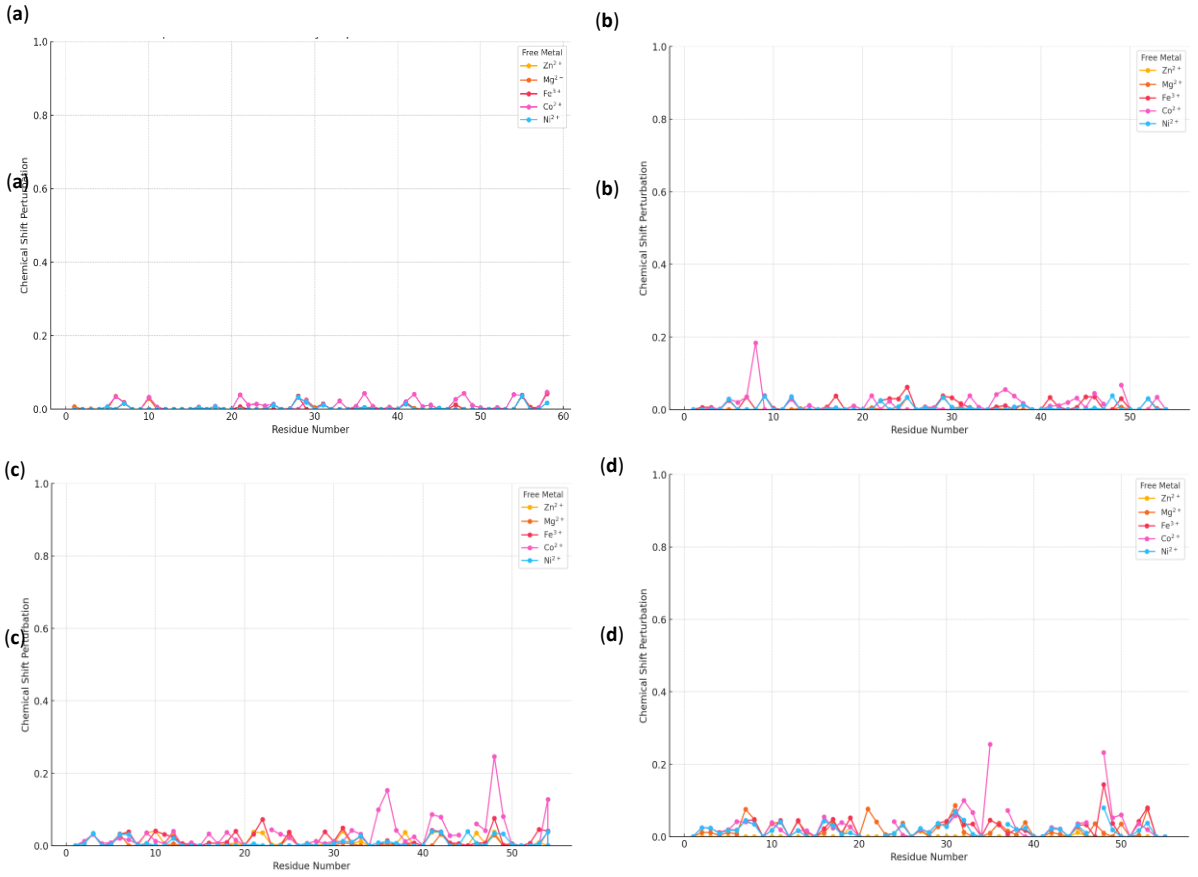
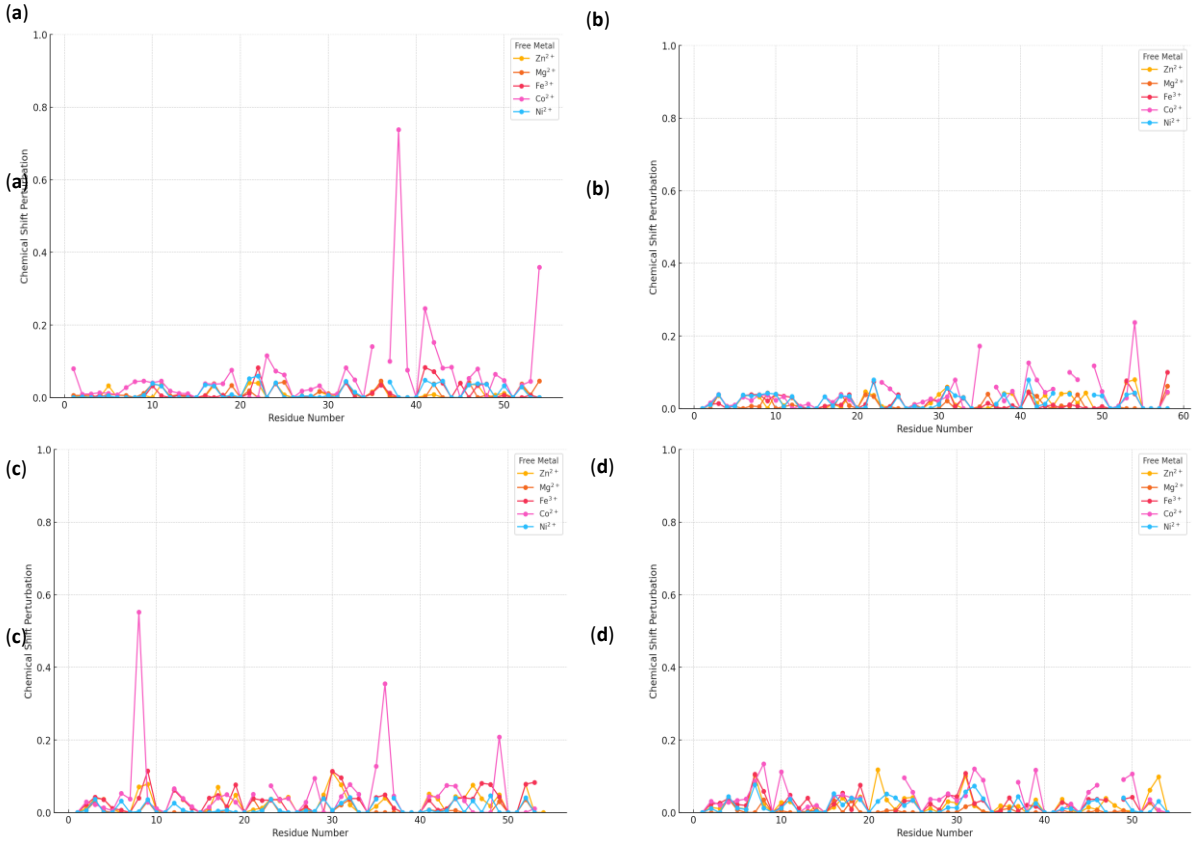
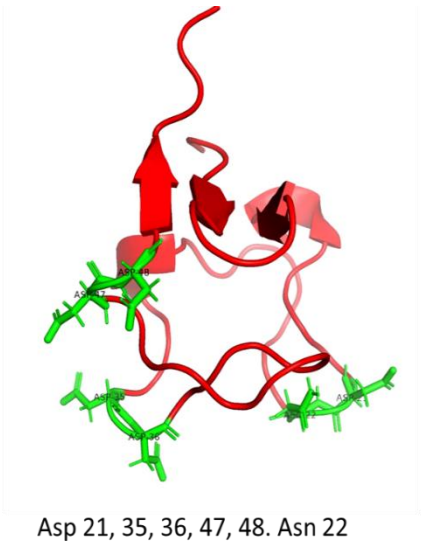


Figure 2.16 – Rubredoxin variants do not bind to transition metals at 2:1 ratio.
*2D ^{1}H-^{15}N HSQC spectrum of (**a**) 80μM Zn-Holo-RbdC, (**b**) 635μM RbdD, (**c**) 1004μM RbdDS1 and (**d**) 1812μM RbdDS1, equilibrated in 100% N_2 overnight, were recorded at 20°C against 2:1 molar ratios of Zn^{2+} (yellow), Mg^{2+} (orange), Fe^{2+} (red), Co^{2+} (pink) and Ni^{2+} (blue) chlorides with a final pH of 4.7±0.15 in unbuffered water. Data was acquired with 32 scans per increment, 2048 total points (t_1), and 15.95ppm sweep width in the direct (^{1}H) dimension and 256 increments with a sweep width of 32.07ppm in the indirect (^{15}N) dimension. While some CSP's were visible, none were consistent with the CSP's observed for the apo-holo transition in RbdC.*





Asp 21, 35, 36, 47, 48. Asn 22

Figure 2.17 – Rubredoxin variants exhibit surface localised CSP's at 10:1 ligand to protein ratios.

*2D ^{1}H-^{15}N HSQC spectrum of (**a**) 80μM Zn-Holo-RbdC, (**b**) 635μM RbdD, (**c**) 1004μM RbdDS1 and (**d**) 1812μM RbdDS1, equilibrated in 100% N_2 overnight, were recorded at 20°C against 10:1 molar ratios of Zn^{2+} (yellow), Mg^{2+} (orange), Fe^{2+} (red), Co^{2+} (pink) and Ni^{2+} (blue) chlorides with a final pH of 4.7±0.15 in unbuffered water. Data was acquired with 32 scans per increment, 2048 total points (t_1), and 15.95ppm sweep width in the direct (^{1}H) dimension and 256 increments with a sweep width of 32.07ppm in the indirect (^{15}N) dimension. CSP's were again not consistent with the CSP's observed for the apo-holo transition in RbdC. (**e**) Common CSP's were visualised on an Alphafold3 RbdD predicted structure, showing the CSP's were localised to surface charge groups and were not consistent with rubredoxin metal binding*

Discussion

In Chapter 1, I outlined the hypothesis that metal ion coordination might represent a universal mechanism enabling the folding and activity of early proteins before the full complement of modern amino acids was available. A test of this hypothesis for the consensus abiotic amino acid set lacking cysteine, is whether stable tetrahedral transition metal-binding sites can form using just this set. In this study, I addressed this question by substituting cysteine residues at the metal-binding site of rubredoxin with aspartate and serine and evaluating their ability to coordinate transition metals within a folded apo-protein with a well-defined transition metal binding site.

Assignment of apo-rubredoxin

A significant first step in this study was the first successful assignment of an apo-rubredoxin. A previously published NMR study of another rubredoxin reported that the apo form was isolated at 20% yield, which was stated to be similar to the zinc bound structure, while 80% was consistent with an unfolded apo state (Gondim *et al.*, 2019). Despite relatively high sequence similarity to that rubredoxin (76%), I observed that the apo form of RbdC exhibited well-dispersed NMR spectra, consistent with a fully folded apo state, and providing a critical baseline for evaluating structural changes upon metal binding. I attributed this success to the use of an extremely large excess of TCEP reducing agent at the apo-creation step and during NMR analysis which prevented the creation of disulfide linkages that likely caused the previously reported lack of structure in apo rubredoxins. In addition, the structural integrity in the apo state allowed me to identify and assign CSPs corresponding to the apo-to-holo transition in metal-

bound RbdC. This enabled definition of the features of successful metal binding in rubredoxin which set me up to confidently identify the absence of metal binding in the cysteine-free variants in later experiments. These results agree and build upon, the small population of folded apo rubredoxin observed in experiments by Gondim *et al*, and those reported by Almeida (Gondim *et al.*, 2019)(Almeida *et al.*, 2021)

pH-dependent folding in RbdD

Another key finding of this study was the pH sensitivity of the RbdD variant, which exhibited folding behaviour strongly dependent on environmental pH. At low pH (pH 4.5), RbdD maintained a well-dispersed NMR spectrum consistent with a folded protein. However, significant spectral narrowing indicated global unfolding at pH 7.0. Chemical shift analysis revealed that this structural collapse was driven by the deprotonation of closely spaced aspartate residues within the metal-binding site, resulting in charge-charge repulsion that destabilised the protein fold.

This observation is potentially relevant to any abiotic metal-binding site reliant on aspartates, as the unfolding behaviour results from the physical proximity of these residues in the deprotonated state when not bound to a ligand. While this might suggest that such sites could be more stable at acidic pH where the aspartates remain protonated, protonation significantly reduces the nucleophilicity of the carboxylate group, limiting its ability to interact with transition metals. The high charge density of deprotonated aspartates in tetrahedral coordination geometries may also explain why such sites are not observed in transition metal binding sites in modern proteins where such high charge

densities do not favour transition metal binding (Dokmanić *et al.*, 2008)(Permyakov, 2021). This issue does not apply to coordination sites using cysteine, whose thiol group lacks significant protonation/deprotonation dynamics below pH 8.1. This charge stability likely drove the evolutionary adoption of cysteine as the key ligand for transition metal coordination in tetrahedral sites where charge density is inherently a greater problem than in octahedral sites due to the smaller coordination area.

Metal on coordination in the absence of cysteine

Cysteine-free variants of rubredoxin successfully folded into structures consistent with the apo and holo forms, indicating that the metal-binding residues were correctly positioned. However, none of the variants coordinated biologically relevant transition metals, such as iron, zinc, cobalt, or nickel, even at a 10:1 molar ratio. Instead, observed interactions were weak and consistent with nonspecific surface interactions rather than the formation of stable, tetrahedral metal-binding geometries.

These results are in alignment with those expected from the Hard and Soft Acids and Bases (HSAB) principle, which postulates that transition metals preferentially bind to thiol groups over carboxylate groups (Pearson, 1963). It was also consistent with crystal field theory, whereby weak field ligands, like carboxylates, do not provide sufficient field splitting for coordination of iron unless at high coordination numbers where ligand quantity compensates for their limited field strength. Despite proper positioning, the carboxylate and hydroxyl moieties of aspartate and serine lacked the nucleophilicity required to replicate cysteine's function when coordination transition metal ions. This indicates that in an abiotic

context, the absence of cysteine would have significantly constrained the emergence of metalloprotein architectures reliant on transition metals.

Supporting these findings, Jordan *et al* investigated the capability of cysteine to spontaneously coordinate iron in the presence of sodium sulfide at alkaline pH to create dissolved Fe_2S_2 and Fe_4S_4 clusters like those seen in Ferredoxin (Jordan *et al.*, 2021). They demonstrated that monomeric cysteine can spontaneously coordinate iron in the presence of sodium sulfide under alkaline conditions, forming Fe_2S_2 and Fe_4S_4 clusters like those in ferredoxin. Consistent with these results, when I conducted qualitative replicates of Jordan *et al*'s experiments, I observed rapid precipitation of FeS clusters when aspartate replaced cysteine (**Appendix A - Figure 02**).

As such, the results presented here align with those of Jordan *et al* and together suggest that the unique thiol chemistry of cysteine provides an irreplaceable capability to enabling the formation of stable Fe-S clusters in abiotic conditions (Jordan *et al.*, 2021). Given that Fe-S clusters, and abiotic transition metal binding sites, may require the presence of cysteine to form, this strongly suggests that niches where cysteines synthesis is possible provide a constraint on where protein-based abiogenesis could occur given how essential metal coordination is to proto-metabolism (Goldford *et al.*, 2019).

Limitations and model constraints

While rubredoxin's simple mononuclear iron-cysteine cluster provides a well-characterised system for probing metal coordination, it represents just one potential solution to metalloprotein formation. This model does not and cannot

encompass the full diversity of early protein folds or environmental factors that may have influenced metal-protein interactions, such as mineral surfaces or phase-separated environments. Additionally, this study focuses exclusively on tetrahedral coordination geometries, which, while common in small metalloproteins, do not represent the full range of biologically relevant metal-binding architectures. Notably the RbdD mutant did not bind Mg^{2+}, which per the HSAB principle RbdD should have been capable of coordinating to. This is most simply explained by rubredoxins limited capability to support octahedral metal coordination geometries which Mg^{2+} strongly favours. Future investigations could expand on this work by examining whether abiotic amino acid residues, such as aspartate or serine, can facilitate stable metal binding within these alternative coordination geometries.

One additional limitation of this study is my exclusive focus on replacing cysteine while omitting the histidine imidazole group. There is evidence that abiotic pathways for imidazole synthesis may exist; for example, Vázquez-Salazar et al. proposed a condensation mechanism involving Ala-Ser-Gly peptides (Vázquez-Salazar *et al.*, 2017). However, as the authors themselves acknowledge, their relatively complex multistep mechanism of synthesis does not resolve, and is perhaps related to, the fact that histidine has not been detected in abiotic environments. This makes the role of histidine in early protein evolution much more speculative. Furthermore, histidine synthesis requires conditions and intermediates that may be less plausible in the abiotic context compared to thiol-based amino acids like cysteine, which have clearer geochemical plausibility in

sulfur-rich environments from the conversion of serine to cysteine (Foden *et al.*, 2020).

The limitations of AlphaFold were also directly evident here. Despite high overall confidence scores for the ligand bound Alphafold3 predictions of RbdD, DS1 and DS2, none of these proteins were observed in the metal bound state in solution. This finding is consistent with the algorithm's reliance on previously resolved protein structures in its training data and its inability to explicitly model metal coordination environments during its prediction pathway.

An additional limitation arose in our protein choice that became evidence only after the analysis of the Apo-RbdC variant. hetNOE data revealed that the RbdC displayed a rigid backbone even in the apo state, contrary to my initial predictions of increased flexibility in the absence of the iron sulfur site which I believed would in principle make the variants sufficiently flexible to both intake and coordinate the metal ions of interest. This rigidity may have constrained the ability of aspartate, glutamate, or serine to reorganise the binding site for effective metal coordination as well as acquire a metal ion from the solution. As a result, the failure of these residues to support metal binding in the rubredoxin context may reflect the limitations of the scaffold, rather than a general property of the amino acids themselves. For a fair test of alternative ligation strategies in future studies, it will be necessary to use a model system with greater conformational flexibility in the apo state. Equally however when measured on balance, it is my assessment that it was the limited ligand field stabilisation offered by the carboxylate or hydroxy ligands in the tetrahedral coordination geometry which most significantly affected binding potential.

Conclusion

These findings demonstrate that cysteine likely plays a unique role in enabling the early evolution of metalloproteins, supporting the hypothesis that abiogenesis may have been favoured in sulfur-rich environments. This work directly supports Hypothesis 1, which indicates that cysteine is a particularly effective route for stable transition metal coordination under plausible prebiotic conditions. In the context of Aim 2 (modelling abiotic-like proteins using only abiotic amino acids) and Aim 4 (considering cysteine's impact on abiotic protein function), the present results suggest that aspartate, glutamate, and serine are significantly less effective at supporting transition metal binding in the absence of a thiol-bearing sidechain.

However, it should be emphasised that the "simple substitution" strategy employed here, where cysteine was replaced by alternative residues within the same rubredoxin scaffold, may not fully capture the range of possible binding environments that could arise in different protein folds or through more extensive sequence redesign. The observed rigidity of the rubredoxin fold and the limited scope of the substitutions constrain the conclusions that can be drawn about the broader potential for non-thiol ligands in transition metal binding. Alternative approaches, such as de novo design or the use of entirely different starting folds, may be necessary to fully test the ability of abiotic amino acids to stabilise metal centres. For this reason, the findings should be interpreted as a case study within a specific structural context, rather than a universal constraint on all possible protein architectures.

Chapter 3: Abiotic protein stability and solubility

Introduction

Study rationale

Proteins synthesised abiotically or by very early life would have been composed exclusively from abiotically available amino acids. In Chapter 1, I discussed potential challenges to the folding, and thus function, of these proteins. Previous studies suggest that abiotic amino acids constitute a 'foldable set', with sufficient side chain chemistry to enable protein folding and function, however their experimental systems retained key 'biotic' positively charged and aromatic residues (Longo *et al.*, 2013). The charge interactions and aromatic hydrophobic interactions retained in these proteins was likely the cause of the observed folding capability, and it has been previously noted that such charge interactions are not abiotically plausible (McDonald and Storrie-Lombardi, 2010). Without charged and aromatic amino acids, the abiotic set is much less likely to constitute a foldable set and hence requires additional components for life to emerge. Whether or not this abiotic set has the capacity to form foldable proteins is the subject of the forthcoming chapter.

Although Chapter 2 shows that cysteine's thiol group may be a vital requirement for enabling abiotic protein function, a broader investigation of a range of proteins with compositions more closely resembling those proposed for early abiotic systems is required. Specifically, I sought to determine whether the predictions I made of these abiotic proteins, that they possess a molten globular fold, are sensitive to acidic pH, insensitive to salinity changes, and capable of metal binding if cysteine were available, are indeed relevant to abiotic proteins.

Abiotic proteins and the challenge of limited sidechain chemistry

Unlike modern proteins which exploit a range of side chains to function, abiotic amino acids lack the aromatic and charged residues that contribute to hydrophobic folding, electrostatic interactions, and catalysis (Zaia *et al.*, 2008). This poses the question: can proteins composed predominantly of abiotic amino acids adopt stable and functional structures, or are they restricted to either non-foldable disordered ensembles or require extreme environmental niches to fold? (Tretyachenko *et al.*, 2022)(Farías-Rico and Mourra-Díaz, 2022).

On an *a priori* basis, some initial predictions of their structural propensity can be made. Modern proteins possess two major secondary structural types - β-sheets - which are stabilised by backbone hydrogen bonding between strands of amino acids in extended conformations, or helices which are stabilised by stacked hydrogen bonds between amino acids within a stretch of sequence where helical formation is strongly dependent on the type of amino acids present to ensure correct spatial distancing between hydrogen bond forming sites for helical formation (**Fig 3.01**) (Sun *et al.*, 2004).

Given that β-sheets rely on structural stabilisation available to all amino acids, i.e. backbone hydrogen bonding, one can immediately infer that this more probable in abiotic proteins given that α-helices require higher sequence conservation to ensure the residues can support the tight turns and hydrogen bond stacking. Proteins that contain secondary structure, but a limited

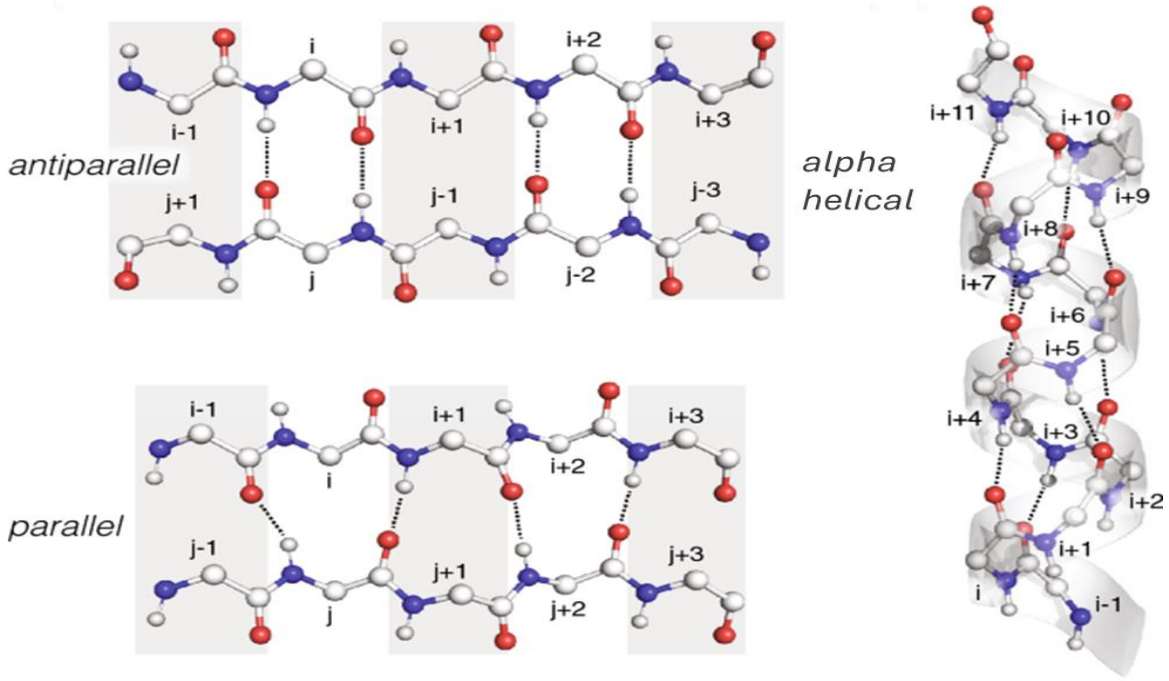


Figure 3.01 – Common secondary structural types observed in modern proteins.
Adapted from (Friedrich *et al.*, 2020)

hydrophobic cores are called molten globules. Typically, the secondary structure can be detected by Circular Dichroism (CD), but the protein does not undergo a cooperative unfolding event upon heating (Christensen and Pain, 1991). As discussed previously metal binding may provide an opportunity to add structure to these proteins, making them less 'molten,' and either indirectly through structure, or directly through chemistry, facilitate catalysis (**Fig 1.05**).

Previous attempts to characterise abiotic proteins.

Two major studies have attempted to characterise abiotic protein stability. However, their methodologies limit the generalisability of their conclusions. McDonald *et al*. conducted a bioinformatics-based search to identify modern proteins that lack basic charges. They searched 16 modern extremophile genomes to identify proteins that lacked lysine and arginine. While these proteins demonstrate functionality in modern organisms, the authors themselves noted that their detections were neither entirely devoid of basic residues, retaining histidine, nor were they very representative of abiotic proteins, containing 18/20 of the modern amino acid set (McDonald and Storrie-Lombardi, 2010). McDonald also noted that negative-charge-dominated proteins, while functional in modern microbes, are unlikely to interact effectively with nucleic acids, a limitation that would constrain their roles in early biochemistry. However more recent work suggests that this could be solved by metal ion interactions (Giacobelli *et al.*, 2022).

Similarly, Longo *et al*. engineered β-trefoil proteins to an 80% abiotic amino acid content (Longo *et al.*, 2013). Although these constructs demonstrated co-operative folding into a tertiary structure with a rudimental hydrophobic core,

they required high salt concentrations for stability, comparable to a drying puddle (2M NaCl). As such, they argued that halophilic environments, such as evaporating seawater pools, may have been a crucial environment for early proteogenesis. However, the retention of ≈20% non-abiotic residues, including asparagine, arginine, and phenylalanine, allowed these constructs to preserve hydrophobic and charge-based interactions that would have not been available to fully abiotic proteins. As such, the relevance of this work to the question of general abiotic protein folding is moot.

An inherent limitation in both studies, and indeed a limitation of our own study, is that any abiotic-like proteins observed today are the beneficiary of evolutionary optimisation over possibly billions of years. Concisely, are present day 'abiotic-like' proteins reasonable proxies for the initial protein constructs? Whilst extant 'abiotic-like' proteins may be imperfect analogues of primordial proteins, they are still useful for determining structural trends, with the caveat that their identification should be based on as broad an expanse of the modern proteome as possible.

A broader search

A constructive mechanism to build upon these works would be to identify modern proteins that are significantly more 'abiotic-like' than the previous searches and investigate more of them. By analysing naturally occurring proteins with a higher percentage content of abiotic amino acid, some of the limitations of the previous approaches can be addressed and a wider landscape of genomes explored than was possible in the earlier work of McDonald *et al* (McDonald and Storrie-Lombardi, 2010). Once a significant number of abiotic-like proteins are

identified, I can imitate the work of Longo *et al* to purify these proteins and verify their folding behaviour across a range of environmental conditions. Through this strategy I integrate the strengths of both previous studies across a much wider landscape at both the informatic and in vitro level.

Approach

To achieve this, a bioinformatic analysis was conducted of the UniProtKB/TrEMBL database in search of naturally occurring proteins whose sequences comprise at least 95% abiotic amino acids, thereby excluding nearly all of the basic and aromatic residues. Searches at 100% are not possible as all modern proteins begin with the biotic amino acid, methionine. This search was conducted both with and without cysteine included as an abiotic amino acid (Cys- and Cys+ runs) to identify what impact, if any, cysteines inclusion would have on the commonly identified structures. After filtering out heavily disordered or tandem repeat sequences, a set of “abiotic-like” candidates remained. Using AlphaFold2, I used their previously predicted structures to analyse their secondary and tertiary structures to identify common structural characteristics that could be generalised to other abiotic proteins (Jumper *et al.*, 2021).

Five candidate proteins from this dataset representing the major secondary types (β-solenoids, and helical mixtures) and the functionally annotated proteins observed were selected for closer study. Of these functionally annotated proteins, there was only one with a relevant previously solved structure. Lysine biosynthesis protein has a previously reported crystal structure on the PDB (3WWL), which has a 56% and 71% sequence similarity to the LysW and LysW-C proteins used here. These proteins were expressed, purified, and

analysed by CD, UV-visible spectroscopy, and solution state NMR to examine their folding, thermal stability, and responsiveness to changes in pH, temperature, and salinity. These proteins were assessed across a range of pH (4-10), salinity (0-1000mM), and temperature (20-90°C) to evaluate their structural stability and folding behaviour under the abiotic solvent conditions discussed in Chapter 1. By conducting these analyses both with and without cysteine present, I could assess the impact of cysteine as an abiotic amino acid, as proposed by Foden *et al*, and what its thiol chemistry could deliver in terms of stabilising early protein folds (Foden *et al.*, 2020). While the functionality of the tested proteins was not interrogated, given the need for even transient folding to occur for protein function, the structural stability of these proteins provides a useful proxy to establish their functional potential.

Methods

Abiotic-like protein selection

*The Python scripts used here are presented within **Appendix C.2**.*

The UniProtKB/TrEMBL database (245,871,724 entries at time of analysis) was processed to identify proteins with a high abiotic amino acid content (**Fig 3.02**). To facilitate this large-scale analysis, the database was split into multiple ≈5 GB FASTA chunks for computational ease (**splitfasta.py**).

For each protein, the abiotic index, defined as the percentage of residues per protein containing one Gly, Ala, Asp, Glu, Val, Pro, Ile, Leu, Ser, or Thr, was calculated (**abiotic_index_calculator_multithreadversion.py**). For the Cys+ analysis, cysteine was included in the set. Only sequences with a minimum length of 30 residues and an abiotic index ≥95% were retained. Accession numbers from retained sequences were extracted (**extractacsimple.py**) and their AlphaFold-predicted structures were retrieved from UniProt alongside their associated meta data (**experimentstructurefetcher.py, basic_analysis_tool4.py**). Secondary structure analysis of AlphaFold predictions was performed using DSSP (**dssppipe.py**). Proteins with excessive (>50%) unstructured regions were excluded.

An abiotic score was calculated by summing the absolute differences between each amino acid's observed fraction within each protein and the meteoric amino acid abundance from Higgs and Pudritz and then normalising by the number of amino acids considered (Higgs and Pudritz, 2009). Higher scores therefore indicate increasing deviation from the meteoric ratio, while lower scores

indicate closer similarity to an abiotic amino acid composition (**abioticscore.py**). Proteins with scores above 0.12 were excluded based on the observation that such sequences appeared to have high enrichment of specific abiotic amino acids. Per-residue pLDDT scores were extracted and proteins averaging below 70% pLDDT were discarded (**extract_plddt.py**). Tandem repeats (2-12 residues) were identified by scanning each protein sequence up to a maximum repeat length of 12aa's per repeat block and identifying and summing further consecutive identical segment. The final merged repeat length was then divided by the total sequence length and multiplied by 100, yielding the tandem repeat percentage (**tandemrepeatdetector.py**). Proteins with >40% tandem repeat percentage were excluded.

After applying all cut-off criteria, 38 (Cys-) and 39 (Cys+) non-homologous proteins remained. Manual inspection identified several structural phenotypes (helical proteins, β-solenoids, β-barrels, α/β ferredoxin-like, and α/β mucin17-like structures).

From these, five representative proteins were selected. a β-solenoid (SALSA), a helical protein (PAST), a β-barrel Lysine Biosynthesis protein W (LysW), a Cys-containing LysW homologue (LysW-C), and a ferredoxin-like protein (FerreC). During early attempts to purify PAST, I observed rapid oligomerisation and precipitation following HisTrap purification. As such no further analyses were performed on PAST. Predictions for each selected protein were regenerated using AlphaFold3, and models were visualised in PyMOL (Abramson *et al.*, 2024).

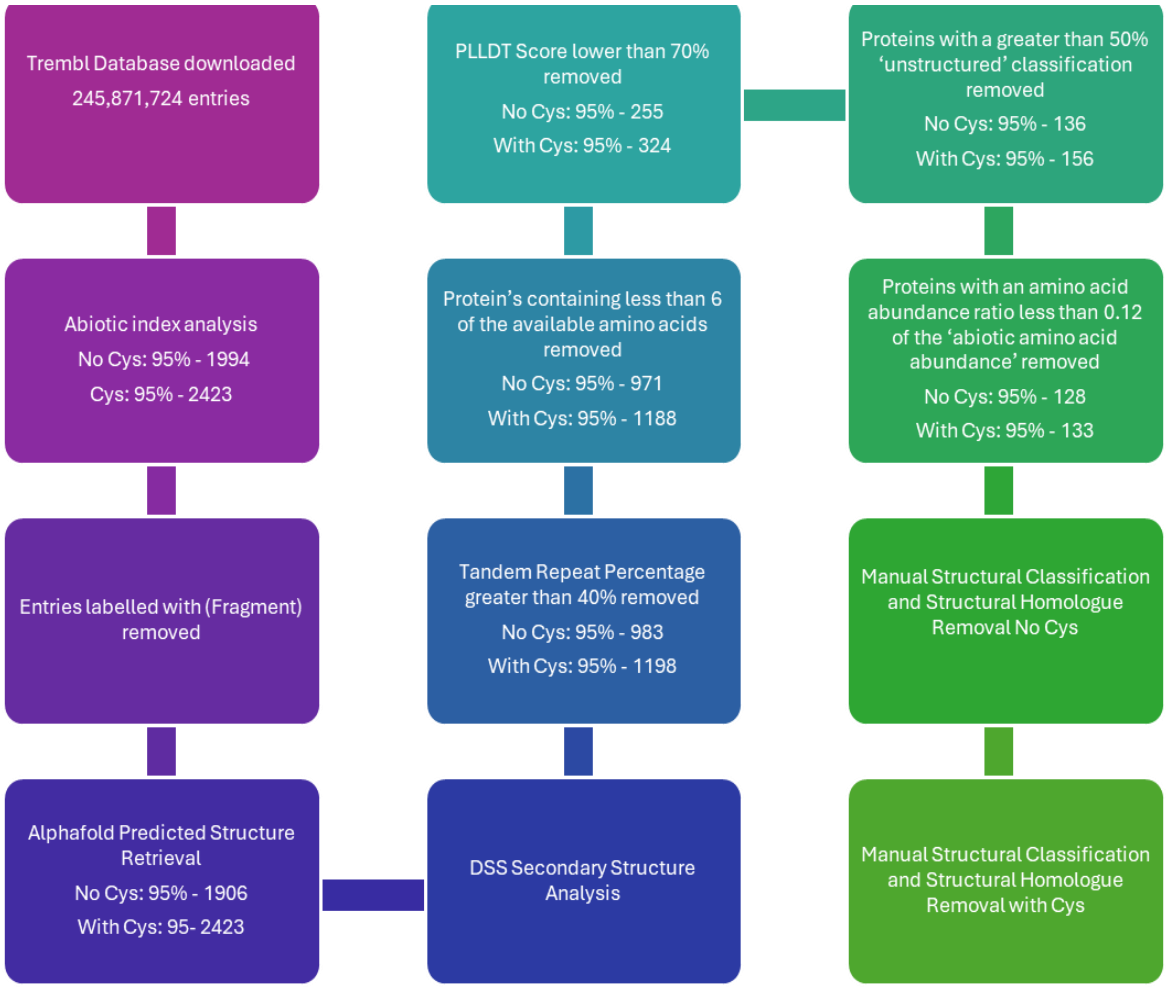


Figure 3.02 – Bioinformatic approach
Summary of bioinformatic approach as described (page 137)

Protein expression and purification

Protein expression

Codon-optimised genes for SALSA, PAST, LysW, LysW-C, and FerreC were synthesised and cloned into pET-11a vectors by GenScript. Plasmids (4µg) were resuspended in 20µl Milli-Q water and stored at -20°C until use.

E. coli BL21(DE3) cells were transformed using the heat shock method: 0.5µl of 200 ng/µl plasmid DNA was incubated at 45°C for 45s, then on ice for 5 min. Transformed cells were plated on LB agar (100 µg/ml ampicillin) and incubated overnight at 37°C. Single colonies were grown in LB broth (100 µg/ml ampicillin) at 37°C until OD_{600} 1.0 was reached, then 20ml of this starter culture inoculated 1L LB high-salt (LB HS) medium (100 µg/ml ampicillin). Cultures were grown at 37°C to OD_{600} 0.6-0.8, induced with 1mM IPTG, and incubated at 18°C for 16-20 h.

Harvested cells were centrifuged (7000x*g*, 15 min, 4°C) and pellets resuspended in 40ml of lysis buffer (20mM Tris-HCl, pH 7.4). Pellets were sonicated on ice (45 amplitude, 5s on/25 s off cycles (QSonica Q700, 700W) and the lysate was centrifuged at 30,000x*g*, 45 min, 4°C to remove insoluble material.

The resulting supernatant was loaded in two 20ml aliquots to a 25ml HisTrap (gravity flow) NiNTA column (cOmplete Ni-NTA His-tag purification resin) pre-equilibrated via 4x wash steps of 30ml 20mM Tris-HCl pH 7.4 flushed through the column. After the supernatant flowed through, the column was washed with 60ml of the same buffer and the flow through was collected. Proteins were eluted with 15ml of 20mM Tris-HCl, pH 7.4, 500mM imidazole, and elution fractions were

concentrated using a Vivaspin 20 (5 kDa cut-off for FerreC, LysW, LysW-C; 30 kDa cut-off for SALSA) and buffer-exchanged into 0.1M phosphate buffer, pH 6.0, by 2x dilution and concentration steps down to 2ml. Concentration was monitored using a Benchtop Nanodrop (Thermo NanoDrop 1000 UV-Vis Spectrophotometer)

Size-exclusion chromatography (SEC) was performed using a Superdex 75 Increase 10/300 GL column (Cytiva) equilibrated with 0.1M phosphate buffer, pH 6.0 on an ÄKTA pure FPLC (Cytiva). Typically, a 1ml aliquot of concentrated protein (≈10 mg/ml) was loaded and eluted at 0.75ml/min over 20 min.

SDS-PAGE analysis

Purity of expressed proteins was initially assessed by SDS-PAGE to confirm approximate molecular weight and detect any degradation and contamination products prior to further analyses. Protein samples (30µl) were mixed with 10µl 4x NuPAGE LDS Sample Buffer (Thermo Fisher Scientific), heated at 70°C for 10 min, and aliquots (10µl) loaded onto NuPAGE 4-12% Bis-Tris gels (Thermo Fisher Scientific) with MES SDS running buffer. Electrophoresis was performed at 200V 110mA for 35 min. Gels were stained with Coomassie Brilliant Blue R-250 and destained. Bands were visualised using a GelDoc system.

UV-Visible spectroscopy

UV-Vis was used to measure metal binding of Fe-LysW-C protein. Samples (≈1mg/ml) in 0.1M phosphate buffer, pH 6.0 were placed in a Hellma quartz cuvette (10mm path). Spectra were recorded from 200-800nm with a Cary

50 Bio spectrophotometer (Agilent) at 1nm intervals and 4800 nm/min scan rate. Baseline correction was performed against the sample buffer only.

Intact mass spectrometry

Intact MS provided confirmation of purification on all expressed proteins. Intact protein masses were measured by ESI-QTOF LC-MS (Agilent). 1µl of 0.05 mg/ml protein was injected into a short linear 1–3-minute ACN/FA gradient (Buffer A: 0.1% FA in H2O, Buffer B: 0.1% FA in ACN) up to 100% ACN. Data was deconvoluted to determine the molecular weight of the protein and compared against predicted masses using the Agilent MassHunter Qualitative Analysis software.

Circular Dichroism (CD) spectroscopy and stability assays

CD was used to initially characterise the secondary structure of each protein and subsequently monitor how the secondary structure content changes with temperature, pH, and salinity. CD spectra were recorded using a Jasco J-815 spectropolarimeter. Samples (1 mg/ml) in 0.1M phosphate buffer (pH as indicated) were placed in a 1mm path length quartz cuvette. Spectra were acquired from 260nm to 190nm at 20°C, with displayed results representing average signal from nine accumulations per sample. Data with voltages above 700V were discarded.

Thermal melts were performed by heating 150µl of 0.01 mg/ml LysW, Fe-LysW-C, SALSA, and FerreC in 10mM sodium phosphate, pH 6.0 from 20-90°C at 2°C increments. Data were collected with two accumulations at each temperature point to assess unfolding.

For pH screens, 1 mg/ml LysW, Fe-LysW-C, FerreC, and 0.1 mg/ml SALSA were adjusted to pH 4, 5, 6, 7, 8, 9, and 10. CD spectra were acquired averaging eight accumulations per sample.

For salinity screens, protein solutions (1 mg/ml for LysW, Fe-LysW-C, FerreC, 0.01 mg/ml for SALSA) were prepared at 0, 250, 500, and 1000mM NaF. CD spectra were collected and averaged across eight accumulations for each condition.

Protein solution NMR spectroscopy

To confirm monomeric purity, Fe-LysW-C metal binding and that Ferredoxin was purified in the apo state, solution state NMR was conducted on isotope labelled proteins. $^{13}C/^{15}N$ labelled samples of Holo(Fe)/Apo LysW-C and Apo LysW, and FerreC (0.5-0.85 mg/ml) in 90% H_2O/10% D_2O, pH 6.2, 0.1M phosphate buffer were analysed. Samples were placed in 3mm NMR tubes and analysed on a 22.3 Tesla spectrometer (Oxford Instruments magnet) equipped with a high sensitivity 5mm TCI cryoprobe and Avance III HD consoles (Bruker) with a SampleJet autosampler at 20°C. Two-dimensional ^{1}H-^{15}N HSQC spectra were acquired ([32 **Sc**], [2048, 256 **Td**], [15.96, 32.07 **Sw**]). Data processing and analysis were performed in TopSpin (Bruker).

Results

Abiotic protein secondary structure propensities

As an initial step, amino acid propensities for secondary structure formation reported by Sen *et al* were re-examined, using only the abiotic amino acids together with cysteine (Sen *et al.*, 2024). The unscaled data suggested comparable propensities for α-helices and β-sheets, implying equal structural likelihoods. However, when scaled according to meteoritic amino acid abundances (**Fig 3.03a**), I observed a decrease in overall β-sheet propensity and a relative increase in turns and loops, alongside a modest reduction in α-helical propensity. Including cysteine in the analysis (**Fig 3.03b**) slightly increased the predicted α-helical fraction but did not fundamentally alter the scaled trends. These predictions were then tested further within the bioinformatics search phase.

Abiotic-like protein identification and structural classification

Naturally occurring proteins whose amino acid compositions contained 95% or more abiotic amino acids were identified. Once identified, proteins were filtered through multiple criteria to identify the most abiotic proteins. These filters aimed to remove proteins that were poorly representative of well-folded abiotic proteins due to being short fragments (30 amino acid minimum filter), having structural model predictions with low confidence (pLDDT filter), tandem repeat proteins (therefore requiring high sequence conservation), or only containing very few of the abiotic amino acids, i.e. low complexity filters to remove proteins with a large number of a few abiotic amino acids (relative abiotic score, distinct amino acid filter).

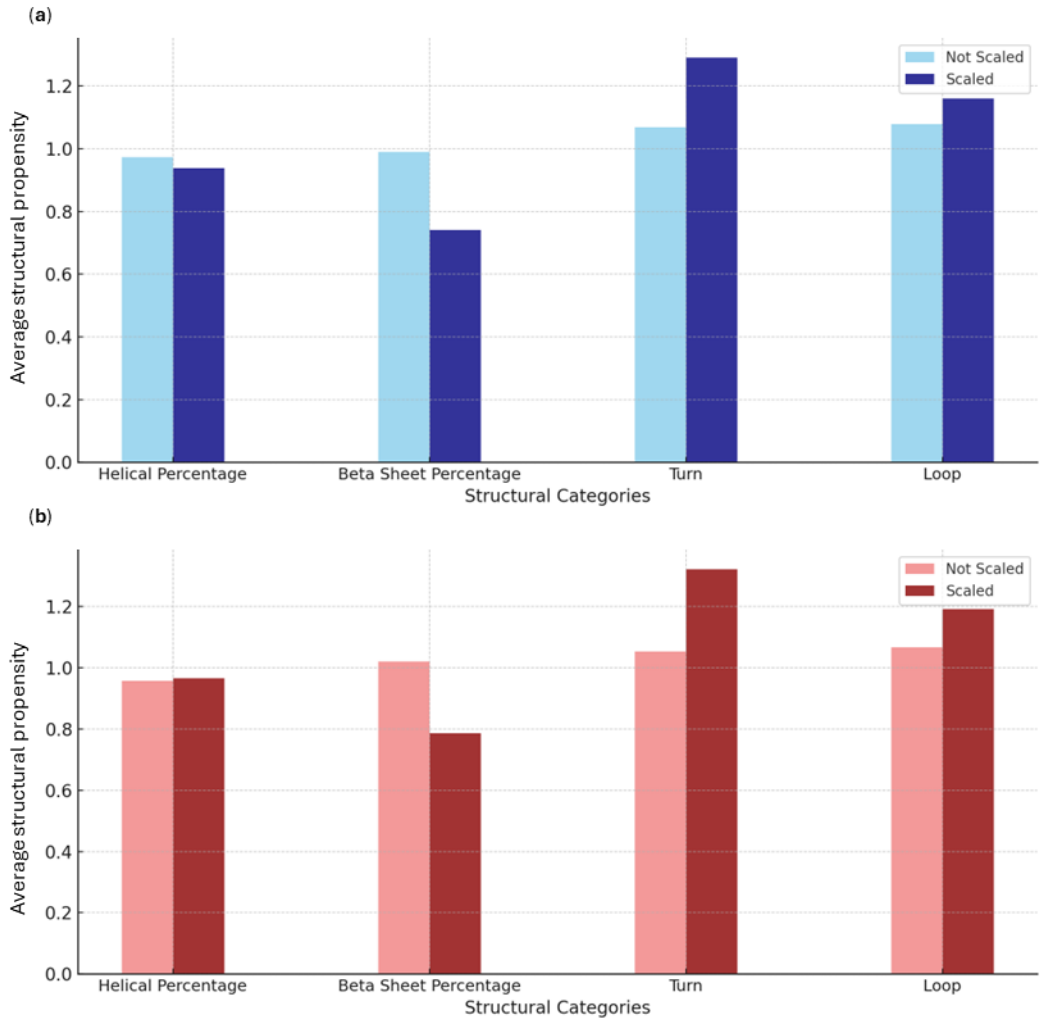


Figure 3.03 – Abiotic proteins will possess a helical propensity based on observed abiotic amino acid abundance.

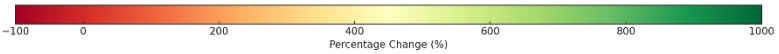
*(**a**) Average structural propensities of proteins composed solely of the 10 abiotic amino acids, categorized into α-helices, β-sheets, turns, and loops based on the amino acid structural propensities reported by (Sen et al., 2024). Unscaled data (left-hand bars) suggests comparable propensities for α-helices and β-sheets, with a smaller contribution from turns and loops. After scaling the propensities according to meteoritic amino acid abundances reported by (Higgs and Pudritz, 2009) (right-hand bars), β-sheet propensity decreases, while turn and loop propensities increase. **(b)** Average structural propensities of proteins composed solely of the 10 abiotic amino acids and cysteine. Unscaled data is consistent to those observed in (**a**) but when scaled for meteoritic abundance, β-sheet depletion persists, while α-helical content shows a slight increase. This addition of cysteine suggests a modest stabilization of α-helices but does not fundamentally alter the observed trends.*

Following a search of over 245 million protein sequences, for the no-cysteine (Cys-) dataset, 1,994 initial proteins were identified which, following filtering, was reduced to 38 non-homologous representatives comprising three major structural phenotypes, α-helical proteins, β-solenoids, and a β-barrel Lysine Biosynthesis protein W (LysW). For the cysteine-including (Cys+) dataset, 2,423 initial proteins were narrowed down to 39 non-homologous representatives encompassing five phenotypes: helical, β-solenoids, a LysW-like β-barrel, a helical/β-barrel mixture resembling a ferredoxin, and a helical/β-barrel mixture similar to mucin17. All identified proteins are displayed in **Appendix B - Table 01a-b.**

Comparisons of amino acid abundances in these abiotic-like proteins (**Fig 3.04a**) showed that alanine (-6 to -10%), glycine (-79 to -78%), aspartate (-40 to -38%), and glutamate (-8 to -2.61%) were depleted relative to the predicted abiotic abundances, while valine (196 to 240%), leucine (847 to 818%), isoleucine (96 to 88%), serine (329 to 259%, and threonine (971 to 909%) were enriched. This over-enrichment is most likely a consequence of sequence adaptations to ensure that these proteins are sufficiently hydrophobic for protein folding to occur in its native environment. While previous authors have suggested that proteins most similar to abiotic proteins in terms of amino acid composition are produced by halophiles, which is supported by their low proteome pK_a, environmental classifications of the host organisms for the identified proteins indicated that most of these proteins originated from mesophilic organisms, suggesting that abiotic-like amino acid compositions are not restricted to extremophilic environments (**Fig 3.04b**) (Longo *et al.*, 2013).

(a)

	A	C	G	V	P	L	I	D	E	S	T
Abiotic Amino Acid Abundance	0.24	0.01	0.41	0.04	0.03	0.01	0.02	0.12	0.1	0.02	0.01
Observed Abundance in Abiotic like Proteins (Excluding Cys)	0.225	nan	0.083	0.118	0.03	0.089	0.039	0.072	0.108	0.046	0.107
Percentage Difference (Excluding Cys)	-6.106	nan	-79.705	196.175	1.503	847.082	95.596	-40.21	8.464	329.948	971.306
Observed Abundance in Abiotic like Proteins (Including Cys)	0.215	0.013	0.09	0.139	0.034	0.083	0.038	0.075	0.097	0.072	0.121
Percentage Difference (Including Cys)	-10.261	28.242	-77.947	240.062	12.465	818.331	88.219	-37.795	-2.61	259.36	909.202



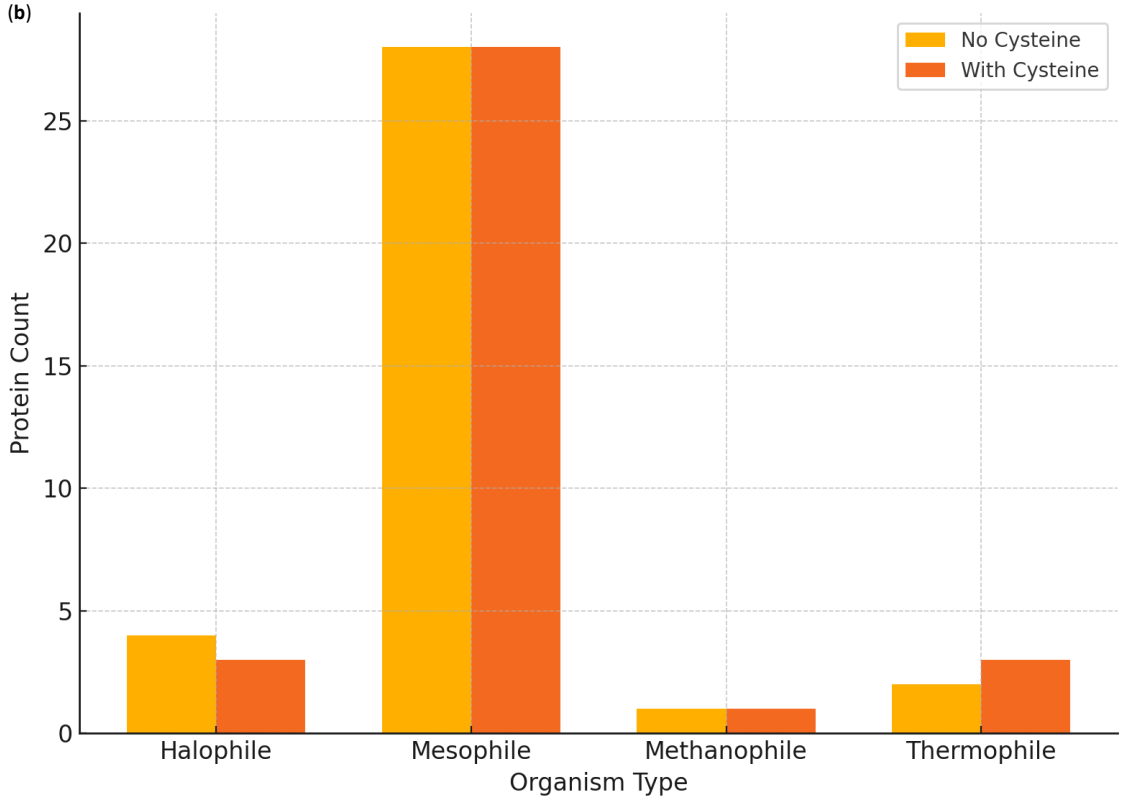


Figure 3.04 – Abiotic-like proteins remain over-enriched in hydrophobic amino acids.

*(**a**) Comparison of predicted abiotic amino acid abundances with observed amino acid abundances in proteins identified from the Cys- and Cys+ datasets. Key differences include depletion of alanine, glycine and aspartate, and over-enrichment of valine, leucine, isoleucine, serine, and threonine. (**b**) The host organism from each identified protein was retrieved and reported. The vast majority of identified proteins originated from mesophilic organisms.*

DSSP analyses of secondary structures (**Fig3.05a**) indicated that α-helices, β-elements, and unstructured regions dominated both Cys- and Cys+ sets. Introducing cysteine (Cys+) slightly reduced β-content and increased unstructured fractions. The relatively high identification of β-content compared to my sequence level predictions (**Fig 3.03**) was attributed to the high amounts of β-solenoids in the data set (**Fig 3.05b**). While these structures appear to be highly abiotic, their stability and defined structures come from highly conserved hydrophobic interactions between valine and leucine in the β-solenoid pore. The high conservation of these complexes, which extends along the entire pore of the solenoids, makes it less likely that these are representative of true abiotic proteins, even if they can likely form with a reduced set of abiotic amino acids. Among these identified proteins, five were selected for further testing.

Three of these proteins represented functional proteins (LysW, LysW-C and Ferredoxin), while PAST and SALSA are instead more representative of structural proteins heavily depleted in biotic amino acids but also found to observe large regions of repeats. While short tandem repeats were removed from the dataset, both these proteins had longer repeat sections which eluded this filter. While this inherently decreases their abiotic relevance given the dependence on conserved repeats for structure, they still provide a useful tool to understand potential abiotic protein functions and thus were carried forward.

1. SALSA (**Fig 3.06a**), predicted as a β-solenoid.

   MEGGGSLTVGEEGGGSLTVGLLLGGTTTGDPEELALTTTGDPEELALT
   TTGDPEELALTTTGDPEELALTTTGDPEELALTTTGDPEELALTTTGDS
   EELALTTTGDPEELALTTTGDPEELALTTTGDSEELALTTTGDSEELALT
   TTGDPEELALTTTGDPEELALTTTGDSEELALTTTGDPEELALTTTGDP
   EELALTTTGDPEELALTTTGDPEELALTTTGDPEELALTATGDPEELALT

ATGDPEELALTATGDSEELALTTTGDSEELTLTATGDSEELALTTTGDP
EELALTATGDPEELALTATGDPEELALTATGDSEELALTATGDSEELALT
TTGDSEELALTTTGDSEELALTTTGLPEELVLIVEIPLLL

2. LysW (**Fig 3.06b**), a β-barrel Lysine Biosynthesis protein

   MADTITAEDPLTGEAIELPADVEVGEIVDSPATGAELEVVSLDPVTLEEA
   PELEEDLGE

3. PAST (**Fig 3.06c**) predicted as predominantly helical.

   MLAAVAEVDALLAEVLAAVAEVEALLAEVLAAVAELEALLAEVLAAVAEL
   DASLAEELAAVAELDALLAEVLAAVAELEALLAEVLAAVAELADLSAEVL
   AAVAELEALLAEVLASVAELADLSAEELAAVAELDALLAEVLAAVAELEA
   LPAEVLASVAELAASLAEEPAAVAELDASLAEVLAAVAELDALVAEVLAA
   VAELDALLAEVLAAVAELDALLAEVLAAVAELADLSAEVLAAVAELEALL
   AEELAAVAEVEALLAEVLASVAELADLSAEELAAVAELDALLAEELATVA
   ELVASLAEVLAAVAEEVAAVAELSA

4. LysW-C (**Fig 3.07a**), a cysteine-containing variant of LysW

   MTVECPECAATVELPADVEVGEILVCPDCGVELEVVGLEPVAVDLAPE
   VEEDWGE

5. Ferredoxin-like protein (**Fig 3.07b**), with a helical/β-barrel mixed fold

   MTAKVDADTCTGCESCVEACPLEAIELVDDIAVVDEDTCGDCGECVDV
   CPVEAITVDE

Although PAST was identified as a helical candidate, it rapidly oligomerised and formed a visible white precipitate following HisTrap purification. As a result, subsequent experimental characterisation focussed on SALSA, LysW, LysW-C, and the ferredoxin-like protein to determine how abiotic-like compositions and the presence of cysteine influence stability and folding under abiotic-like conditions.

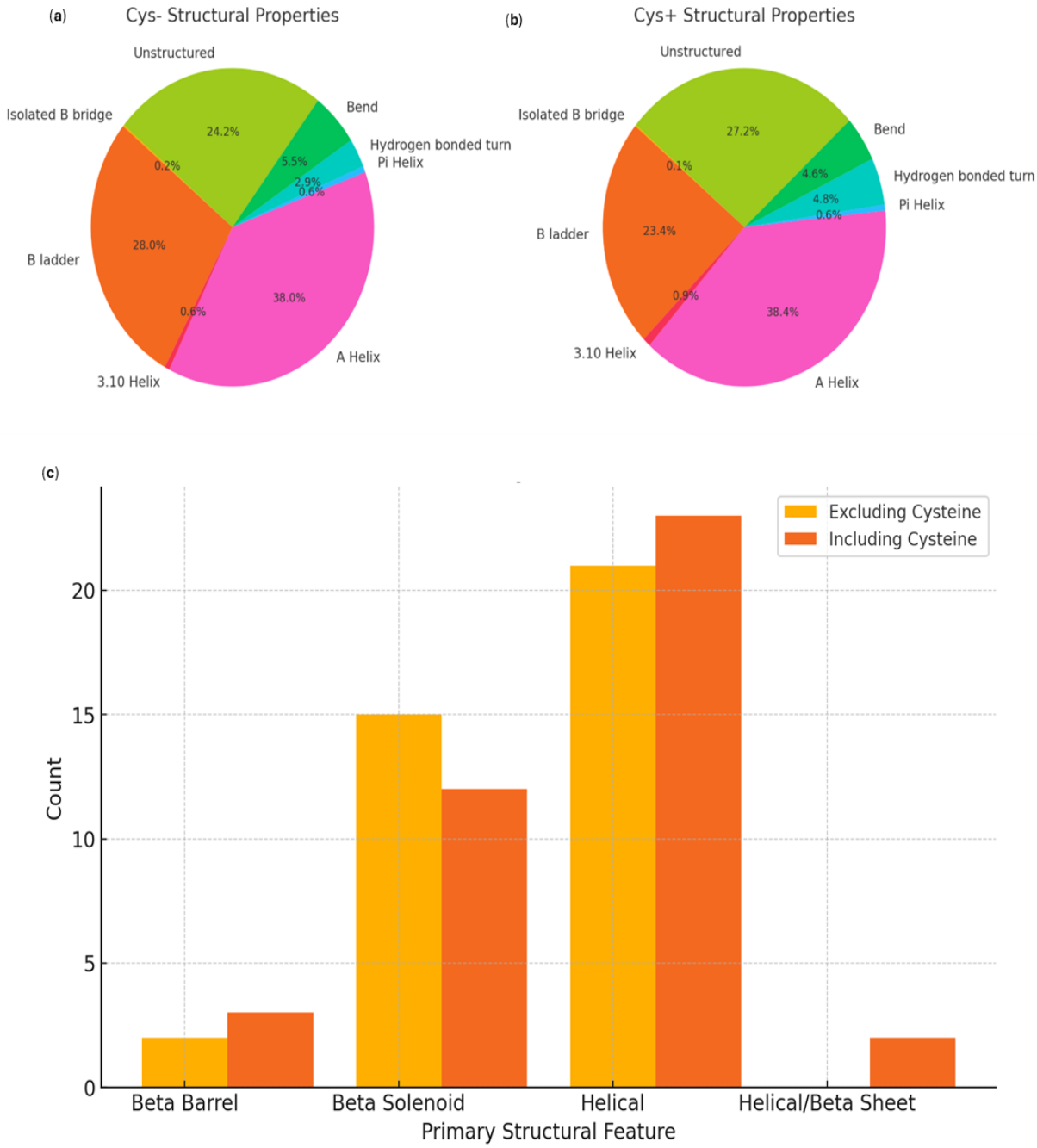


Figure 3.05 – Incorporation of cysteine within the search parameters increases the number of functional proteins identified.

*(**a-b**) Inclusion of cys+ decreases the average beta sheet content in observed abiotic-like proteins. DSSP analysis of proteins identified from the Cys- and Cys+ dataset was conducted on Alphafold2 predicted structures of the proteins. A reduction in the average beta sheet content was observed with a corresponding increase in average unstructured content in the Cys+ dataset. (**c**) a tally of the observed primary structural feature of each protein reveals this is likely through a reduced detection of beta solenoids.*

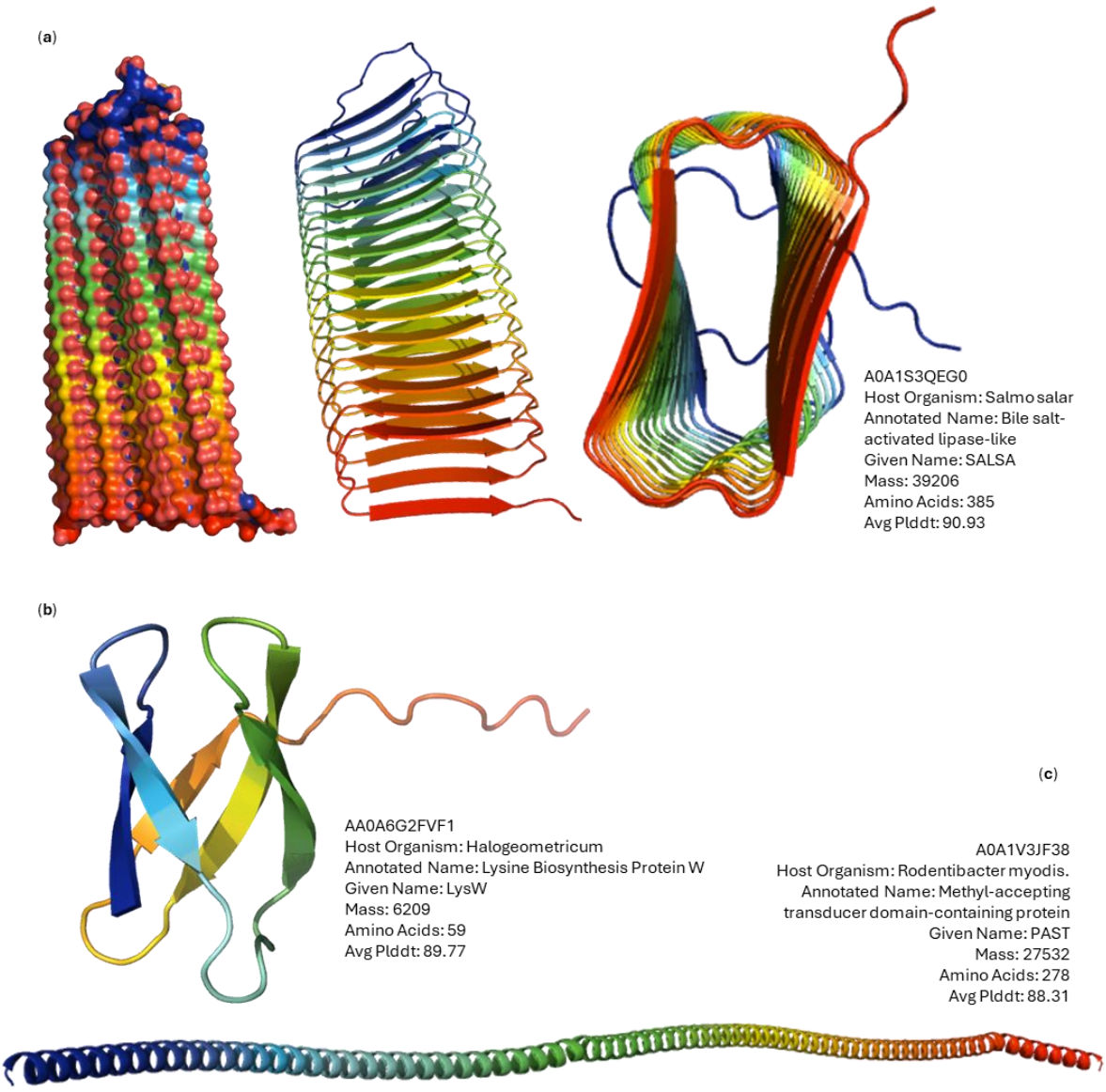


Figure 3.06 – Proteins identified within the Cys- search are dominated by extended helices and beta solenoids.

*Alphafold3 predictions were generated of the amino acid sequences from accession numbers A0A1S3QEG0 (**a**), AA0A6G2FVF1 (**b**) and A0A1V3JF38 (**c**). Structures were download into and images exported from Pymol (educational licence)*

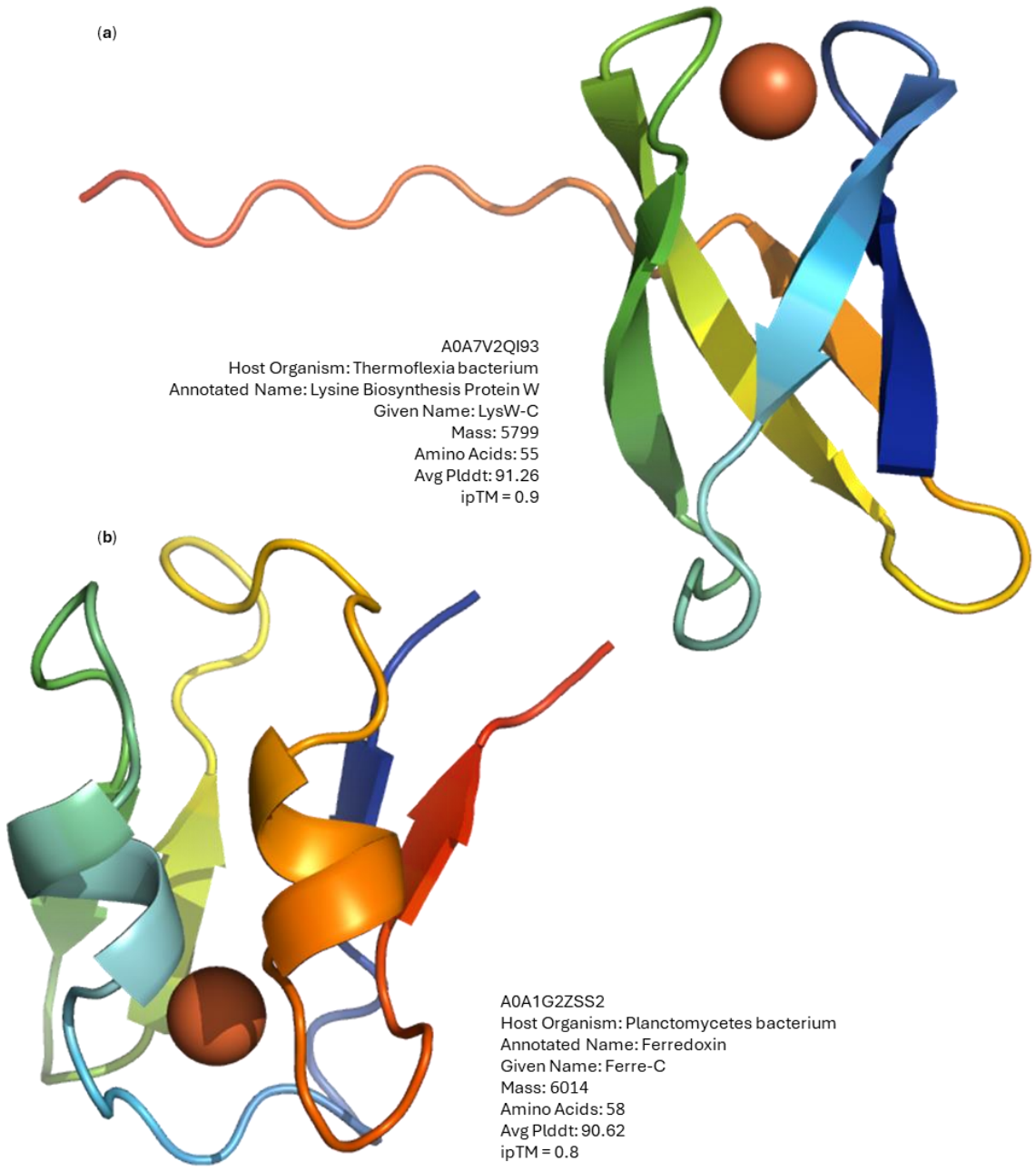


Figure 3.07 – The Cys+ search reveals two potentially iron binding abiotic-like proteins.

*Alphafold3 predictions were generated of the amino acid sequences from accession numbers A0A7V2QI93 (**a**), A0A1G2ZSS2 (**b**) alongside an Fe^{3+} ligand. Structures were download into and images exported from Pymol (educational licence)*

Abiotic-like protein analysis

A β-solenoid?

While the high sequence conservation in the abiotic-like β-solenoids made it less representative of abiotic like protein behaviour, it provided a useful initial test model for this study given its enrichment in abiotic amino acids. SALSA, which was predicted to form a stable β-solenoid fold, was expressed and purified (**Fig 3.08a-b**). Following SEC, SDS-PAGE confirmed a single band near the predicted ≈39 kDa mass and a second higher mass object at approximately 70kda (**Fig 3.08c**). The higher mass product was assigned to an oligomer forming in the gel, which at this point had been identified for nearly all proteins analysed by this technique. The presence of monomeric SALSA was confirmed by intact mass spectroscopy (**Fig 3.08d**). Unlike the predicted secondary structure from Alphafold2 and 3 (**Fig 3.06a**), CD spectra indicated a strongly α-helical signature rather than extensive β-content (**Fig 3.09a-b**). BestSel analysis of the resulting spectra confirmed the dominance of α-helical structure in this structure (**Fig 3.09c**) (Micsonai *et al.*, 2022)**.** While this difference for the observed spectra versus the Alphafold predicted structure was surprising, it was not entirely unexpected. During manual structural classification, I observed several proteins that had identical functional UniProt annotations, but which possessed an extended helix in one structure and a β-solenoid in the other, suggesting that the sequences that result in β-solenoids and the extended helices have enough sequence overlap to result in this discrepancy.

CD melts revealed gradual, cooperative unfolding of SALSA between 50-60°C which suggests that this protein adopts a tertiary structure more consistent

with a conventional protein (**Fig 3.10**). pH titrations for SALSA demonstrated that this structure was stable over the pH range 4-10 (**Fig 3.11**) but with possible sensitivity to salinity when screened against NaF, which was used to avoid signal interference from the Cl^- ion in CD (**Fig 3.12**), although the signal-to-noise ratio of these spectra prevented any confident conclusions on this.

Together this data show that SALSA, despite being predicted to adopt a β-solenoid conformation, instead exhibits a predominantly α-helical fold under our expression conditions and displays moderate thermal stability and pH resilience.

Cysteine increases structural propensity and stability of the LysW β-Barrel.

Lysine biosynthesis protein W was previously identified as an abiotic like protein by McDonald *et al*, although it has not been well-characterised, with one crystal structure in the PDB (3WWL). As such, LysW and Fe-LysW-C, its cysteine-containing counterpart, were investigated here. Purification and analysis by SEC (**Fig 3.13a-b**) and SDS-PAGE (**Fig 3.13c-d**) and intact mass spectrometry (**Fig 3.13e-f**) confirmed that both proteins were isolated as monomeric species at their predicted masses. Notably Fe-LysW-C had a red colour when purified while LysW was colourless.

UV spectra (**Fig 3.14a**) indicated that Fe-LysW-C exhibited an absorbance peak at 350nm which was absent in LysW, consistent with an iron-sulfur interaction. Solution state NMR spectra of Fe-LysW-C showed only a small proportion of the expected signals, consistent with bound iron, which is paramagnetic and suppresses the signals of nearby nuclear spins (**Fig 3.14b**). In

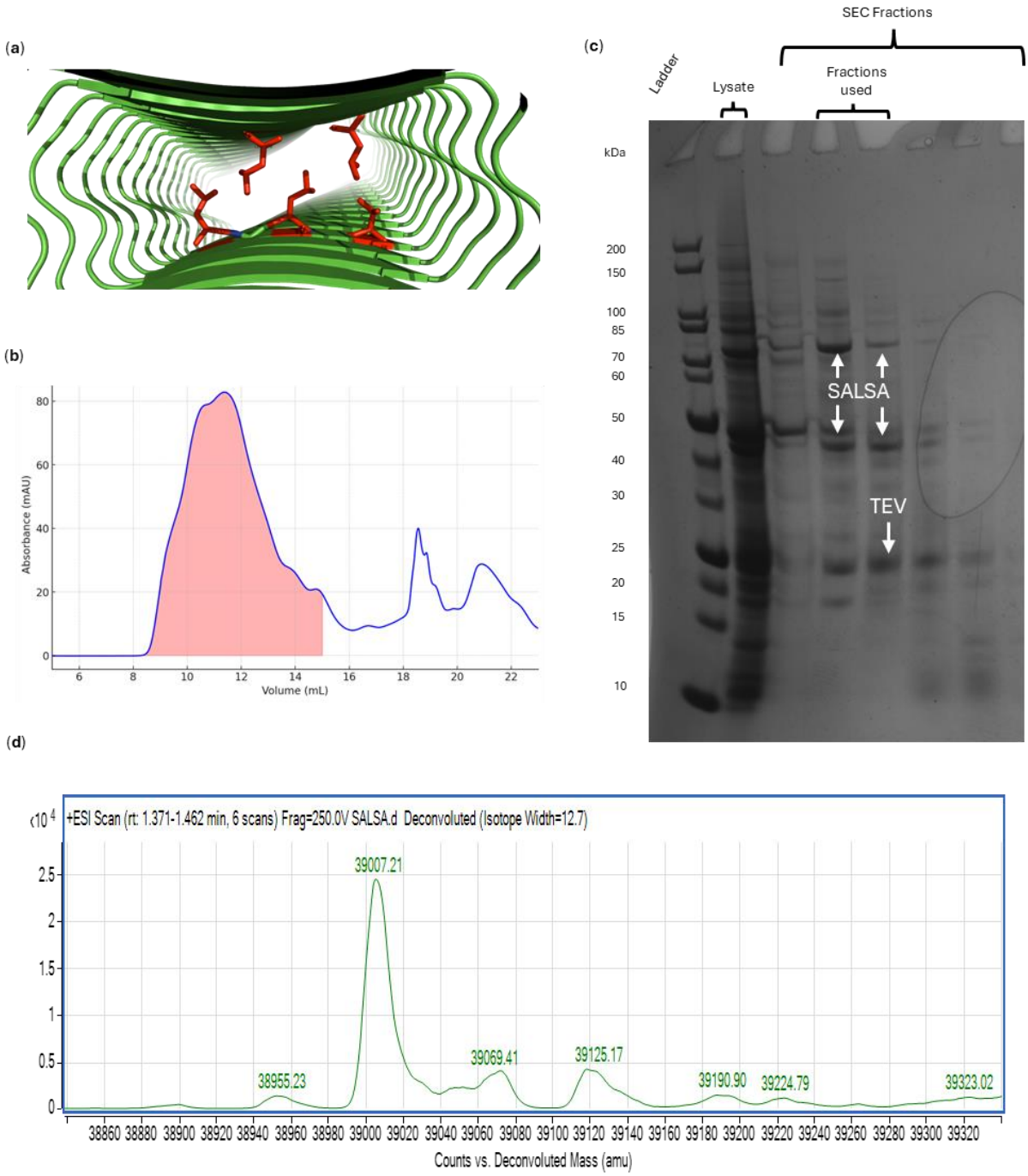


Figure 3.08 – SALSA purifies both monomerically and dimerically by SDS-PAGE

*(**a**) Beta solenoid and SALSA stability is driven by complexing of leucine and threonine residues within a central pore. Alphafold3 model of accession number A0A1S3QEG0 and displayed in Pymol (educational licence)(**b**) SALSA was purified by gravity flow NiNTA HisTrap and dialysed overnight into 0.1M phosphate buffer, concentrated via a 25kda Vivaspin concentrator, and 1ml was injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1 M phosphate buffer, pH 6.0. An asymmetric large peak was observed from 9-15ml. (**c**) SDS-PAGE analysis of the resultant pooled fractions following SEC indicated the presence of monomeric and dimeric SALSA and TEV following TEV cleavage. (**d**) purification of SALSA was confirmed by intact TOF mass spectrometry, revealing a mass hit within 3da of the predicted mass. Given that the Histag was judged as unlikely to strongly influence protein behaviour, future purifications opted not to cleave the his-tag of this sample to prevent final sample contamination.*

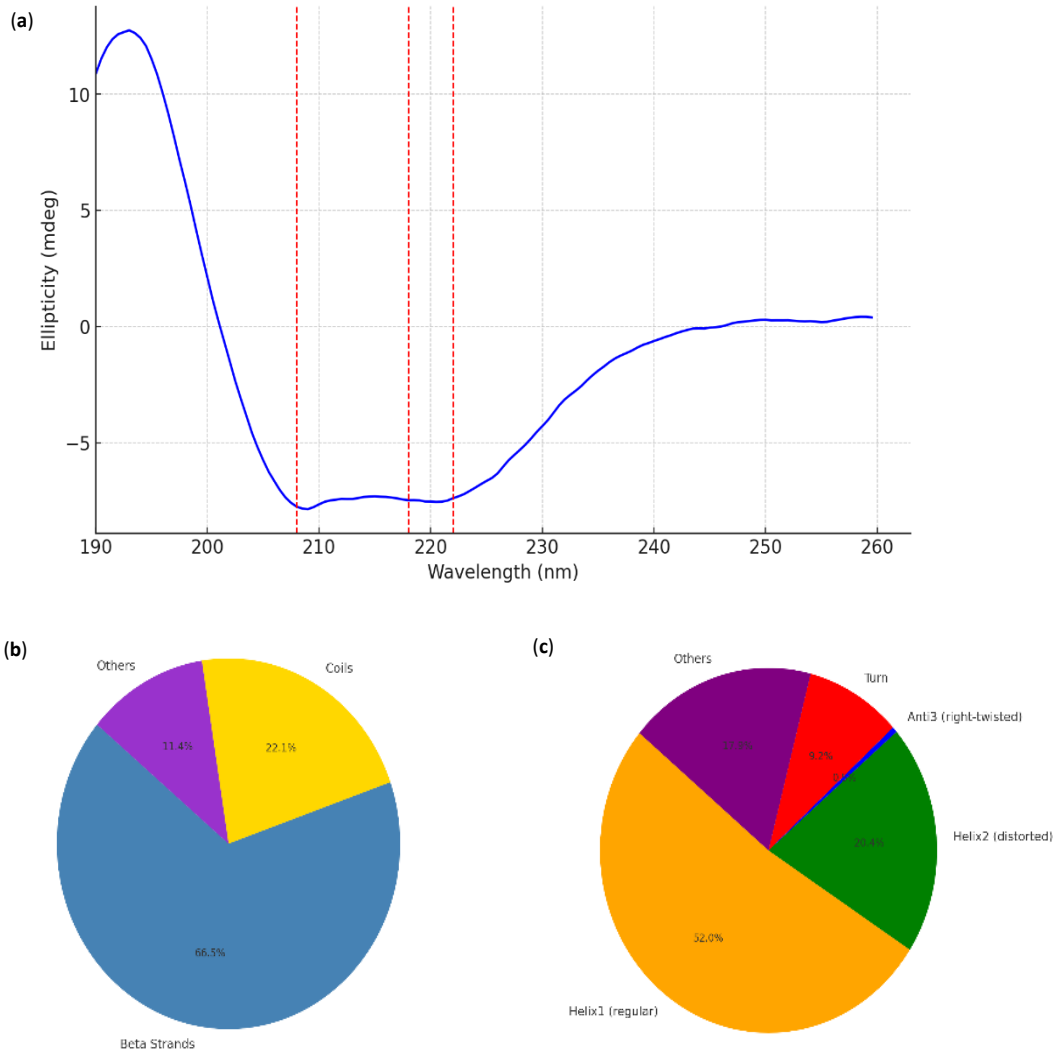


Figure 3.09 – SALSA purifies as an alpha helical protein.

*(**a**) CD spectra of 1.0mg/ml SALSA was recorded using a Jasco J-815 CD spectropolarimeter at 20 °C. The spectrum, obtained in 0.1 M phosphate buffer (pH 6.0), is characteristic of a predominantly α-helical structure. Key secondary structural wavelengths indicative of helical (208, 222nm) and beta sheet content (218nm) are indicated in red (**b**) The second structure of SALSA was predicted within the BestSel package using SALSA's Alphafold3 predicted structure. (**c**) The spectra of purified SALSA were submitted for analysis within the BeStSel package, which defines secondary structure components based on a derivative of DSSP. Beta-sheets are classified as parallel or antiparallel, with antiparallel sheets subdivided into left-hand twisted (anti1), relaxed (anti2), and right-hand twisted (anti3). The regular portion of alpha-helices (helix1) is distinguished from the distorted ends (helix2), 3_{10}-helices are categorized as "others," and the definition of turns follows that of DSSP. SALSA appears to purify as a majority alpha helical protein.*

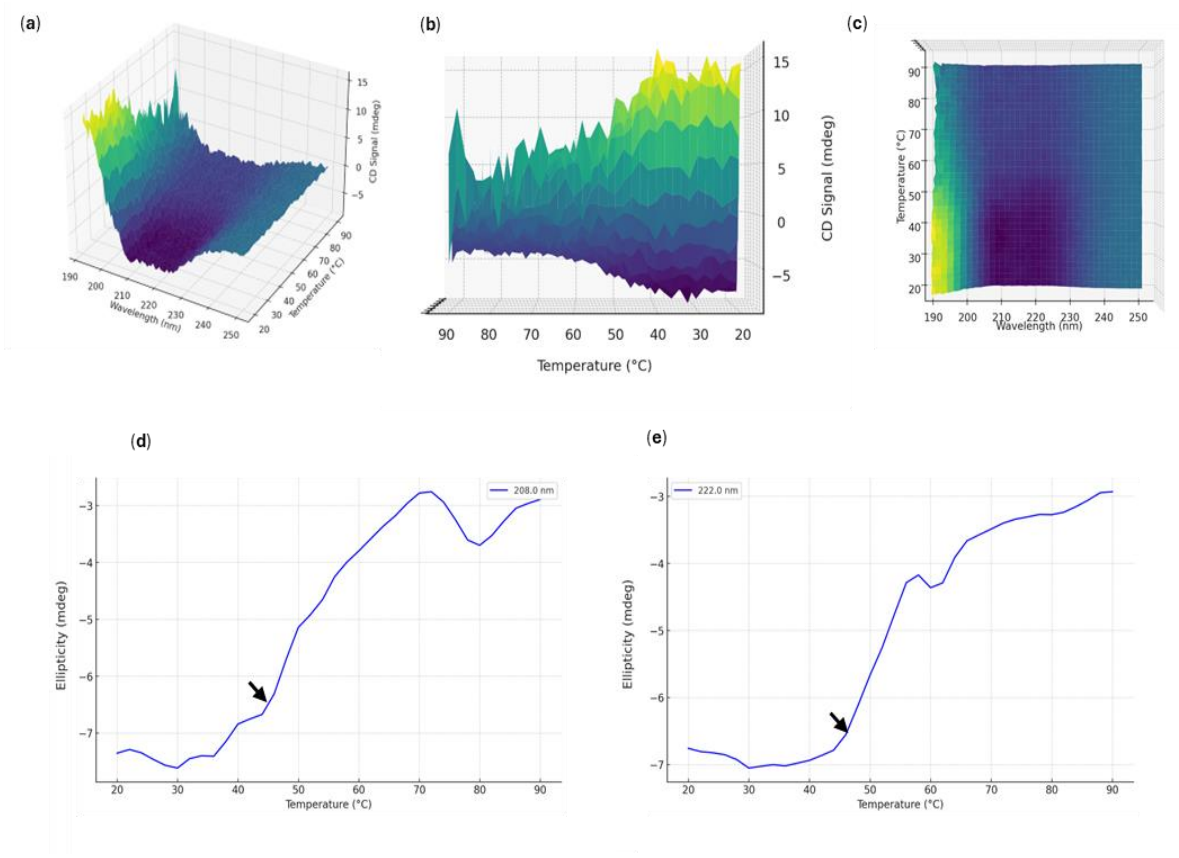


Figure 3.10 – SALSA undergoes cooperative unfolding.

*(**a-e**) 3D CD spectra of 0.1mg/ml SALSA is shown as recorded using a Jasco J-815 CD spectropolarimeter across a temperature range from 20-90°C at 2°C increments and 2 accumulations per increment. Spectra were plotted in Python and displayed in various viewpoints for analysis. Ellipticity as a function of temperature at 208 and 222nm, hallmarks of alpha helical structure are shown, revealing a thermal transition inflection point takes place from 40-50°C up to 70°C degrees (arrows) consistent with global loss of alpha structure.*

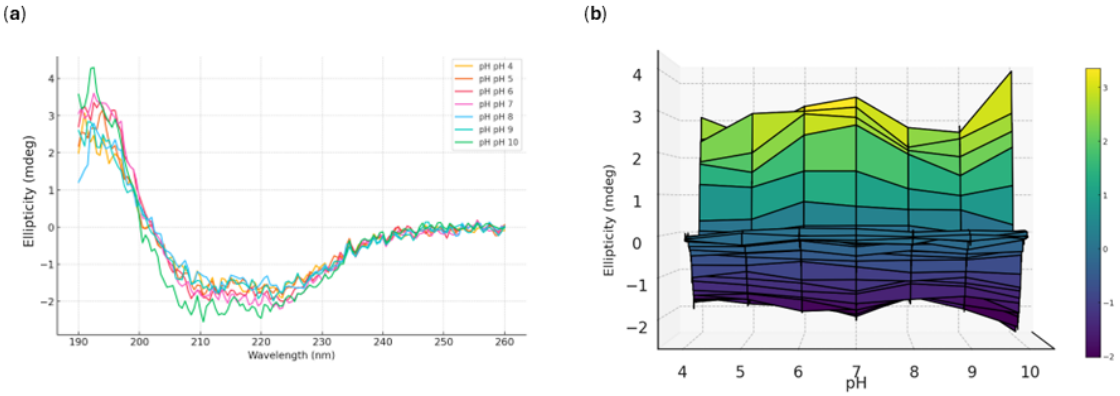


Figure 3.11 – SALSA is structurally stable against changes in pH from pH 4-10.

*(**a-b**) 3D CD spectra of 1mg/ml SALSA is shown as recorded using a Jasco J-815 CD spectropolarimeter across a pH range from pH 4-10 established by addition of 0.1M NaOH or HCl to each sample. Resulting spectra are averages from 8 accumulations per data point. Minimal changes in ellipticity are observed, consistent with SALSA being structural stable against pH changes in this range.*

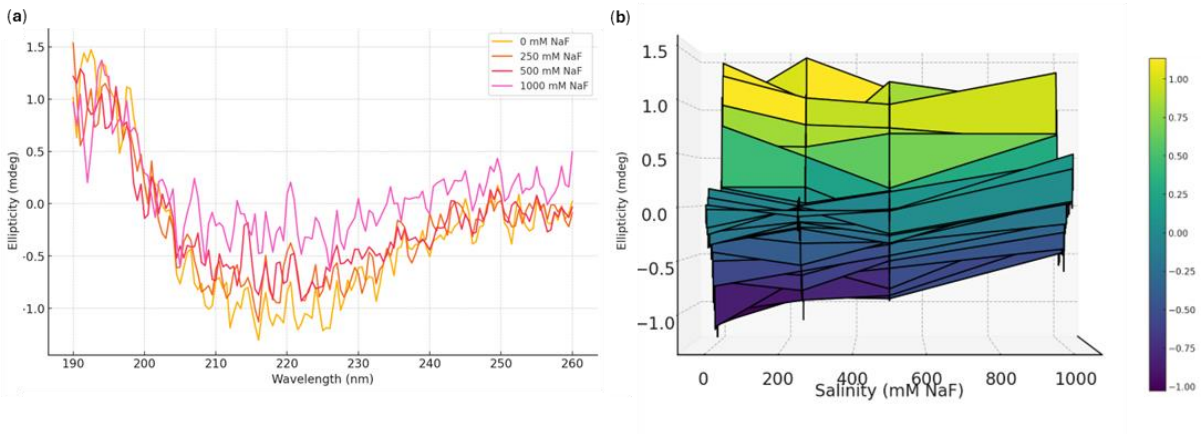


Figure 3.12 – SALSA begins to precipitate upon increasing salinity.

*(**a-b**) 3D CD spectra of 0.1mg/ml SALSA is shown as recorded using a Jasco J-815 CD spectropolarimeter across a pH range from a salinity range of 0-1000mM NaF established by addition of anhydrous NaF to 250, 500 and 1000mM. Resulting spectra are averages from 8 accumulations per data point. While low signal to noise challenges analysis of this sample, a trend is observed with decreasing ellipticity over the salinity range, indicating that the protein loses secondary structural content as salinity increases, resulting in precipitation and thus signal loss.*

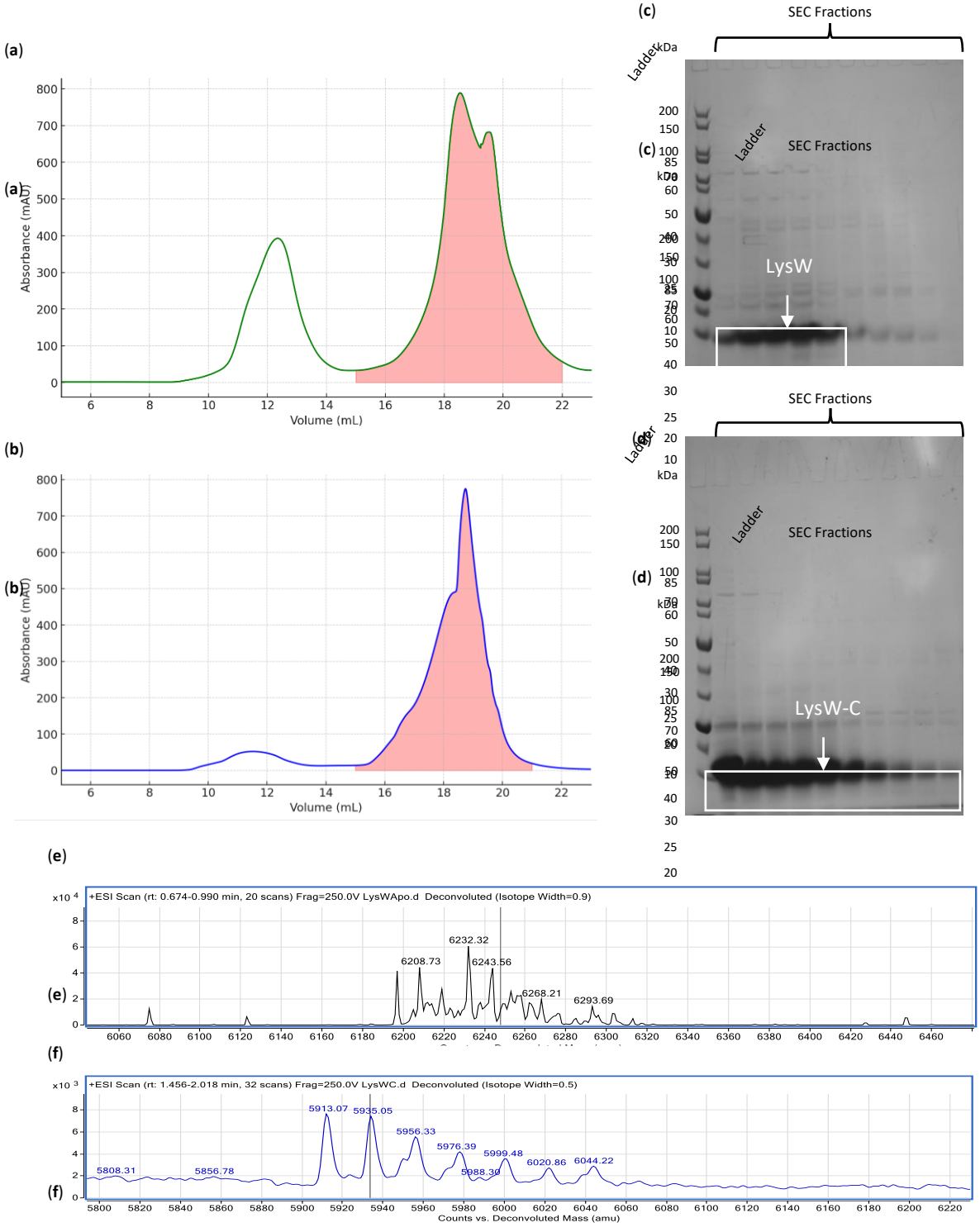


Figure 3.13 – LysW and Fe-LysW-C purify monomerically.

*(**a-b**) LysW and Fe-LysW-C were purified by gravity flow NiNTA HisTrap, dialysed overnight into 4l 0.1M phosphate buffer, concentrated via a 25kda Vivaspin concentrator, and 1ml was injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1 M phosphate buffer, pH 6.0. An asymmetric large peak was observed from 14-21ml (**a**) and 14-22ml (**b**). SDS-PAGE analysis of LysW (**c**) and Fe-LysW-C (**d**) following SEC indicated the presence of monomeric LysW with a band below the 10kda ladder, and dimeric LysW-C when run on the gel with a band between 10-15kda. Later investigations by NMR confirmed LysW-C is monomeric in solution. Successful purification of LysW and Fe-LysW-C was confirmed by intact qTOF mass spectroscopy with identification of a peak at 6208.73, compared to a predicted mass of 6208.71 for LysW (**e**) and a hit of 5913.07 compared to a predicted mass of 5913.63 for Fe-LysW-C (**f**)*

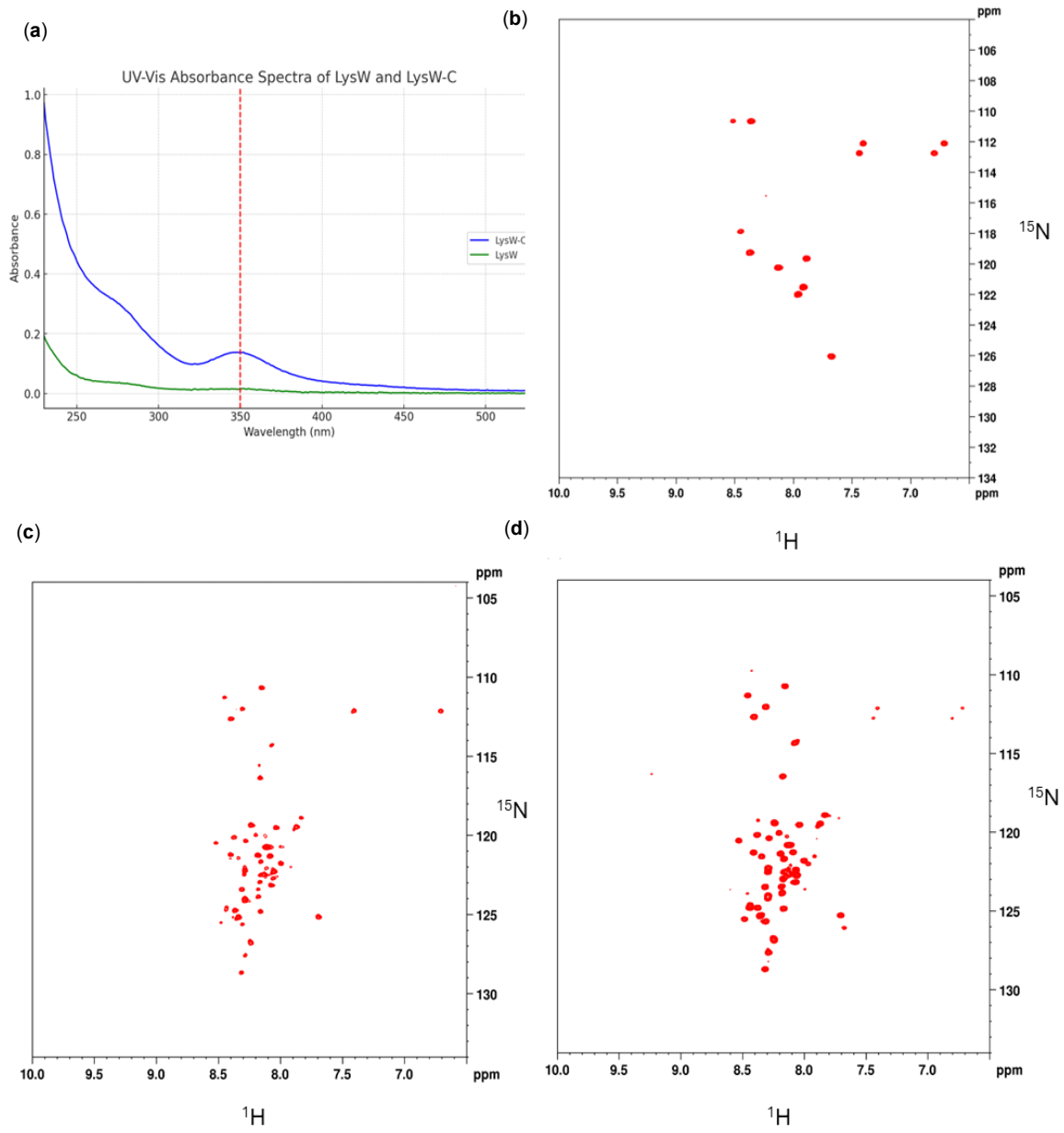


Figure 3.14 – Fe-LysW-C binds to iron

*(**a**) UV-visible spectroscopy of RbdC was recorded with a Cary 50 Bio spectrophotometer, showing absorption bands centred at approximately 350, consistent with a metal-sulfur binding site. (**b**) 2D ^{1}H-^{15}N HSQC spectrum of 0.65mg/ml Fe-LysW-C generated a spectrum with far fewer resonance than would be expected. Data was recorded at 20°C and pH 6.0 in 0.1M phosphate buffer with 32 scans per increment, 1276 total points (t_1), and 16.03ppm sweep width in the direct (^{1}H) dimension and 320 increments with a sweep width of 30.00ppm in the indirect (^{15}N) dimension. Iron was removed from Fe-LysW-C via TCA precipitation to create Apo-LysW-C and the sample was replaced into the spectrometer, revealing a spectra (**c**) significantly more consistent with spectra collected from LysW at 0.65mg/ml under identical conditions (**d**), confirming that Fe-LysW-C is metal bound at the point of purification and the loss of iron appears to result in reduced signal dispersion, and thus likely reduced protein structure. While the identity of the bound metal was not investigated further, based on the brick red sample colouration that persisted even after dialysis, I predicted that this metal was Fe^{3+}.*

contrast, LysW spectra showed approximately the expected number of signals (**Fig 3.14d**). Upon TCA precipitation, LysW-C regained a large number of LysW resonances (**Fig 3.14c**). Having confirm metal binding, LysW-C was subsequently called Fe-LysW-C for consistency.

CD spectra suggested that LysW possessed partial β-structure but also significant flexibility, indicated by a strong negative peak at ≈202nm (**Fig 3.15a**). BeStSel predicted this as a majority anti parallel β-sheet and helical mixture (**Fig 3.15b**)(Micsonai *et al.*, 2022). This was similar to the one observed for rubredoxin (**Fig 2.04**). Fe-LysW-C displayed slightly altered CD signature consistent with an increase in secondary structure content (**Fig 3.15a,c**). The predicted secondary structural content was more similar with the CD secondary content prediction of an X-ray crystallographic structure of a Zn-LysW-C protein homologue bound to zinc (PDB ID: 3WWL) (**Fig 3.15d**)

Co-operative unfolding could not be detected in thermal melts of LysW and Fe-LysW-C (**Fig 3.16**), suggesting that LysW adopt molten globular states. While Fe-LysW-C also did not undergo a thermal transition, this could also be explained by extreme structural stabilisation through iron binding, which prevented any thermal transition even within the tested range. pH titration experiments also showed that both LysW and Fe-LysW-C precipitated at acidic pH, with exposure to pH 4 causing complete protein precipitation in both proteins (**Fig 3.17**). For LysW this behaviour was also observed at pH 5, and with only partial recovery at pH 6 before regaining an ellipticity signal that was broadly consistent up to pH 10, indicate stability against alkaline pH. In the case of Fe-LysW-C, the protein acquired more of its native state ellipticity at pH 5 and full regained it by pH 6,

indicating that metal binding had resulted in increased stability against changes in pH in this protein. Notably there was an apparent change in ellipticity again at pH 7 before the trend in ellipticity resumes at pH 8. this sudden change was assigned as an experimental anomaly rather than reflecting a specific change for this protein that only happens at pH 7.

The observation that metal binding increases the ability of Fe-LysW-C to resist changes in local pH was also observed in salinity screens, whereas for LysW a steady reduction in signal is observed as anhydrous NaF was added, which was not observed in the Fe-LysW-C sample (**Fig 3.18**). This was assigned to a decreasing solubility of LysW as salinity increases, which was not seen in the more structured Fe-LysW-C and was most likely due to reduced solvent access for charged residues that would be affected by changes in salinity.

Overall, LysW demonstrates a partially β-structured, molten globule-like conformation that is significantly destabilized by acidic pH and slightly sensitive to salinity up to 1M NaF. While molten globular proteins typically adopt alpha helical secondary content, the assessment of molten globularity here is driven by LysW's differing CD spectra from Fe-LysW-C which indicates that it has less secondary structure, its lack of cooperative unfolding, and its poorly dispersed NMR spectra, all of which are consistent within at least a more molten globular-like fold. In contrast, Fe-LysW-C shows clear evidence of metal coordination, via its cysteine thiols, which correlates with an increase in secondary structure content and enhanced resistance to both acidic pH and high salinity. This indicates that cysteine-mediated metal binding provides a stabilising effect in an otherwise molten globule-like protein, although it does not yield a fully cooperative

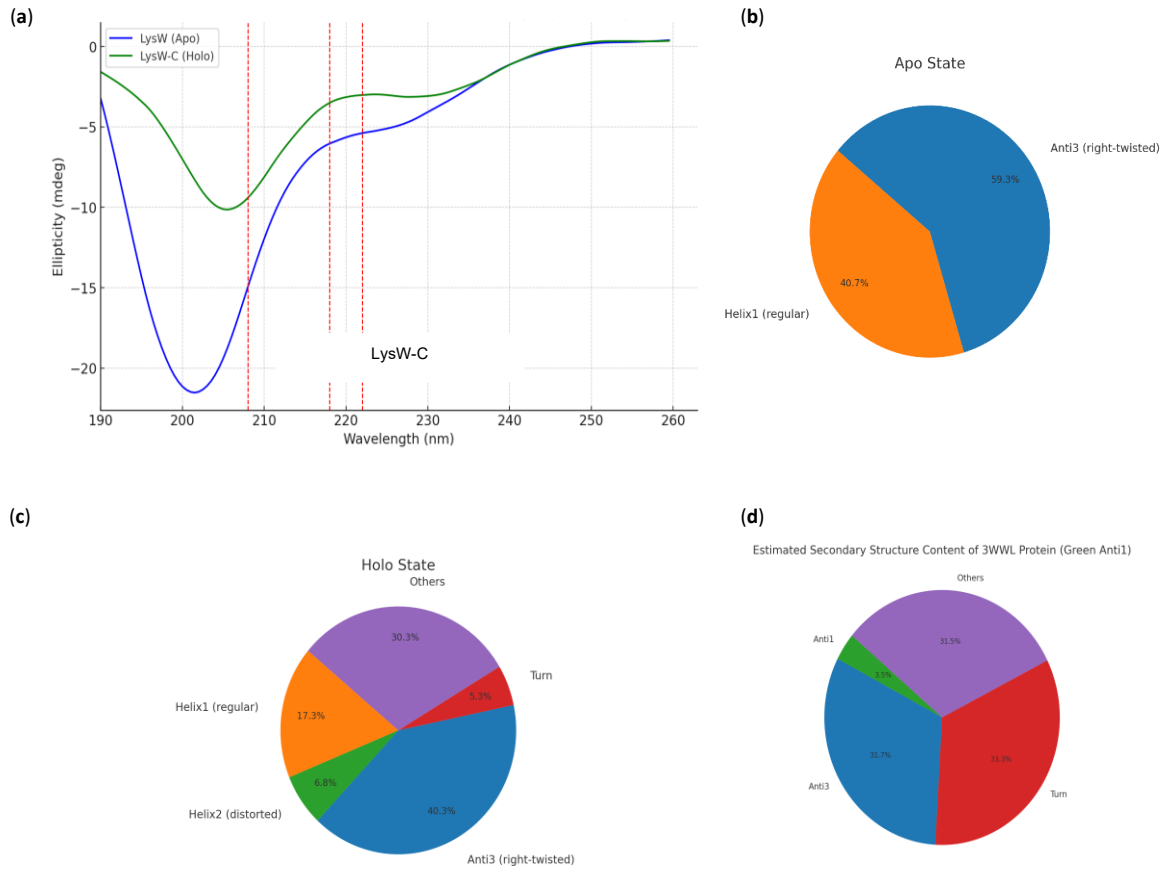


Figure 3.15 – Metal binding causing increased structure in Fe-LysW-C.

*(**a**) CD spectra of 1.0mg/ml LysW and Fe-LysW-C was recorded using a Jasco J-815 CD spectropolarimeter at 20 °C. The spectrum for LysW is consistent with a protein containing some level of secondary structure, with disordered regions. Fe-LysW-C on the other hand gave a spectrum more consistent with a mixed helical/beta sheet protein. Key secondary structural wavelengths indicative of helical (208, 222nm) and beta sheet content (218nm) are indicated in red. The resulting spectra were submitted for analysis within the BeStSel package, which defines secondary structure components based on a derivative of DSSP. Beta-sheets are classified as parallel or antiparallel, with antiparallel sheets subdivided into left-hand twisted (anti1), relaxed (anti2), and right-hand twisted (anti3). The regular portion of alpha-helices (helix1) is distinguished from the distorted ends (helix2), 3_{10}-helices are categorized as "others," and the definition of turns follows that of DSSP. The predicted structural percentages for LysW (**b**), Fe-LysW-C (**c**), and the BeStSel-predicted secondary structural percentage of the 3WWL crystal structure of Lysine Biosynthesis protein from Thermus thermophilus (**d**) are shown. structure of Lysine Biosynthesis protein from Thermus thermophilus (**d**) are shown.*

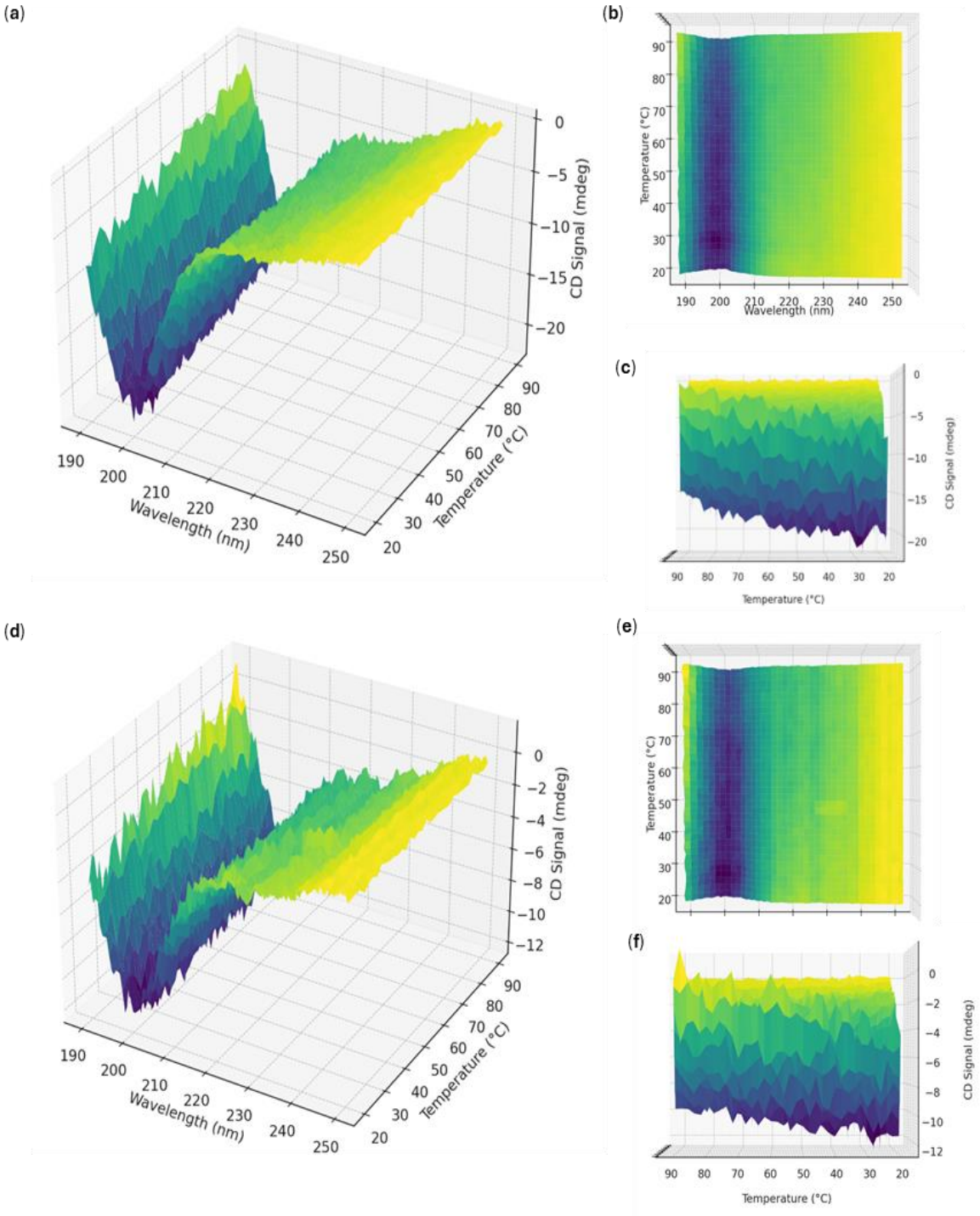


Figure 3.16 – LysW and Fe-LysW-C does not undergo cooperative unfolding with increasing temperature.

*3D CD spectra of 1mg/ml LysW (**a-c**) Fe-LysW-C (**d-f**) is shown as recorded using a Jasco J-815 CD spectropolarimeter across a temperature range from 20-90°C at 2°C increments and 2 accumulations per increment. Spectra were plotted in Python and displayed in various viewpoints for analysis. No significant rapid onset change in ellipticity is observed, suggest that both proteins do not unfold via cooperative unfolding and possess molten globular structures.*

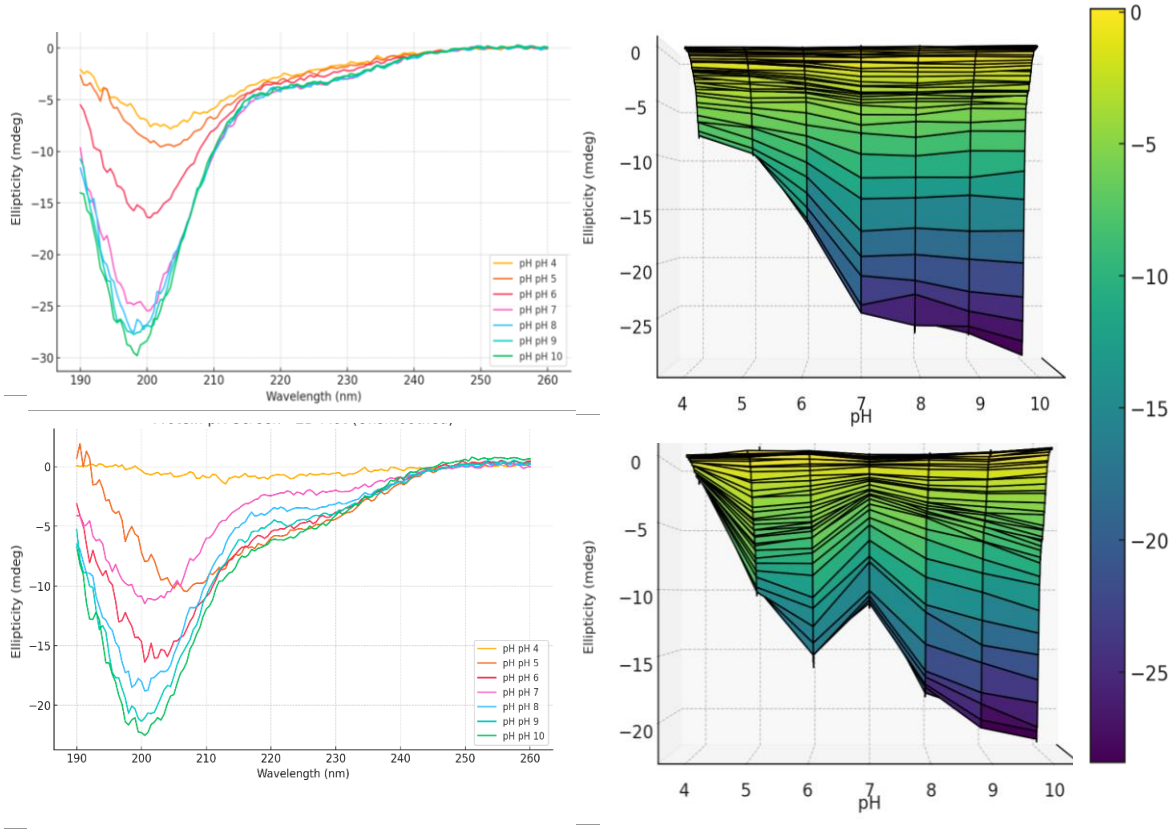


Figure 3.17 – Metal binding increases pH stability in Fe-LysW-C.

*Overlays and 3D Circular dichroism (CD) spectra of 1mg/ml LysW (**a-b**) and Fe-LysW-C (**c-d**). Spectra were recorded using a Jasco J-815 CD spectropolarimeter across a pH range from pH 4-10 established by addition of 0.1M NaOH or HCl to each sample. Resulting spectra are averages from 8 accumulations per data point. Both proteins experienced significant loss of ellipticity at pH 4 and 5, with the presence of a metal ion reducing the degree Fe-LysW-C's loss of signal at pH 5, suggesting that metal binding stabilises to the protein at pH 5, compared to LysW.*

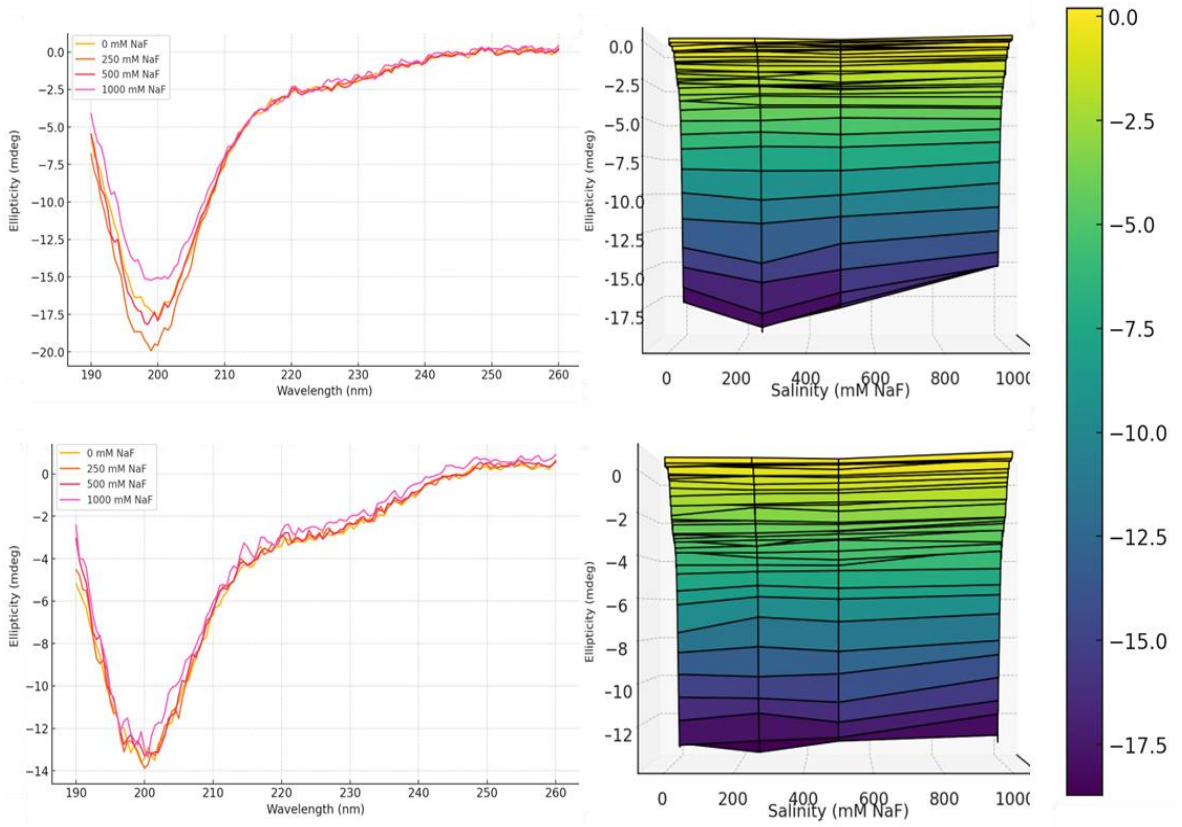


Figure 3.18 – Metal binding increases Fe-LysW-C resistance to salt induced protein unfolding.

*(**a-b**) 3D CD spectra of 1mg/ml LysW (a-b) and Fe-LysW-C is shown as recorded using a Jasco J-815 CD spectropolarimeter from a salinity range of 0-1000mM NaF established by addition of anhydrous NaF to 250, 500 and 1000mM solution concentration. Resulting spectra are averages from 8 accumulations per data point. An increase in ellipticity at 1000mM NaF is observed in 1000mM, indicating loss of signal due to protein precipitation which is not observed in the Fe-LysW-C spectra, suggest that metal binding increases resistant to salt induced protein unfolding in this protein.*

tertiary fold. The extra structural order and environmental resilience point to the importance of cysteine in enabling early proteins to achieve at least partial stability and functionality.

The ferredoxin-like protein is a molten globule and sensitive to acidic pH.

Whilst annotated by UniProt as Ferredoxin, given its lack of an *E.coli* specific Fe-S cluster recruitment signal, needed for Fe_4S_4 complex incorporation into Ferredoxins, the expression of this protein by *E.coli* as a functional metalloprotein was unlikely (Wachnowsky *et al.*, 2018). Ferredoxin was purified by SEC (**Fig 3.19a**) and confirmed by SDS-PAGE (**Fig 3.19c**) and QTOF mass spec (**Appendix A - Figure 03**). Although purified as a monomer (**Fig 3.19d. Appendix A - Figure 04**), metal binding was not observed by UV (**Fig 3.19b**) nor did the protein display the characteristic holo-ferredoxin coloration. This lack of metal binding was assigned to an incompatibility of this protein's recruitment of the *E.coli* Fe_4S_4 recruitment pathway which is likely very different from the *Planctomycetes bacterium* that this protein originates (Hidese *et al.*, 2014).

Similarly to LysW (**Fig 3.15a**), CD spectra of ferredoxin were consistent with a mixed α/β protein (**Fig 3.20**),and no cooperative unfolding transition could be detected indicative of a molten globular structure (**Fig 3.21**). pH titrations again showed very behaviour similar to the LysW constructs, with complete protein precipitation at pH 4, and strong destabilisation at pH 5 with structural recovery occurring by pH 6. In contrast, Ferredoxin exhibited a significant reduction of ellipticity at 198nm at pH 10 consistent with an increase in flexibility of the protein at this pH (**Fig 3.22**). Salinity screens revealed unchanged ellipticity across the 0-1000mM NaF, suggesting that this protein was stable to changes of salinity in

this range and tolerated ionic strength better than LysW in the apo state (**Fig 3.23**).

Thus, LysW, Fe-LysW-C, and the ferredoxin-like protein adopt molten globule-like states and are destabilized by acidic pH.

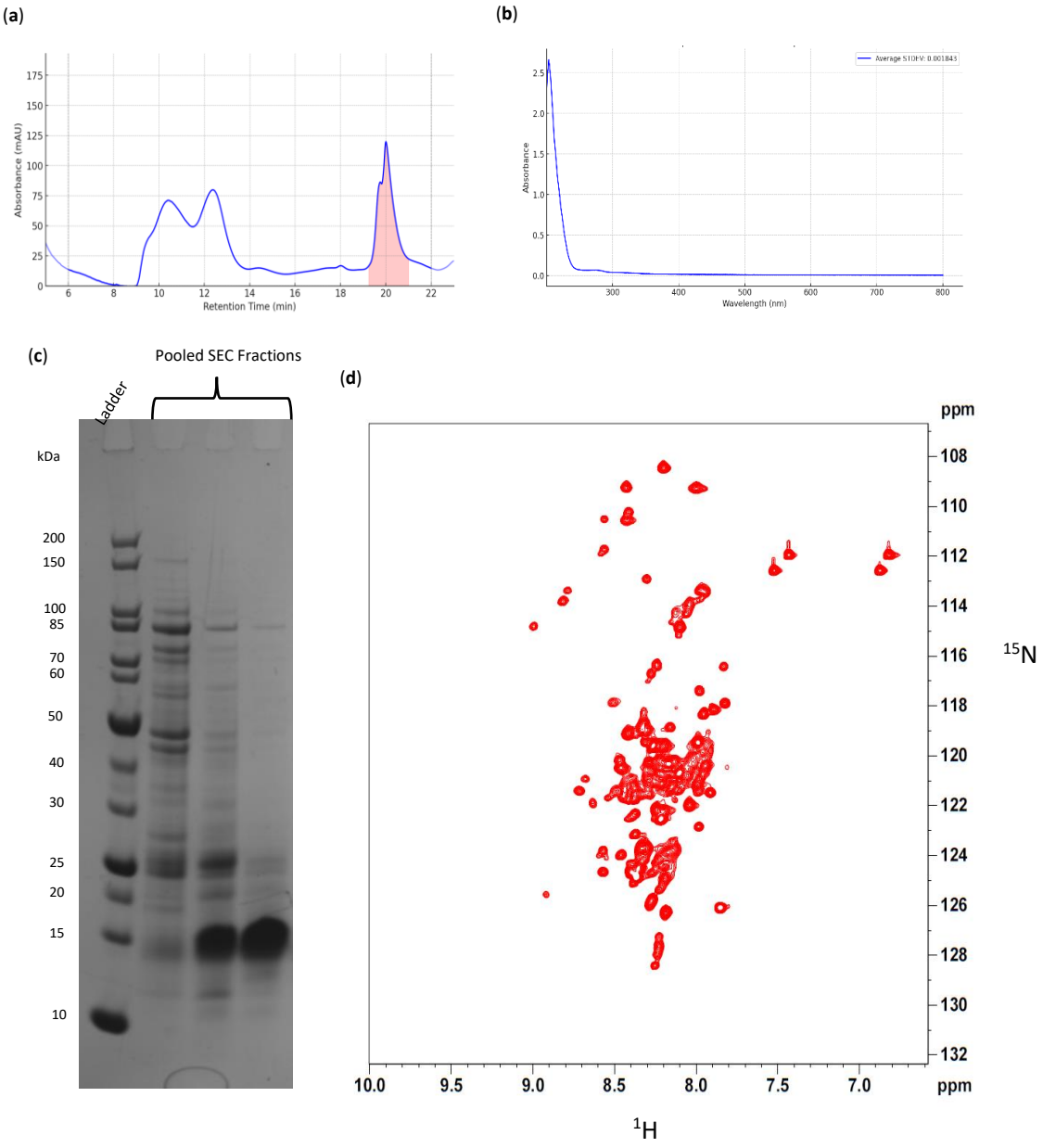


Figure 3.19 – Ferredoxin is purified in the apo-state when expressed in *E.coli*.

*(**a**) Ferredoxin was purified by gravity flow NiNTA HisTrap, dialysed overnight into 4l 0.1M phosphate buffer, concentrated via a 5kda Vivaspin concentrator, and 1ml was injected on to a Superdex 75 Increase 10/300 GL column equilibrated with 0.1 M phosphate buffer, pH 6.0. An asymmetric peak was observed from 19-21ml similar to LysW elation profile which had a similar mass within 2kda. (**b**) UV-visible spectroscopy of Ferredoxin (blue) was recorded with a Cary 50 Bio spectrophotometer, showing no significant of iron sulfur absorption bands from 300-800nm, consistent with no iron-sulfur cluster being present. (**c**). SDS-PAGE analysis of Ferredoxin following SEC indicated the presence of dimeric Ferredoxin with a strong band observed at 15kda, consistent with the dimeric mass of Ferredoxin of 16.4kda. (**d**) 2D ^{1}H-^{15}N HSQC spectrum of Ferredoxin generated a spectrum consistent with the number of resonances (78) expected from Ferredoxin (79), indicative of monomeric protein but also lacking any metal binding capability (peak picked spectra shown in **Appendix A - Figure 5**). Data was recorded at 20°C and pH 6.0 in 0.1M phosphate buffer with 32 scans per increment, 1276 total points (t_1), and 16.03ppm sweep width in the direct (^{1}H) dimension and 320 increments with a sweep width of 30.00ppm in the indirect (^{15}N) dimension.*

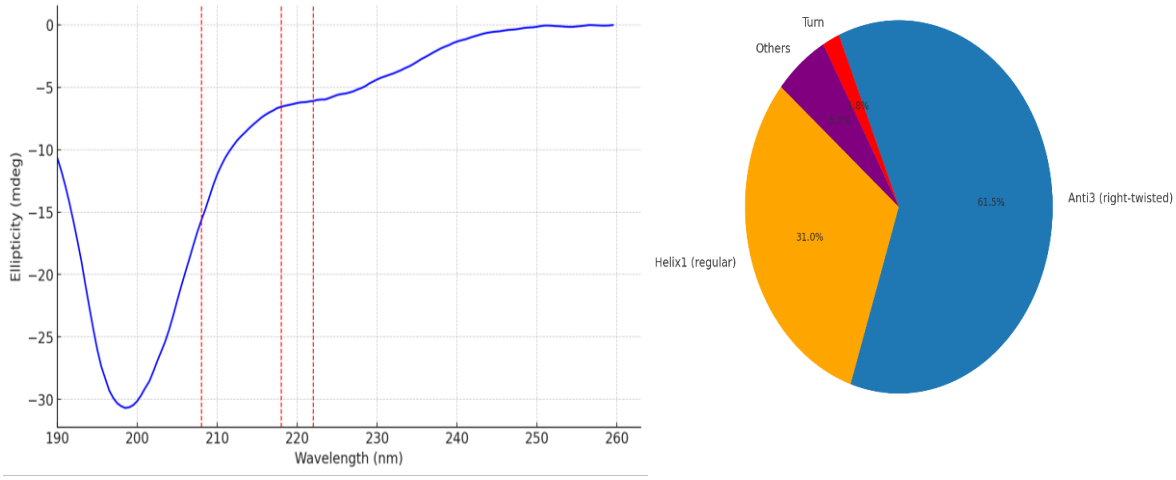


Figure 3.20 – Ferredoxin adopts a mixed α/β structure upon purification in *E.coli.*

(a) CD spectra of 1.0mg/ml Ferredoxin was recorded using a Jasco J-815 CD spectropolarimeter at 20 °C. The spectrum for LysW is consistent with a protein containing some level of secondary structure, with disordered regions. Key secondary structural wavelengths indicative of helical (208, 222nm) and beta sheet content (218nm) are indicated in red. The resulting spectra were submitted for analysis within the BeStSel package, which defines secondary structure components based on a derivative of DSSP. Beta-sheets are classified as parallel or antiparallel, with antiparallel sheets subdivided into left-hand twisted (anti1), relaxed (anti2), and right-hand twisted (anti3). The regular portion of alpha-helices (helix1) is distinguished from the distorted ends (helix2), 3_{10}-helices are categorized as "others," and the definition of turns follows that of DSSP. This suggested a mixed helical/anti parallel beta sheet structure, similar to that predicted for LysW.

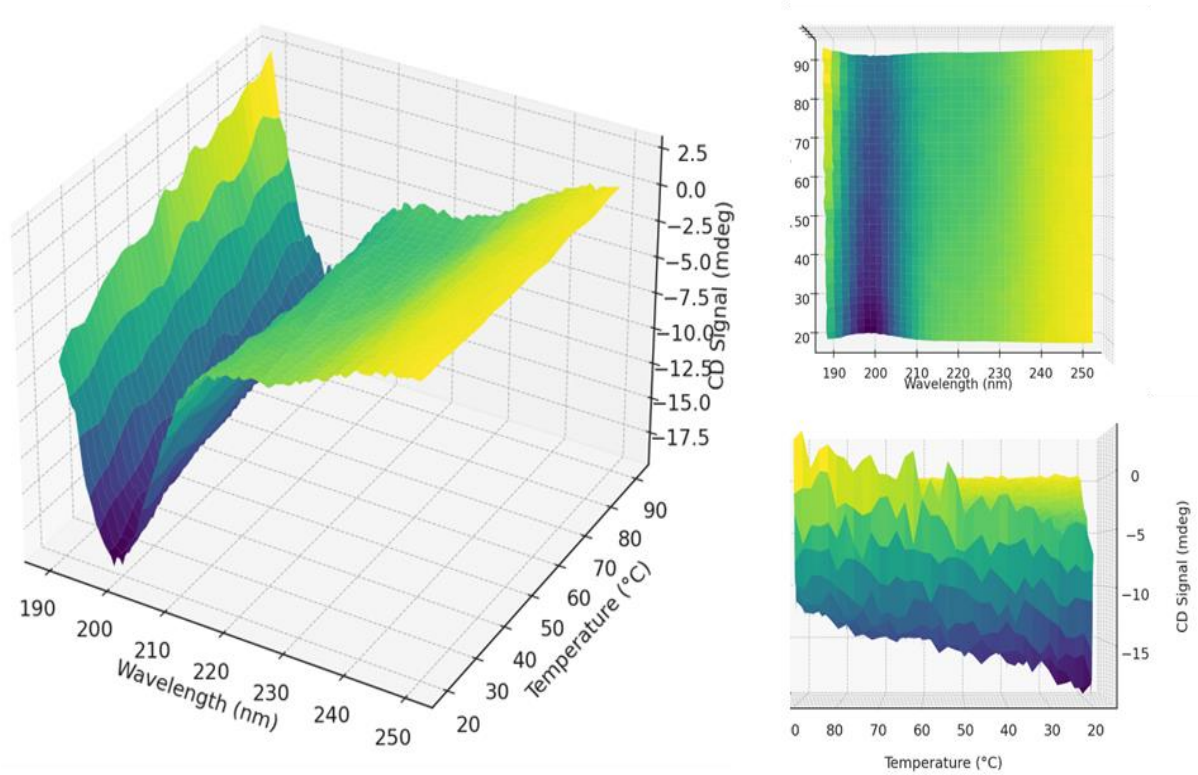


Figure 3.21– Ferredoxin does not undergo cooperative unfolding up to 90℃.

*3D CD spectra of 1mg/ml Ferredoxin (**a-c**) is shown as recorded using a Jasco J-815 CD spectropolarimeter across a temperature range from 20-90°C °C at 2°C increments and 2 accumulations per increment. Spectra were plotted in Python and displayed in various viewpoints for analysis. No significant rapid onset change in ellipticity is observed, suggest that ferredoxin does not unfold via cooperative unfolding and possesses a molten globular structure.*

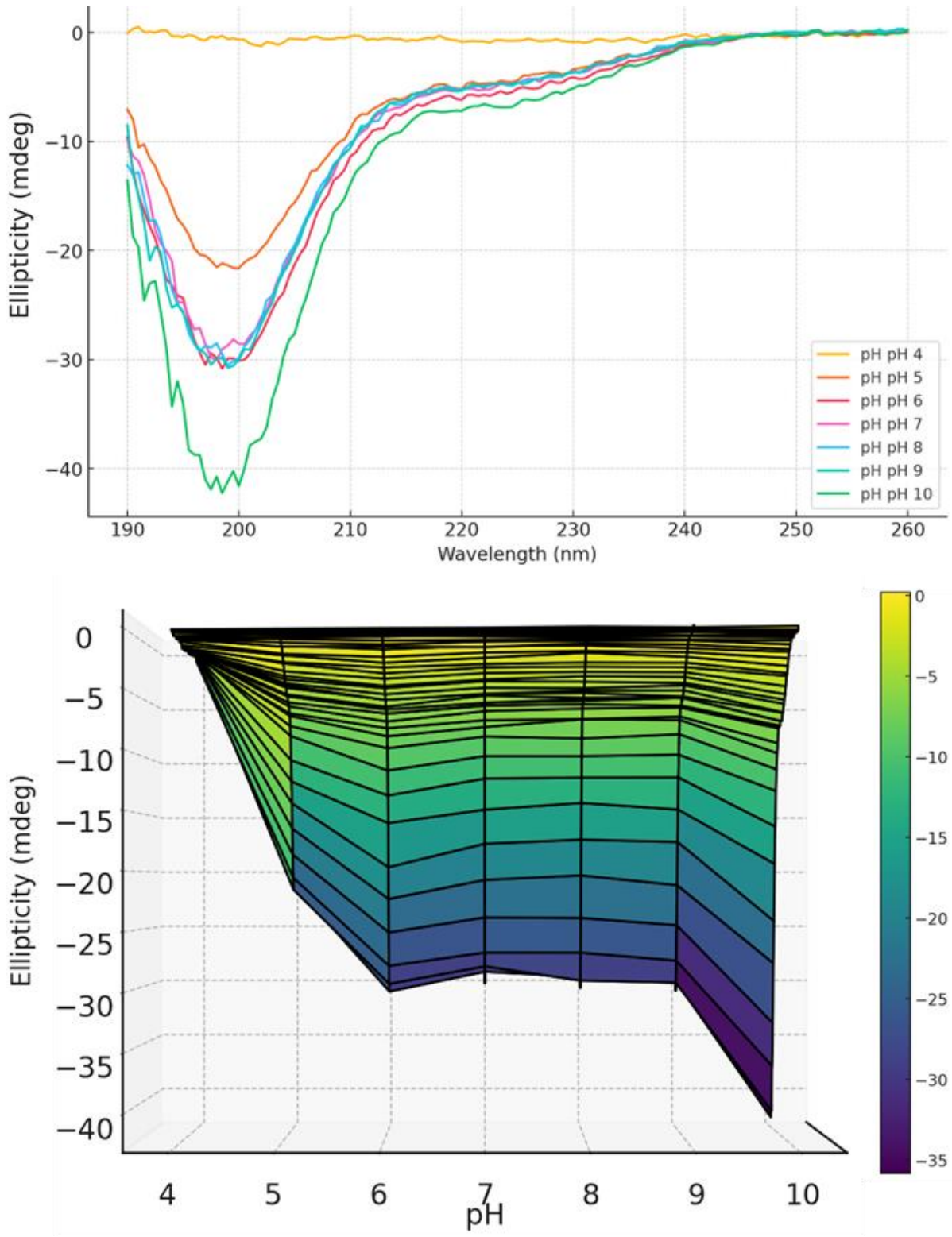
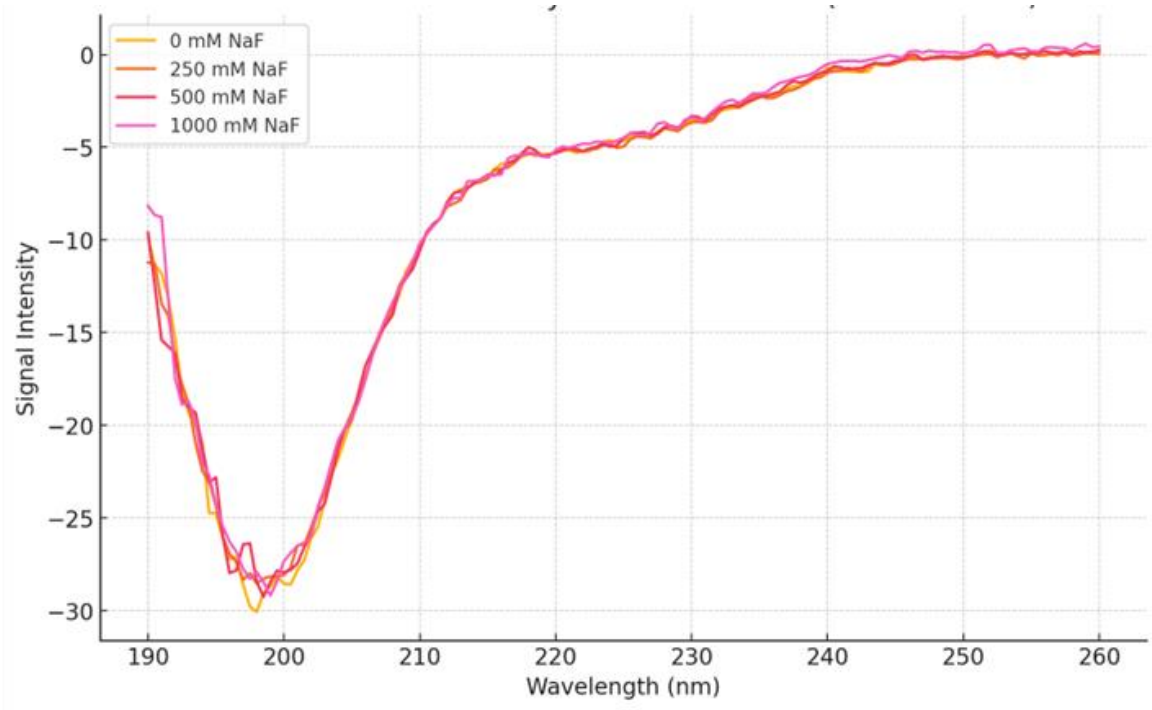


Figure 3.22 – Ferredoxin is sensitive to acidic pH's.

*Overlay (**a**) and 3D CC (**b**) spectra of 1mg/ml Ferredoxin. Spectra were recorded using a Jasco J-815 CD spectropolarimeter across a pH range from pH 4-10 established by addition of 0.1M NaOH or HCl to each sample. Resulting spectra are averages from 8 accumulations per data point. Loss of signal at pH 4 was consistent with precipitation of the protein similarly to that observed in LysW and Fe-LysW-C.*



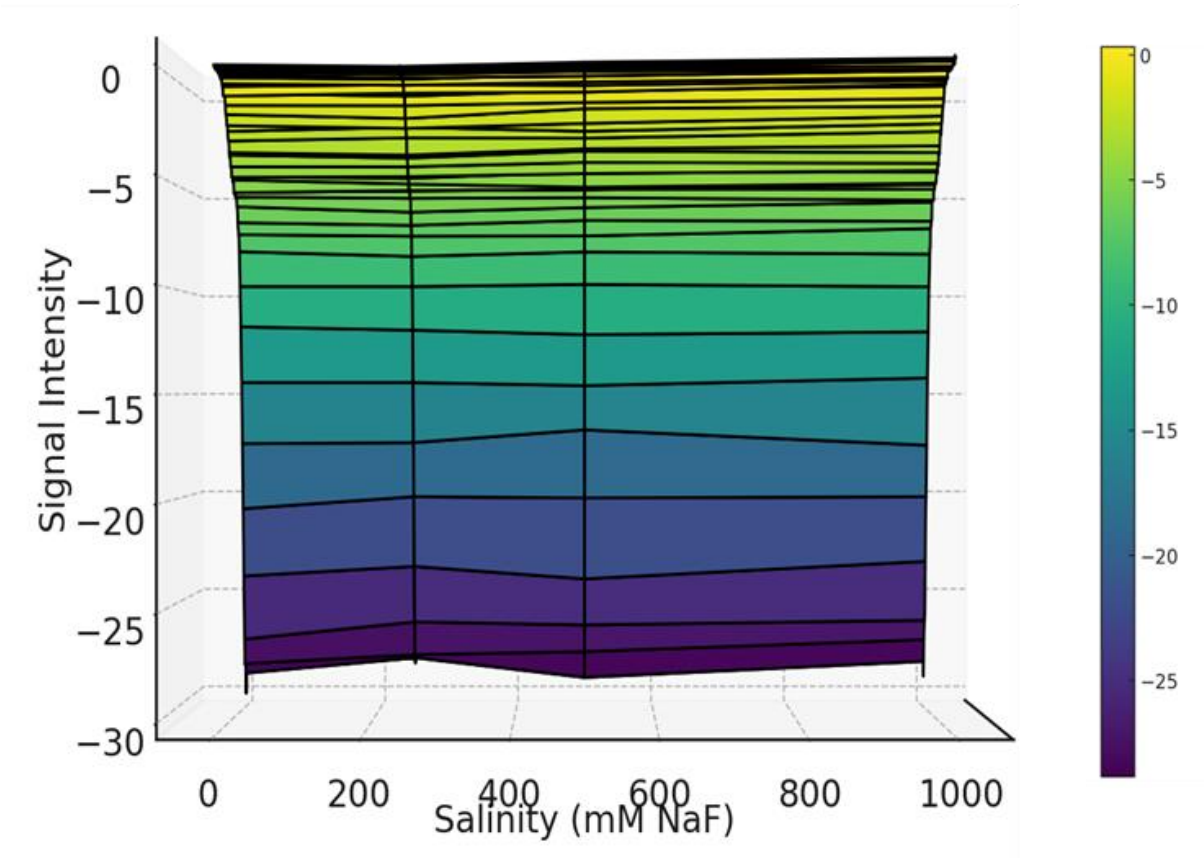


Figure 3.23 – Ferredoxin is insensitive to changes in salinity.

*(**a-b**) An overlay and 3D CD spectra of 1mg/ml Ferredoxin is shown as recorded using a Jasco J-815 CD spectropolarimeter from a salinity range of 0-1000mM NaF established by addition of anhydrous NaF to 250, 500 and 1000mM solution concentration. Resulting spectra are averages from 8 accumulations per data point. No significant change in ellipticity is observed over the salinity range, indicating the ferredoxin is insensitive to changes to salinity in this range.*

Discussion

The ability of abiotic proteins to form transient structures determines their functional capacity, which would be needed to bring about an RNA world within the RNA-peptide world model (Tagami and Li, 2023). In this chapter, the structural stabilities of proteins composed of abiotic amino acids were determined using an integrated bioinformatics and *in vitro* experimental approach. The proteins identified through this approach potentially offer a wider view on the 'foldability' of abiotic like proteins than the single-rubredoxin approach in Chapter 2. In addition, the model proteins provide a test of *a priori* predictions of abiotic protein folding characteristics.

A large-scale bioinformatics approach identified 38 non-homologous proteins with ≥95% abiotic amino acid content (Cys-) and 39 proteins when cysteine was included (Cys+). This analysis improved on earlier attempts by significantly increasing the scale of analysis, with 250 million sequences interrogated compared the 10 genomes used by McDonald and Storrie-Lombardi. The resulting proteins included several distinct structural classes, namely β-barrels, β-solenoids, α-helical proteins, and α/β folds.

The observed structural tendencies of abiotic-like proteins did not align with my *a priori* predictions but were consistent with the propensity level data that abiotic proteins would be depleted in β-sheet content and enriched in loops, turns, and helical content. This consistency suggests that despite the variability in unstructured sequences the overall folding behaviour of these proteins is governed by the type and abundance of amino acids available. Scaled propensities for secondary structure, based on meteoritic amino acid

abundances, indicated that helical and unstructured regions were dominant, while β-sheet content was reduced. However, a significant number of β-solenoids were still identified in my bioinformatic analysis. These were attributed to a limited need for biotic amino acids to drive folding for these structures, which instead fold by hydrophobic interactions of abiotic amino acids. While this may suggest that such structures are an easy route by which early proteins could achieve structure, their requirement for high sequence conservation negates this point. In total these observations suggest that abiotic-like proteins are biased towards helical and disordered regions most likely due to the abiotic enrichment in alanine, which has the highest propensity among all amino acids for helical structure, and glycine which has a high propensity for loops and turns. Notably the inclusion of cysteine in the analysis did not significantly affect the structural outcomes, implying that cysteines inclusion in the abiotic amino acid set might be strictly as a functional contributor rather than a structural stabiliser. Overall, these data provide support the abiotic proteins being predominantly composed of loose strands and coils with helical content.

Abiotic-like proteins are pH sensitive and possess molten globular structures.

SALSA, while initially identified as a β-solenoid candidate, proved difficult to purify without contaminants. As a result, it was not possible to draw confident conclusions, so I turned to LysW, Fe-LysW-C, and the ferredoxin-like protein.

Purification of LysW and its cysteine-containing counterpart, Fe-LysW-C, was significantly more successful and revealed a mild structural stabilisation effect of cysteine that enhanced the diversity of structural elements. Both proteins exhibited β-barrel structures with molten globular characteristics, and Fe-LysW-

C displayed increased resistance to acidic conditions and salinity changes compared to LysW, reflecting the stabilising effects of metal binding facilitated by cysteine thiols. While this increased stability was small, even minor increases in folded stability may provide enough structural coherence to enable ligand and substrate binding for catalysis. The ferredoxin-like protein demonstrated similar behaviours, exhibiting molten globular characteristics and acidic pH sensitivity. However, its cysteine-rich sequence did not provide the structural coordination necessary to support spontaneous iron-sulfur cluster formation. An alternative explanation for this however might be that the protein does not undergo thermal unfolding until above the tested thermal range. However, on the basis of evidence from the poorly dispersed NMR spectra, and its overall limited amount of secondary structure, I assess this explanation as less likely than the argument that it is a molten globular like protein.

The structural trends observed across LysW, Fe-LysW-C, and the ferredoxin-like protein (**Table 3.01**) align with the prediction that proteins with abiotic amino acid composition will be sensitive to acidic pH (**Thesis hypothesis 2**), adopt molten globule states (**Thesis hypothesis 3),** and be insensitive to changes in salinity (**Thesis Hypothesis 4**). As predicted, all proteins possessed molten globular-like states were sensitive to pH from pH 4-5 and the CD spectra were largely unchanged up to 1M NaF. This confirms that multiple acidic side chains with minimal buffering from basic residues can lead to strong pH-dependent instability and that the already looser molten globular folds are not strongly destabilised by moderate changes in salinities.

		Sensitive to acidic pH (4-5)	Sensitive to increasing salinity	Capable of and stabilised by metal ion coordination	Molten globular structure
Abiotic-Like Protein	SALSA	No	Yes	N/A	No
	LysW	Yes	Yes		Yes
	Fe-LysW-C	Yes	No	Yes	Potentially
	Ferredoxin	Yes	No	No	Yes
Tally		3/4	2/4	1/2	3/4

Table 3.01 – Summary table of observed characteristics in abiotic-like proteins
Fe-LysW-C is listed as potentially as I was unable to distinguish if metal binding had increased thermal stability to beyond the tested range, or if the protein retained an apo-like molten globular-like folding state.

The observed pH sensitivity suggests that the activity of abiotic proteins would be confined to neutral and alkaline environments. It is likely that low pH results in aspartate sidechain protonation, which reduces protein solubility and promotes thiol-based reactions (Mirceta *et al.*, 2013). Notably a growing amount of literature also points to alkaline areas being more early life than acidic locations. For example, Fujikawa et al., investigated the pH stability of previously reconstructed proteins from LUCA, namely nucleoside diphosphate kinases and ribosomal uS8s proteins, which suggested these proteins function preferentially in alkaline environments compared to acidic ones (Fujikawa *et al.*, 2024).

My findings suggest that while the abiotic amino acid set could support basic structural elements, they do not provide a true 'foldable set.' As such, functional abiotic proteins likely required additional stabilising features, such as metal coordination or environmental constraints, e.g., high salinity, to achieve greater complexity.

The presence of cysteine in the environment emerges as potentially significant, not just for added protein stabilisation, but also for its potential to support metal-mediated functionality with abiotic proteins. These findings provide support for the hypothesis that abiotic-like proteins could achieve partial folding sufficient for rudimentary functions, such as ligand binding or catalysis, even in the absence of fully cooperative structures.

Evaluating the hypothesis of metal-guided protein folding

In Chapter 1, I proposed that cysteine could enable the folding of abiotic proteins by creating a metal-guided protein folding pathway whereby transition

metal ions drove structural collapse as opposed to the hydrophobic effect. The data collected in this chapter does not directly confirm the hypothesis of metal-guided protein folding as a nucleation mechanism for abiotic proteins. While Fe-LysW-C demonstrated slightly greater secondary structure content, metal binding, these effects were insufficient to induce fully cooperative folding observable by CD. Similarly, the ferredoxin-like protein, despite its cysteine-rich sequence, lacked the structural organization required to coordinate transition metals effectively. This outcome demonstrates the challenge of replicating metal-guided folding in systems like this. Future work using de novo sequence design or modeling is needed to properly interrogate this possibility.

Limitations

A potential limitation of this chapter's work is that it relies on modern proteins with high abiotic amino acid content as proxies for abiotically synthesised proteins. Although these proteins provide useful insights into the folding behaviours of sequences heavily enriched in abiotic residues, they are still products of billions of years of evolutionary optimisation. This evolutionary history introduces structural adaptations, sequence biases, and stabilising interactions that may not have been accessible in early Earth conditions. For example, the hydrophobic interactions observed in β-solenoids represent evolutionary adaptations rather than intrinsic folding tendencies of the abiotic amino acid set. Similarly, beta amyloids are often used as examples of small peptide structures that can create functional structures, however the most common amino acids in beta amyloids are among the least abundant abiotically (Rout *et al.*, 2022). A viable alternate route that I could have exploited in this study was to conduct

ancestral sequence reconstruction on previously identified ancient proteins to create abiotic like proteins that would have overcome this limitation.

Additionally, this chapter focuses solely on the structural properties of abiotic-like proteins, without including functional tests that would provide stronger constraints on the impact of cysteines incorporation into Fe-LysW-C. While structural stability is a prerequisite for function, the catalytic or binding capabilities of these proteins remain unexplored.

Another important limitation of the Chapter 3 approach is its heavy reliance on AlphaFold2 predictions. Candidate proteins were ranked almost entirely on AlphaFold confidence scores, chiefly pLDDT, or on metrics derived from the predicted models. The sequences selected here were intentionally enriched in abiotic amino acids and are, by design, poorly represented in structural databases. For such out-of-distribution inputs, even a high pLDDT value cannot guarantee correct global folding topology prediction. This screening approach should therefore be viewed as ‘triage’ step only which I then expanded through experimental interrogation of each protein.

Finally, the experimental section of this work was limited by the inability to purify two of the five selected proteins (PAST and SALSA) in sufficient quality for detailed analysis. The remaining three proteins, LysW, Fe-LysW-C, and the ferredoxin-like protein, were successfully purified and characterised, but this reduced the protein fold diversity and therefore limits the ability to generalise my interpretations extensively.

The bioinformatic and experimental results in this chapter, therefore, illustrate what is possible within the abiotic amino-acid set but cannot define the full range of folding behaviours that might have existed on the early Earth.

Conclusion

In summary, the findings presented in this chapter contribute to a growing body of evidence that abiotic-like proteins exhibit inherent structural and functional limitations. The molten globular states observed across multiple proteins, coupled with their sensitivity to pH and insensitivity to changes in salinity demonstrates that pH represents the greatest barrier to abiotic protein function, and that acidic conditions are likely directly destabilising to many abiotic proteins.

Notably, the introduction of cysteine imparts a measurable increase in structural order and environmental resilience by enabling metal binding, as observed in Fe-LysW-C. Specifically, the thiol-mediated coordination of iron correlated with enhanced secondary structure content and improved tolerance to both acidic pH and high salinity relative to LysW. Although these effects were modest, they support the view that cysteine's thiol chemistry could have been essential on early Earth for stabilising early proteins by conferring rudimentary structure and supporting metal driven protein folding. Given the observation that cysteines increased structural order here, and its requirement for simple metal binding identified in chapter 2, I now move on to integrate these findings to investigate how cysteine could have facilitated the emergence of life which is explored in the subsequent chapter.

Chapter 4: Investigating the role of cysteine in stereoselective surface-catalysed peptide synthesis.

Introduction

Previous chapters have demonstrated the potential importance of cysteine in abiotic protein function, and I have suggested that the abiotic availability of cysteine was indispensable for early protein activity. However, the instability of cysteine's thiol group in oxidising laboratory environments, where it is prone to oxidation and degradation, has historically led to its exclusion from abiotic peptide synthesis studies (Mayer-Bacon *et al.*, 2022). This exclusion leaves a gap in our understanding of whether cysteine can engage in peptide synthesis under realistic prebiotic conditions, such as those occurring in hydrothermal vent environments. While Foden *et al*. demonstrated that the cysteine thiol has the capacity to catalyse ligation of other peptides in neutral water, it has not been tested under conditions representative of hydrothermal systems or on mineral/solvent interfaces (Foden *et al.*, 2020).

Study rationale

To address whether cysteine enhances abiotic peptide synthesis, I investigated its inclusion under conditions analogous to white smoker hydrothermal vents, which are alkaline, moderately saline, and reducing environments. This study aims to determine if cysteine can overcome the limitations of low peptide yields and limited stereoselectivity observed in previous experiments. The binding of other amino acids to mineral surfaces is via carboxylates, which is likely far weaker than thiol-based interactions.

Abiotic peptide synthesis and limitations

As discussed in Chapter 1, mineral surfaces can catalyse peptide synthesis by adsorbing amino acids, increasing local monomer concentrations,

and enabling bond formation (**Fig 1.03)** (Bernal, 1949)**.** The efficiency of this process depends on the residence time of amino acids on the mineral surface, which is influenced by the amino acid/mineral interaction strength and the electrostatic environment (Lambert, 2008)(Kitadai *et al.*, 2017). Peptide synthesis in alkaline hydrothermal models using the abiotic amino acid set typically yields low peptide bond formation rates and produces short peptides of about 2-6 amino acids (Cleaves *et al.*, 2009)(Kitadai and Maruyama, 2018)(Lemke *et al.*, 2009)(Robinson *et al.*, 2021)(Otake *et al.*, 2011). This limitation arises from the weak association between peptides and mineral surfaces which decreases as a function of peptide length due to the peptide's decreasing molecular charge density and the formation of chain-terminating products. Additionally, stereoselective peptide synthesis, whereby L-amino acids are incorporated preferentially into the growing peptide, a factor that may have contributed to the homochirality of L-amino acids in modern proteins, is generally absent (Chen and Ma, 2020). Previous demonstrations of stereoselective syntheses rely on chiral templating minerals which may not have been prevalent in prebiotic environments (Hazen *et al.*, 2001). Therefore, no general peptide synthesis mechanism currently provides both high yield and significant stereoselectivity under plausible prebiotic conditions (Kitadai and Maruyama, 2018).

Cysteine's potential in abiotic peptide synthesis

The inclusion of cysteine in abiotic peptide synthesis models offers a potential solution to these limitations. Its thiol group is significantly more nucleophilic than other amino acid side chains, which offers new routes to enhance interactions with mineral surfaces (Brotzel and Mayr, 2007). This

increased reactivity and mineral residence time could lead to higher peptide yields and the synthesis of longer peptides. Furthermore, the ability of cysteine to coordinate transition metals was likely a key enabler for early metabolic functions (Vallee *et al.*, 2017)(Lukács *et al.*, 2021).

Thiol-peptide selective peptide ligation

In this chapter, I also investigated a proposed novel pathway for peptide elongation called Thiol-Peptide Selective Peptide Ligation (TSPL). Here, thiol-bearing peptides spontaneously coordinate transition metals, forming metal-thiol complexes under alkaline conditions. Within these complexes, peptides are concentrated around metal ions, enabling the ligation of thiol bearing peptides via their N- and C- terminal amino acids (**Fig 4.01**). This would allow the selective elongation of only thiol-peptides with amino acid sequences that have the correct spatial separation of cysteine to allowing spontaneous coordination, (i.e. CXXC, or more rarely CXC or CXXXC, where X is any other amino acid) without the need for genetic coding mechanisms (Lukács *et al.*, 2021). Thus, this pathway is partially sequence selective. This reaction could plausibly proceed both within the bulk fluid, or where one of the thiol-peptides is covalently bonded to the mineral surface, where higher amino acid and peptide concentrations likely further enhance the reaction rate.

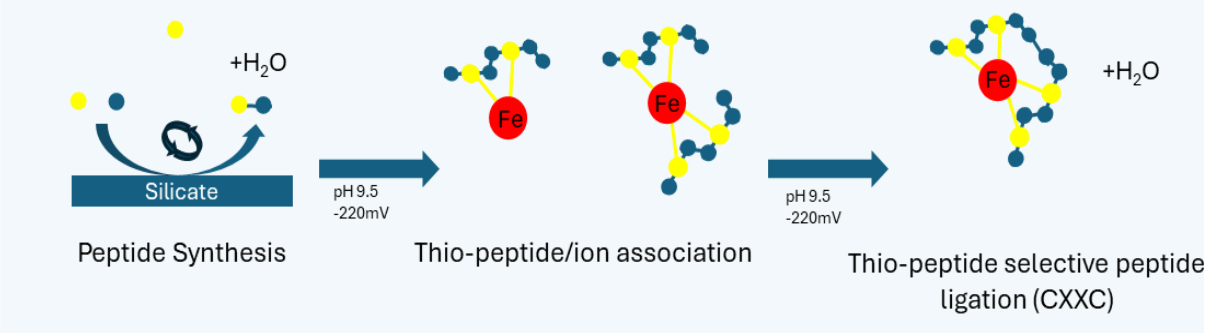


Figure 4.01 – Proposed mechanism of thiol-peptide selective peptide ligation

A core requirement for this pathway, however, would be the abiotic generation of cysteine enriched peptides with sufficiently high yields that the probability of a randomly sequenced peptide containing two cysteine amino acids with the sufficient separation in the amino acid sequence (CXXC) becomes reasonably high (i.e. >0.01 probability per 100mer long protein). As observed in the previous chapters, proteins containing aspartate or glutamate as the only anionic amino acid are insufficiently nucleophilic for the spontaneous coordination of transition metals in solution (Dokmanić *et al.*, 2008).

TSPL would thus introduce a fundamentally new route for elongating peptides, bypassing the reliance on wet-dry cycles typically required for abiotic synthesis of longer peptides (Rodriguez-Garcia *et al.*, 2015). TSPL thus provides a compelling, abiotically plausible pathway for overcoming the limitations of short peptide synthesis, paving the way for the emergence of early functional proteins and catalytic system.

Approach

To test the two hypotheses (cysteine acceleration of peptide synthesis and TSPL), I designed experiments simulating alkaline hydrothermal vent conditions. I conducted surface-catalysed peptide synthesis reactions with, and without, cysteine, using basaltic silicate minerals as catalytic surfaces under reducing conditions at elevated temperatures. Peptide yields were initially assessed using UV spectrophotometry, and formation was confirmed through mass spectrometry and ^{1}H-^{1}H and ^{1}H-^{13}C NMR. I also examined stereoselectivity by comparing reactions involving L-cysteine and D-cysteine. The mechanisms of the observed effects were elucidated by solid state NMR and polarisation resolved IRRAS.

These experiments were built upon with a test of the ligation of an 7mer peptide identified in the synthesis experiments to create a 14mer peptide with and without metal ions present to investigate if thiol-bearing peptides can enable the novel TSPL pathway.

Methods

Peptide synthesis

Scourie Mòr (N.W Highlands, Scotland) pyroxenite was crushed and mesh filtered to less than 106µm. 50g of the resulting powder was cleaned by washing and drying at 120°C for 2 hours with 2x 50ml 60% ethanol (diluted from 100% Stock; VWR) and 2x 50ml MilliQ H_2O across four heat-drying cycles. After cleaning, the pyroxenite was heated for a further 24 hours at 200°C. 0.06g/ml (of reactant) of cleaned sterilised Scourie Mòr pyroxenite was added into to 100ml borosilicate crimp neck vials (VWR, 548-0609 (50ml experiments)) or 15ml polypropylene screw cap tubes (SARSTEDT, 62.554.502 (5ml experiments)) to simulate a mineral/ocean interface near a hydrothermal system. For experiments using different oxide minerals, spectroscopy pure SiO_2, AlO_2, MgO_2 and FeO_2 was added to each reaction vessel as supplied (Sigma-Aldrich).

20mM total L- or L+D (1:1) meteoric amino acid mixture (MAAM) solutions were prepared from crystalline powders as supplied (>98%, Thermo Fisher Scientific) in the following proportions in line with the amino acid proportions reported in **Fig 1.01**: glycine (0.41), alanine (0.24), aspartate (0.12), glutamate (0.10), valine (0.04), serine (0.02), isoleucine (0.02), leucine (0.01), proline (0.03), and threonine (0.01). For experiments using isotopically labelled amino acids, crystalline L-cysteine (99% 3-^{13}C, Cambridge Isotope Laboratories), glycine ^{15}N (98% ^{15}N, Sigma-Aldrich), L-methionine (^{13}C-methyl) (Sigma-Aldrich) and L-alanine ^{15}N (98% ^{15}N, Cambridge Isotope Laboratories) were used. The amino acid mixtures were dissolved in 0.1M bicarbonate/carbonate buffer (pH 9.55±0.1) that had been prepared from anhydrous sodium carbonate (Fisher Scientific) and

sodium bicarbonate (ACS grade, Sigma). Once all amino acids were fully dissolved, 500mM NH_4OH was added from a 28-30% NH_4OH stock solution (ACS grade, 28-30% in water, Thermo Fisher Scientific) and the pH adjusted with 33% HCl or 5M NaOH to 9.55±0.1.

20mM stocks of each 'excess' amino acid to be tested (>98% purity, Thermo Fisher Scientific) were prepared separately in 0.1M bicarbonate/carbonate buffer (pH 9.55±0.1). Solutions contained 20mM Na_2S (ACS grade, MP Biomedical) or 20mM TCEP (Fluorochem) prior to mixing to prevent amino acid oxidation before or after peptide synthesis. The measured redox potential was typically 90-100mV following synthesis. All solutions were sparged for 20 minutes in N_2 followed by a further 30-minute equilibration in a 100% N_2 atmosphere (<10 O_2 ppm) in a positive pressure anaerobic hood (COY Laboratory Products).

Following N_2 equilibration, the MAAM solutions and the excess amino acid solution were added to the borosilicate or polypropylene reaction vessel with 0.06g/ml pyroxene or other oxide mineral. The solutions were diluted with 0.1M carbonate buffer (pH 9.55±0.1) to a final concentration of 5mM MAAM, 5mM TCEP or Na_2S, 250mM NH_4OH, and 5mM of the excess amino acid. The reaction vessels were crimp sealed, or screw capped and heated at 90°C (Heratherm OGS100).

At reaction completion, the reaction vessels were removed from the oven and uncapped within a fume hood while cooling to room temperature. Solutions were vacuum pump filtered through a 0.22µm polyethersulfone (PES) membrane

(MilliPore). The resulting amino acid/peptide mixtures were stored at -80°C until analysed by UV-spectrophotometry or lyophilised in 20 or 1ml aliquots.

Peptide purification and analysis

C18 peptide purification

Post-reaction samples for TIMS-TOF MS were purified by C18 chromatography prior to loading. 4ml aliquots of post-reaction sample were lyophilised and resuspended into 400µl of 0.1% trifluoracetic acid (TFA) (Sigma-Aldrich). Separately, C18 SPE tips were prepared by packing 2µl of C18 resin (Empore SPE C18 48mm 12µm particle size) into 10µl micropipette tips. The resin was activated by washing twice with 60µl of 100% acetonitrile (HPLC grade, 99.9%, Sigma-Aldrich) and 60µl of 0.1% TFA (Sigma-Aldrich) with a 2-minute spin at 4000xg after each addition. 400µl of sample in 80µl aliquots was then pipetted into each SPE tip with a 4-minute spin at 5000xg after each addition. The flow-through was retained, and the column washed with 60µl of 0.1% TFA. Peptides were eluted in 160µl of 60% acetonitrile in 80µl aliquots. Elution fractions were lyophilised and stored at 4°C prior to analysis.

UV-Vis spectrophotometry

Post-reaction samples in 0.1M carbonate at pH 9.5 were diluted 1 in 20 with 0.1M carbonate buffer at pH 9.5 (prepared as described above) to 200µl. 200µl of the diluted aliquots were added to an HellmaAnalytics High Precision Quartz SUPRASIL UV-Vis cell (light path 10x2mm) and were scanned at wavelengths of 200-800nm in a Cary 50 Bio UV-Visible Spectrophotometer with a scan rate of 4800.000nm/min and 1.0nm data intervals in dual beam mode with a baseline correction against a 0.1M carbonate pH 9.5 solution. Three technical

repeats were recorded for each sample and the mean reported as a single biological repeat for comparison between synthesis runs. The UV-Vis cell was cleaned between samples with 1ml of 2% Hellmanex III (Sigma Aldrich) and thoroughly rinsed with Milli-Q water between each experiment and dried with N_2 gas. Biological repeats were normalised relative to L-alanine. Results were exported with the absorbance at 230.04nm plotted for each experiment using GraphPad Prism Software. For measurement of rates of apparent peptide bond formation, peptide bond concentrations were estimated using 10mM dialanine (A_{230} = 0.45) as an internal calibration standard.

Statistical analysis of UV-Vis results

The statistical significance of results acquired by UV-Vis was determined by one way ANOVA with Tukey Post Hoc analysis of the mean absorbance values (2-3 technical repeats per sample) of post-reactive solutions, following heating as reported for **Fig 4.03a**. An unpaired two-tailed T test was used instead for data reported in **Fig 4.03b**.

TIMS-TOF MS

Note: The following samples were resuspended and loaded on to the instrument by Dr Rod Chalk within the Mass Spectroscopy Analysis Facility (Centre State for Medicines Discovery, University of Oxford). Spectral analysis was conducted by the author.

TIMS-TOF MS was conducted on purified peptide solutions following C18 SPE purification (described above). 1µl of purified peptides in 0.1% FA was injected onto a Bruker nanoElute HPLC equipped with a ThermoTrap Cartridge

Guard Column (Thermo Scientific 160454) and a Bruker PepSep Fifteen elution column operating at 50°C (Bruker, 1893473). Peptides were eluted on to a Bruker timsTOF Pro 2 Mass spectrometer by gradient elution at a flow rate of 0.55ul/min (Buffer A - 0.1% FA, Buffer B - 100%ACN 0.1% FA) on a 30 minute 0-40% ACN gradient). Detections were analysed within the Bruker Direct Analysis software. Peak lists were analysed using a novel Python approach (**Appendix C.3**). This consisted of **massmatcher.py**, which compared exported mass detections to a list of every possible amino acid combination of peptides up to 20 amino acids long (**Peptidesequencepredictor.py**, **peptidemasspredictor.py**), and candidate peptides with >5ppm deviation in mass were discarded. For detections where multiple sequence combinations were possible, the identification with the lowest ppm error, while still no greater than the 5ppm cutoff criteria, was retained. Detections identified also in the negative control were removed. Detections with elution times less than 18 minutes (≈20% ACN) or greater than 28.5 minutes (≈32% ACN), and those with signal-to-noise ratios <2, were also removed.

Solution state NMR

Lyophilised crude or C18 SPE purified post-reaction solutions (1ml original volume) were resuspended into 200µl of water with 5% D_2O and 0.4mM DSS (Cambridge Isotope Laboratories) at pH 5.5 or in 99.9% D2O (Sigma-Aldrich) with 5mM benzoic acid (ACS grade 99.5%, Thermo Fisher Scientific) and 0.4mM DSS (Cambridge Isotope Laboratories) at pH 5.5. The resulting 5x concentration solutions were loaded into 3mm Bruker SampleJet NMR tubes. NMR data were recorded in magnets operating at 17.6 or 22.3 Tesla (Oxford Instruments magnets) and equipped with high sensitivity 5mm TCI cryoprobes and Avance III

HD consoles (Bruker). 2D $^{1}H,^{13}C$ heteronuclear single quantum correlation (HSQC) spectra were acquired with 64 scans, and the total points and sweep widths were 1274 and 320, and 11.934ppm and 65.00ppm, for ^{1}H and ^{13}C, respectively. 2D total correlation spectroscopy (TOCSY) data were acquired with 128 scans, sweep widths of 10.000ppm, and total points of 2048 and 320 in the direct and indirect ^{1}H dimensions, respectively. qNMR to measure depletion of monomeric amino acids was performed by integrating and comparing the $^{1}H\alpha$ signals from monomeric amino acids before and after reaction. The integrated triplet signal of benzoic acid centred at 7.403ppm at 20˚C was used as an internal standard to enable quantitative comparison between experimental runs. The results were analysed with TopSpin 4.3.0.

Confirmation of amino acid stock ratios

Confirmation of both monomer concentration parity and purity of the relevant stocks between samples containing different enantiomers of excess amino acids was performed by measurement of peak height of the ^{1}H-Cα resonance by solution NMR. Concentration differences in L- vs D-cysteine stocks ranged from 0.999-1.001 between samples.

Probing cysteine silicate surface interactions

Preparation of SiO_2/amino acid complexes

SiO_2/amino acid complexes were prepared using the approach of (Lopes *et al.*, 2009). Solid amino acid/SiO_2 samples were prepared in a 100% N_2 gas atmosphere (<10ppm O_2) to eliminate the requirement for reducing agents during the mixing stage. 500mM solutions of L- or D-cysteine were prepared by dissolution of crystalline cysteine (>99%, Thermo Fisher Scientific) in Milli-Q

water, the pH was adjusted to 1.5 or 10.5 using 33% HCl or 5M NaOH. For each reaction, 3.0g of fumed SiO_2 (Specpure, Alfa Aesar) was weighed out and placed into a 50ml polypropylene tube (Sarstedt). Amino acid solutions were added to the SiO_2 powder and the tubes screwed-capped. The samples were stirred with a magnetic stir bar for 48 hours at 23°C.

Following mixing, the solutions were vacuum pump filtered through a 0.22µm PES membrane (MilliPore) to remove free amino acids. The SiO_2 solid left on the disk was gently washed and filtered three times with 30ml of 0.1M carbonate at pH 10.5 or pH 1.5. Following washing, the SiO_2 solid was scraped from the membrane and put into a 50ml polypropylene tube and lyophilised prior to storage at -80°C before analysis.

^{13}C Solid state NMR

Note: The following NMR spectra were acquired by Dr Nick Rees at the solid-state NMR facility (Department of Chemistry, University of Oxford). Spectral analysis was conducted by the author.

Solid SiO_2/cysteine complexes, and control powders of free cysteine (99%, Thermo Fisher Scientific) and cystine (99%, Thermo Fisher Scientific) were loaded into a 100µl rotor. Solid state NMR data were recorded at 9.4 Tesla magnet equipped with an Avance III console (Bruker). ^{13}C cross polarisation experiments were recorded with 4096 points and a sweep width of 405 ppm. Data on post-reaction samples were recorded with 80,000-100,000 scans, whereas data for the L-cysteine and L-cystine controls were recorded with 64 scans. NMR data was analysed with TopSpin 4.3.0.

PR-IRRAS sample production and data collection

Note: The following Ti/Au/SiO_2 microfilm slides were prepared by Dr Richard Makin at the Thin Films Facility (Department of Physics, University of Oxford). PR-IRRAS spectra were collected by Dr Robert Jacobs at the Surface Analysis Facility (Department of Chemistry, University of Oxford). Spectral analysis was conducted by the author.

Ti/Au/SiO_2 microfilm slides mounted on a standard glass slide were prepared at the Thin Film Facility in the Department of Physics, University of Oxford. Glass microscope slides were thoroughly cleaned prior to coating with Decon 90 and rinsed with deionised water, acetone, and IPA. Slides were loaded into a Leybold L560 evaporation coater and cleaned by Argon glow discharge for 10 minutes (with an Ar pressure of $6x10^{-2}$ mbar). Ti was subsequently evaporated using an electron beam at high vacuum ($4x10^{-5}$ mbar) and deposited at a rate of 0.3nm/s to a final thickness of 2nm. Au was evaporated by thermal evaporation at ($2x10^{-5}$ mbar) and deposited at a rate of 0.5nm/s to 10nm thickness. Finally, SiO_2 was evaporated using an ebeam ($2.2x10^{-5}$ mbar) and deposited at a rate of 0.25nm/s to a thickness of 10nm. Following preparation, the slides were moved to an anaerobic hood and placed at the base of a polypropylene container.

Separately, 25mM solutions of D- and L-aspartate and solutions of 50mM D- and L-cysteine and D-and L-alanine were made up in Milli-Q water under a 100% N_2 atmosphere. Solution pH was adjusted to 10.0 using 5M NaOH and each added to a container with a Ti/Au/SiO_2 slide. The set up was incubated for 36 hours at room temperature. Following incubation, the slides were removed and washed using 20ml of pH 10.00 degassed Milli-Q water before being dried

at room temperature prior to analysis. PR-IRRAS spectra for the Ti/Au/SiO_2 amino acid complexes were recorded on a Bruker Vertex 80 FT-IR spectrometer equipped with a liquid N_2 cooled MCT detector and a Pike VeeMax accessory. The spectrometer was referenced to a blank Ti/Au/SiO_2 slide, and the samples were assessed with s-polarised and p polarised light at incidence angles of 30˚, 60˚ and 80˚. Absolute reflectance of p polarised beams, which only reflected from surface bound species, was plotted using GraphPad Prism software, compared against the s polarised light which provided the background reflectance from the material itself.

Monte Carlo simulations

Monte Carlo simulations were conducted to model the impact of increased cysteine reactivity on peptide synthesis and the probability of CXXC motif formation within random systems. The code used to conduct this is available in **Appendix C.4**. Simulations were performed using a Python script to generate randomised peptide sequences based on the determined abiotic amino acid probabilities and adjusted reactivity factors based on the observed enhancement of yield (7.8x) when cysteine was present (Higgs and Pudritz, 2009). Base amino acid probabilities (mole fractions) were assigned as follows: glycine (0.41), alanine (0.24), aspartate (0.12), glutamate (0.10), valine (0.04), serine (0.02), isoleucine (0.02), leucine (0.01), proline (0.03), and threonine (0.01). Two values for cysteine concentration were initially used: a lower bound value of 0.01, and an upper bound value of 0.035. To determine the cysteine concentration required for there to be a 5% probability of every peptide having a CXXC motif, simulations were conducted where the cysteine probabilities were systematically varied

across ten experiment runs at equally spaced intervals between 0.001 and 0.1 mole fraction, with corresponding adjustments to the other amino acid probabilities to maintain a total probability of 1.

Reactivity factors were applied to adjust the likelihood of amino acid inclusion in peptides based on their reactivity. Cysteine was assigned a reactivity factor of 7.8, as derived from its experimentally observed enhanced reactivity as observed in **Fig 4.02**. Aspartate was assigned a reactivity factor of 1.68, based on the percentage enhancement of A_{230} in peptide synthesis experiments containing excess aspartate compared to alanine. All other amino acids were assigned a baseline reactivity factor of one. Two simulation conditions were tested for each cysteine concentration: one assuming uniform reactivity factors of 1 for all amino acids, and another incorporating the specified reactivity factors.

Each simulation generated five million peptides of 10 residues in length per run, with 50 runs conducted per cysteine concentration, yielding 250 million peptide sequences per condition. Peptide sequences were generated using a NumPy script (version 1.21.2) by random sampling of amino acids according to the adjusted probabilities. The percentage frequency of CXXC motifs and peptides containing at least one cysteine was tabulated using Python's Counter module, and the results were normalised to the total number of peptides generated and plotted as a function of cysteine concentration.

Peptide Ligation

A cysteine bearing peptide (CGSCST) identified by timsTOF was modified with one additional serine, to make the peptide slightly longer which I assessed

as more favourable to peptide ligation by increasing spatial separation between the metal binding site and the N terminus for rotational freedom, to create the CGSCSST (Arc4) peptide. This was synthesised and used as provided (GenScript Peptide Synthesis Services). This additional amino acid was added based on a prediction that a slightly longer peptide than the 6mer would give the n terminus more flexibility to react with other soluble peptides when bound to a metal ion. An additional peptide identified, with a composition matching one of those detected in the peptide synthesis reaction, Arc3 (GGCPSCGGC) was also synthesised for O_2 stability testing. All preparation steps were conducted within an anaerobic hood at a measured O_2 <10ppm, and the solutions were sparged in 5% H_2 and 95% N_2 for 8 minutes at approximately 5 container volumes/min immediately prior to heating. Failure to sparge with H_2 results in the oxidation of cysteine peptides and ligative failure.

Lyophilised peptides (4.9 mg) were resuspended in 1 ml of pH 9.5 bicarbonate/carbonate buffer. Separately, borosilicate reaction vessels were made up to 2.8 ml with bicarbonate/carbonate buffer 5mM TCEP (pH 9.5). 200µl of 7.61 mM peptide solution were added to achieve a final concentration of 0.5 mM per reaction. Twenty microliters of 1M $FeCl_3$ or $ZnCl_2$ were added to a final concentration of 6.7 mM. The reaction vessels were sealed and heated for 36 hours. Resulting solutions were filtered through 0.22 µm filter membranes to remove metal hydroxide precipitates. The solutions were lyophilised and resuspended in 200 µl of Milli-Q water and analysed by ^{1}H NMR and ^{1}H-^{13}C solution NMR as described above, or by proteomics liquid chromatography-mass spectrometry (LC-MS).

Peptide metal binding

Metal binding capabilities of the Arc4 peptide prior to ligation was investigated by UV-Vis as described above using 0.1mM Arc4 peptide in 0.08M bicarbonate buffer to which $FeCl_3$ and $CoCl_2$ was added to a final concentration of 0.13mM both with and without the peptide present to enable difference spectrum to be calculated. The absorbance of the buffer + metal ion, and buffer + Arc4 was subtracted from the buffer + Arc4 + metal ion spectra and difference spectrum were acquired. Data smoothing was performed using a Savitsky-Golay filter and deconvoluted using the deconvolution function contained within the *SciPy signal* library.

Hybrid Quadrupole-Orbitrap LC-MS

Lyophilised 7-mer and 14-mer peptides were resuspended in 0.1% formic acid. Nano-liquid chromatography was performed using an Ultimate 3000 RSLC system (Thermo Fisher Scientific) coupled to a Q Exactive mass spectrometer with an EASY-Spray source. Peptides were trapped on a C18 PepMap100 precolumn (300µm × 5mm, 100Å) and separated on an EASY-Spray RSLC C18 analytical column (75µm × 50cm, 2 µm particle size) using a linear gradient from 15% to 35% solvent B (0.1% FA in 100% ACN) over 15 minutes at a flow rate of 200 nl/min following a 20min load phase.

The mass spectrometer operated in data-dependent acquisition mode. Full MS spectra were acquired in the Orbitrap (m/z 350-1,500) at a resolution of 70,000, with an automatic gain control (AGC) target of 3×10^6 and a maximum injection time of 50ms. The top five most intense ions were selected for higher-energy collisional dissociation (HCD) fragmentation at a normalised collision

energy of 30%. MS/MS spectra were acquired at a resolution of 17,500, with an AGC target of 5×10^4 and a maximum injection time of 120 ms. Dynamic exclusion was set to 5s. Data analysis was performed using the Mzmine software.

Results

Cysteine increases peptide synthesis reaction rates.

A reaction scheme was established based on the oxide mineral and amino acid reaction systems of (Takahagi *et al.*, 2019). All reaction mixtures contained a racemic or L-meteoric amino acid mixture (MAAM) supplemented with an additional L- or D-amino acid at excess concentration to test its effect on peptide yields. The ratios of amino acids used were based on the mean rank of relative concentrations of amino acids observed in nonbiological contexts reported by (Higgs and Pudritz, 2009). The mineral surface was represented with heat-sterilised pyroxene or fumed spectroscopy grade SiO_2 in borosilicate or polypropylene reaction vessels, and reaction mixtures were heated in reducing conditions at 90°C for the times indicated. Consistent with surface-catalysis the reactions in borosilicate vessels led to 1.8-fold increases in peptide yields when compared to those caried out in polypropylene vessels due to the additional silicate (**Fig 4.02**). The starting stocks of monomers were checked for both purity and parity using solution NMR. All samples were confirmed to be 99-101% ratiometric L vs D mixtures prior to heating, confirming that differences in stock purity did not result in the differing peptide yields observed.

Initial assessment of peptide yield was conducted by UV-spectrophotometry. Statistical significance of differences in absorption was determined through one way ANOVA with Tukey post hoc analysis to determine the effect of each tested excess amino acid on peptide yield compared to an L-alanine negative control (**Appendix B - Table 02**). In borosilicate reaction

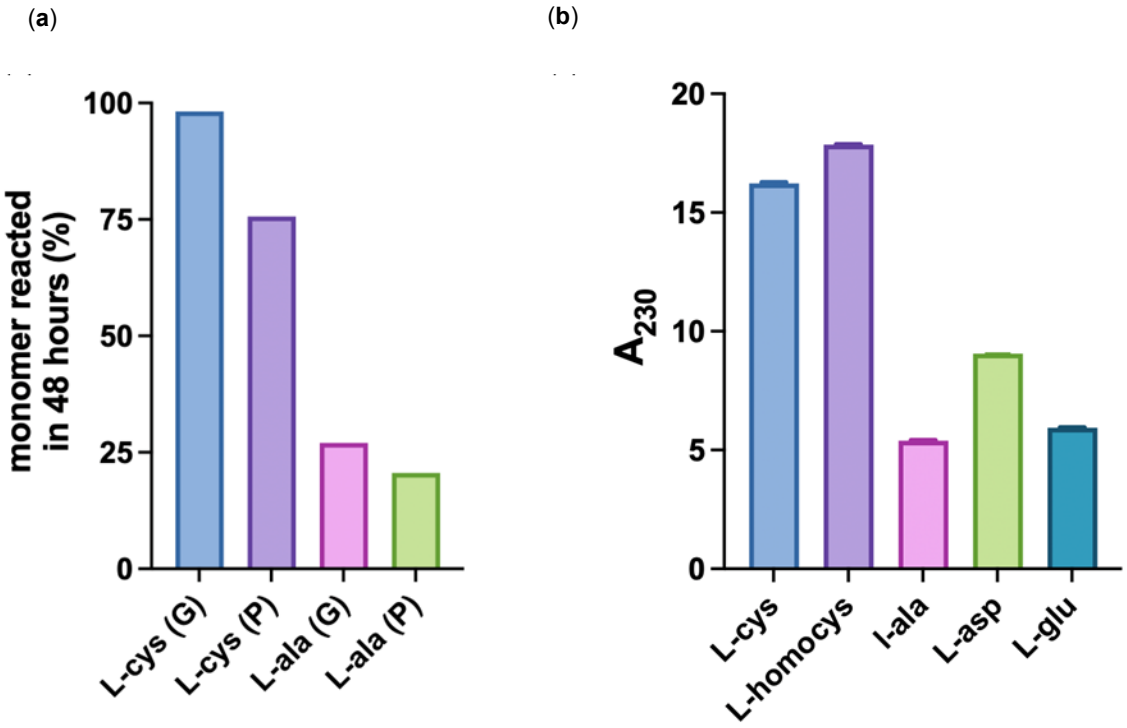


Figure 4.02 – The presence of a free thiol side chain significantly increases peptide yield in the presence of silicate.

*(**a**) The amount of monomeric amino acid removed from solution after 48 hours in the peptide synthesis reaction containing an excess of L-alanine or L-cysteine in borosilicate glass (G) or polypropylene (P) reaction vessels. The amino acids were measured by Quantitative NMR using a benzoic acid internal standard. Error bars display standard deviation of the benzoic acid signal intensity at 7.4034ppm. (**b**) Peptide abundance as determined by A_{230} for reactions carried out in borosilicate reaction vessels and pyroxenite with a 5mM excess of the indicated amino acids. Error bars display the standard deviation of instrument noise at 400-800nm.*

vessels, absorption of 230nm light (A_{230}) increased significantly after 48 hours in the presence of L-cysteine compared to reactions with either L-alanine ($p=<0.0001$) or D-cysteine ($p=0.0006$). In contrast, synthesis with excess L- or D-alanine ($p=0.7172$), and L-alanine or L-methionine ($p=0.7021$) did not result in significantly increased peptide yields, indicating that these enantiomers absorbed to silicates similarly at alkaline pH. When compared to the mean absorbances for L-alanine, D-alanine and L-methionine, the presence of L-cysteine increased peptide bond yield 7.8-fold (±0.6), and D-cysteine increased it 3.9-fold (±0.6) (**Fig 4.03a)**.

The presence of silicate surfaces significantly amplifies the catalytic effect of cysteine on peptide bond formation. In the absence of a silicate surface, peptide bond yields were increased approximately 3-fold, consistent with thiol group catalysed ligation reported by Foden *et al*, and the addition of a silicate surface (pyroxene) increased this yield a further 6-fold ($p=0.0015$, two-tailed unpaired t-test) (**Fig 4.03b, Appendix B - Table 03**) (Foden *et al.*, 2020).

Given the large number of organic species that could have been generated in our model system with absorbance at A_{230}, the synthesis of peptides was further investigated by ^{1}H-^{1}H and ^{1}H-^{13}C solution NMR. Quantitative NMR (qNMR) measurements after 48-hour reactions with either L-cysteine or L-alanine indicated that there was a 7.7-fold increase in monomeric L-cysteine removed than L-alanine (**Fig 4.03c**). By tracking the resonance specific intensity both prior- and post peptide synthesis of the self-same samples, any difference in the initial amino acid concentration or contaminants can be normalised allowing direct comparison of run-products. ^{1}H-^{1}H total correlation spectroscopy (TOCSY)

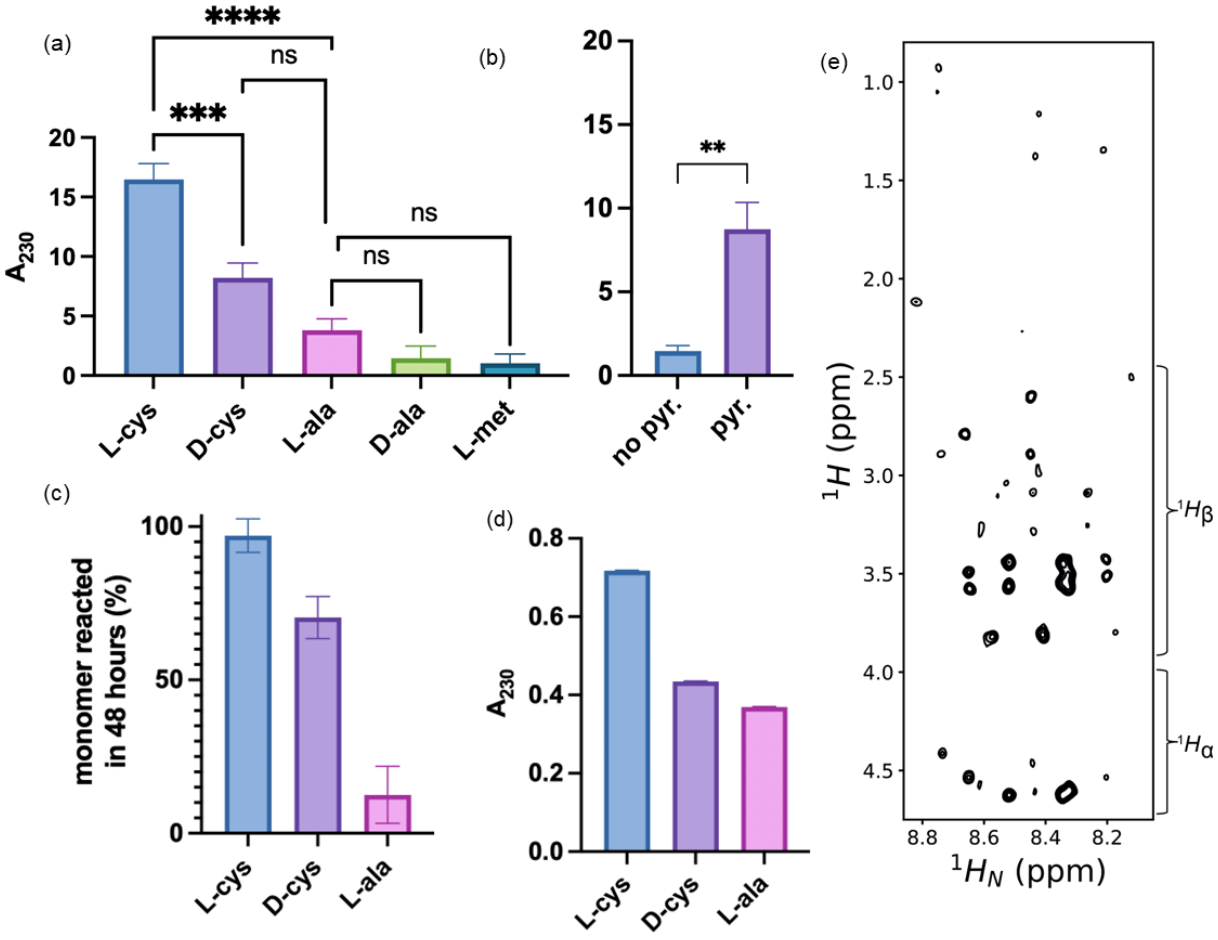


Figure 4.03 – Cysteine stereo-specifically increases peptide yield in the presence of an achiral oxide mineral.

*(**a**) Detection of peptide bond formation by absorbance at 230nm (A_{230}) after 48 hours of heating in borosilicate reaction vessels with MAAM plus an excess of L-cysteine (n=6), D-cysteine (n=5), L-alanine (n=6), D-alanine (n=3), or L-methionine (n=2). Biological repeats were normalised to the A_{230} with L-alanine. Normalisation relative to L-cysteine was used to determine the standard error of the mean (SEM; error bars) of L-alanine biological repeats. The mean control absorbance of L-alanine, D-alanine and L-methionine was 2.10. Significance was determined by one way ANOVA with Tukey post-hoc analysis of absorbance vs L-alanine (L-cysteine, p=<0.0001; D-cysteine, p=0.085; D-alanine, p=0.712; L-methionine, p=0.702) (**Appendix B - Table 02**) (**b**) A_{230} amino acid mixtures reacted in polypropylene reaction vessels with L-cysteine and without or with pyroxene (pyr) (p=0.0015, two-tailed unpaired t-test (**Appendix B - Table 03**)). Error bars indicate SEM. (**c**) Loss of monomeric amino acid reactant by Quantitative NMR with a benzoic acid internal standard after 48 hours of reaction time. Error bars display standard deviation of the benzoic acid signal height at 7.4034ppm. (**d**) A_{230} measurement of stereoselective peptide bond formation following reactions at lower (100μM) amino acid starting concentrations. Reactions were heated at 90˚C for 42 days in borosilicate reaction vessels. Error bars indicate the standard deviation for technical repeats (n=3). (**e**) 2D ^{1}H-^{1}H TOCSY spectrum of a post-reaction mixture enriched in L-cysteine. The spectral region corresponds to intra-residue correlations between amide protons (^{1}HN) and either $^{1}H\alpha$ (~4.0 – 4.7ppm) or sidechain protons (~1.0 – 4.0ppm).*

of the experimental run products revealed new signals consistent with cysteine $^{1}H_{N}$-$^{1}H_{\alpha}$ and $^{1}H_{N}$-$^{1}H_{\beta}$ correlations that were not observed in the unreacted control samples (**Fig 4.03e**). These signals can only emerge following the creation of a peptide bond. As such these confirmed that the increases in A_{230} seen in L and D-cysteine were evidence of the formation of peptide bond and were consistent with results from UV-Vis.

Peptides synthesised with cysteine are compositionally diverse.

The synthesis products were further characterised by repeating the reactions using isotopically labelled amino acids for analysis by ^{13}C NMR spectroscopy. NMR spectra of products formed in the presence of $^{13}C_{\beta}$ L-cysteine after heating at 90°C for 48 hours indicated a large number of unique signals in the chemical shift range expected for non-oxidised cysteine $^{1}H_{\beta}$-$^{13}C_{\beta}$ and $^{1}H_{\alpha}$-$^{13}C_{\alpha}$ correlations, as well as amide ^{1}H to $^{1}H_{\alpha}$ and $^{1}H_{\beta}$, which were consistent with the formation of a diverse set of cysteine-bearing peptides (**Fig 4.04a-b**). Inspection of the pre-reaction solutions confirmed no major contaminants were present in the solution prior to synthesis. Automatic peak picking using TopSpin showed 18 ^{13}C resonances prior to synthesis and 265 ^{13}C resonances post synthesis, a 1472% increase in organic diversity. In contrast to the reactions with cysteine, few unique signals were observed when synthesis was conducted in the presence of ^{13}C-labelled L-methionine (**Fig 4.05**).

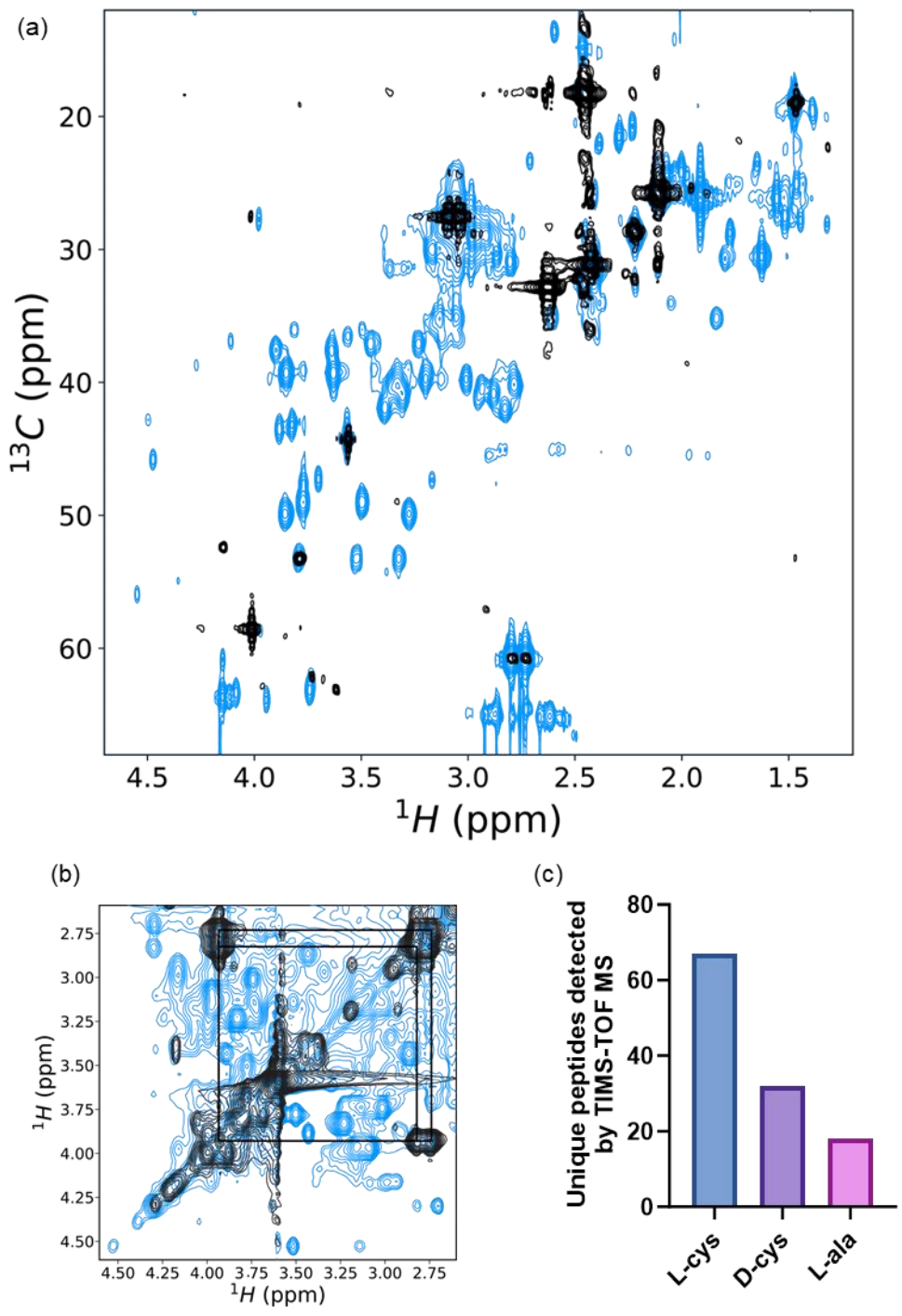


Figure 4.04 – Peptide synthesis with cysteine yields a diverse set of cysteine-containing peptides.

*(**a**) Overlay of 2D ^{1}H-^{13}C HSQC spectra of reactions containing $^{13}C_{\beta}$-labelled L-cysteine before (black) and after (blue) reaction. New crosspeaks in the post-reaction spectrum are attributed to incorporation of $^{13}C_{\beta}$-labelled L-cysteine into peptides adjacent to different amino acid types. Additional crosspeaks arise from natural abundance ^{13}C (1.1%). (**b**) Overlay of 2D ^{1}H-^{1}H TOCSY spectra of reaction mixtures containing L-cysteine before (black) and after (orange) reaction. The intense cysteine monomer $^{1}H_{\alpha}$-$^{1}H_{\beta}$ crosspeaks observed before reaction are indicated by lines. The large number of new post-reaction crosspeaks are attributed to unique peptide species. (**c**) Histogram of the total number of unique peptide products in L-cysteine reactions identified by TIMS-TOF MS after purification by SPE and HPLC prior to MS loading.*

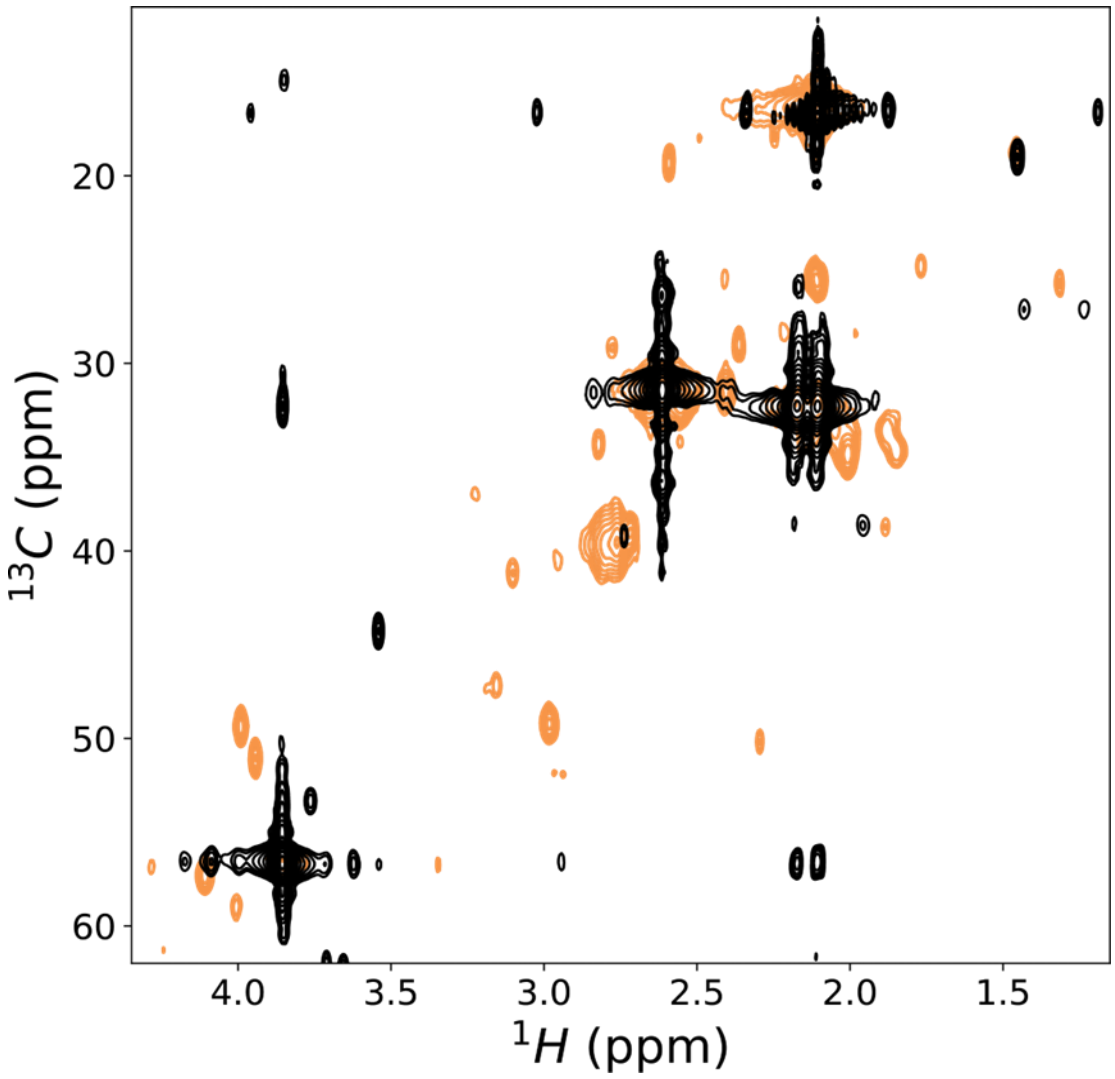


Figure 4.05 – Peptide synthesis in the presence of excess methionine produces a low diversity of products.

*^{1}H-^{13}C HSQC of reactions with excess ^{13}C-labelled methionine (orange).Resonances prior to heating are shown in black. Reactions were set up under identical conditions to those in **Fig 4.04** and carried out in borosilicate reaction vessels for 48 hours at 90˚C.*

Peptide compositions and lengths were evaluated using Trapped Ion Mobility Spectrometry Time-of-Flight (TIMS-TOF) mass spectrometry (MS) after 42 days at 90˚C followed by purification using C18 solid phase extraction (SPE). Synthesis conducted in the presence of L-cysteine generated the largest number of unique peptide detections (**Fig 4.04c**) (**Appendix B - Table 04**). The peptide diversities detected by TIMS-TOF MS were likely underestimated due to the C18 peptide purification conducted prior to HPLC/TIMS-TOF MS, which enriches samples in hydrophobic peptides (**Fig 4.06, Fig 4.07**). In the presence of L-cysteine, significantly larger numbers of unique peptides longer than four amino acids were detected compared to L-alanine (Mann Whitney Test, p=0.0188, **Appendix B - Table 05**), and peptides containing up to 12 amino acids were observed. In the presence of L-alanine the largest observed peptide was an 8-mer, and fewer unique peptides were observed overall. Increases in the number of detectable peptides were also observed with D-cysteine compared to L-alanine, but not to statistical significance (p=0.1563) (**Fig 4.08**, **Appendix B - Table 05**). The increased detections of peptides in reactions with L-cysteine compared with D-cysteine and L-alanine were consistent with the yields determined from A_{230} and qNMR and supported a stereoselective bias toward L-cysteine.

Increased peptide yields with cysteine are due to its thiol group.

The increased peptide yields with L- and D-cysteine compared to L-alanine correlate with the high nucleophilicity of the sidechain thiol, which results in strong attachment to mineral surfaces. Reactions in the presence of L-homocysteine showed a comparable level of peptide bond formation to L-cysteine (**Fig 4.02a**)

whereas L-methionine showed no statistically significant increase in peptide bond formation compared to L-alanine (**Fig 4.03b**). Reaction yields with excess L-aspartate were higher than the yields with excess L-alanine (**Fig 4.02b**), suggesting that attachment through the carboxylate sidechain can also contribute to increased yields. However, the yields with L-aspartate were much lower than that of reactions with excess L-cysteine, consistent with the much lower nucleophilicity and weaker silicate binding of the sidechain carboxylate group.

A high correlation between peptide yield in the presence of different metal oxides and the corresponding metal-sulfur bond dissociation energies was observed. This confirms the strength of attachment via the sulfur group to the mineral surface is the dominant factor affecting peptide yield (**Fig 4.09**). In contrast, silicate-sulfur bonds have very high dissociation energies (619kJ/mol (ΔHf_{298}))and the yield in the presence of silicate was lower than expected from the trend observed for metal oxides, which may indicate that adsorption was no longer rate-limiting.

NMR and MS results supported the observations by UV spectroscopy that cysteine stereoselectively enhances peptide yield when in the presence of an oxide mineral surface. The stereoselective bias toward L-cysteine persisted at the lowest concentrations of amino acids tested (55μM) (**Fig 4.03d**), indicating that peptides generated abiotically in the presence of racemic cysteine monomers would be enriched with L-cysteine at a ratio similar to that of the observed peptide yield increase.

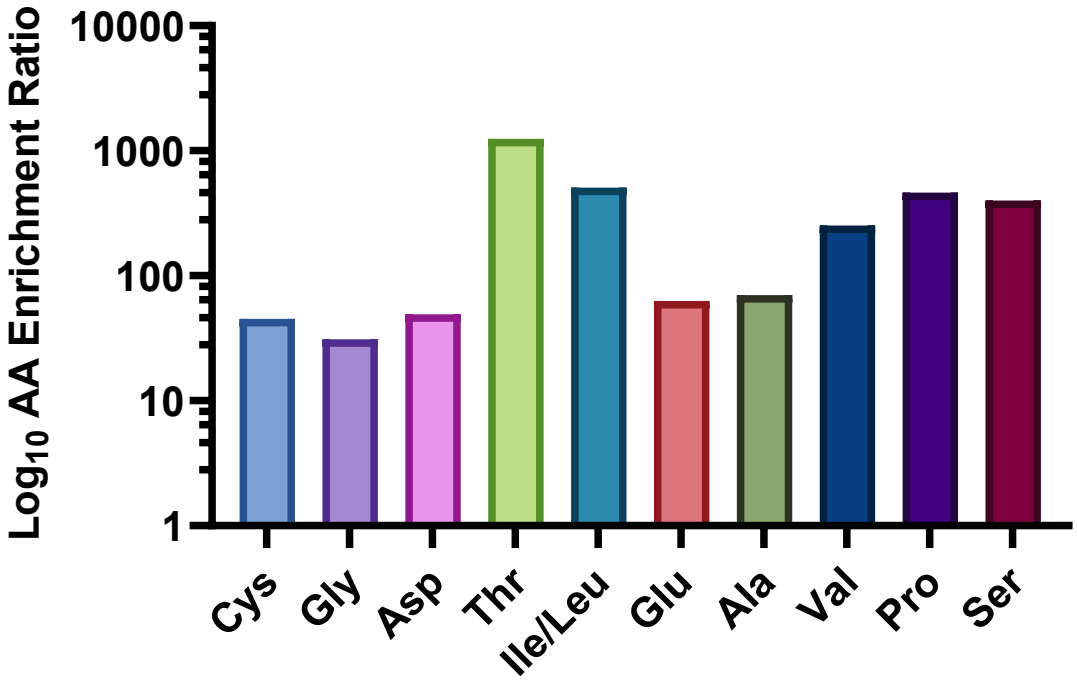


Figure 4.06 – Hydrophobic amino acids are enriched in peptides C18 SPE purified and detected by TIMS-TOF MS.

Peptide reaction mixtures were C18 SPE purified and analysed by TIMS-TOF MS. Amino acid enrichment was determined by dividing the observed amino acid abundance in the detected peptides by the relative abundance added to the reaction mixtures.

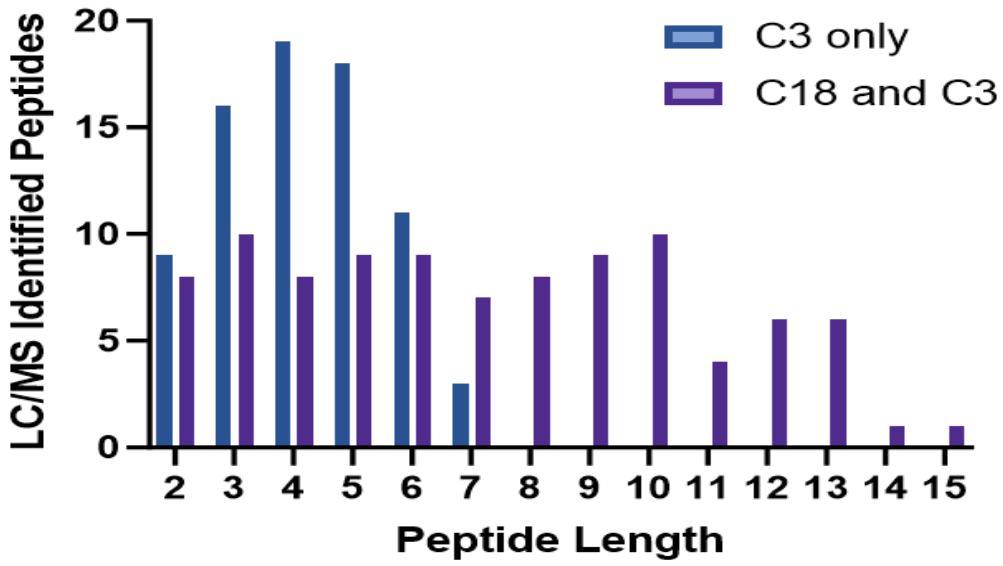


Figure 4.07 – C18 purification prior to injection increases the length of detected peptides by TOF MS

Peptide reaction mixtures were C18 SPE purified and or analysed by HPLC-TOF MS equipped with a C3 column. Identified proteins were tallied based on peptide length.

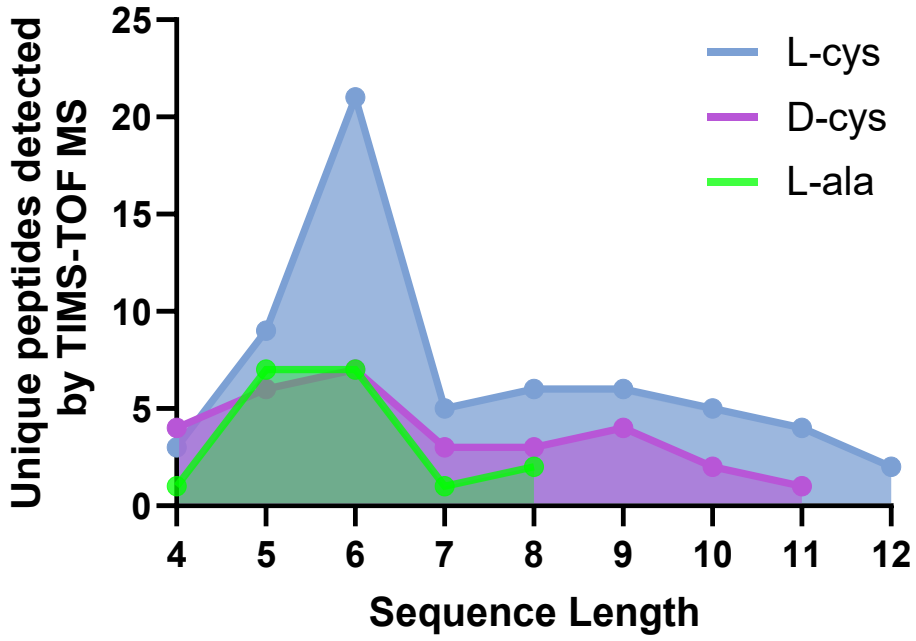


Figure 4.08 – Synthesis with L-cysteine increases peptide yield and the length of detected peptides.

The distribution of peptide lengths for reactions in the presence of excess L-cysteine or L-alanine for peptides detected by TIMS-TOF MS. Peptides were synthesised with excess L-cysteine, D-cysteine, and L-alanine, and then C18 SPE purified and analysed by TIMS-TOF MS.

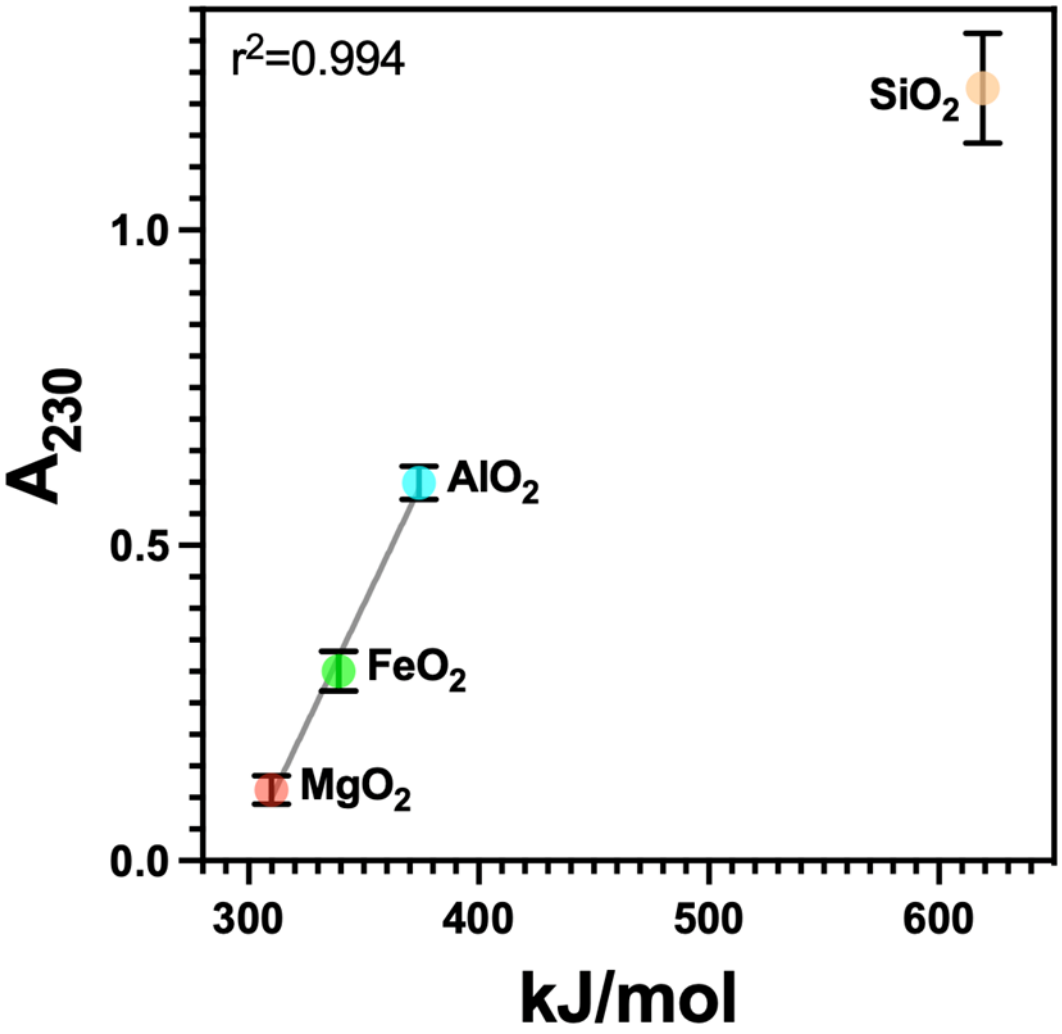


Figure 4.09 – Abiotic peptide bond formation in the presence of L-cysteine correlates with metal-sulfur bond dissociation enthalpies.

Peptides were synthesised as described in polypropylene reaction vessels with different silicate minerals present for 2 days. Peptide formation was assessed by A_{230}, and metal-sulfur bond enthalpies are taken from (Cottrell et al. 1958). Error bars display the standard deviation of instrument noise at 400-800nm. The linear regression fit, and correlation coefficient shown were calculated for the metal-containing silicates only.

Cysteine silicate bivalent adsorption at alkaline pH

I hypothesised that bivalent surface attachment to mineral surfaces, via both the carboxylate and sulfur groups and simultaneously with deprotonation of the amino group at alkaline pH, results in a mechanism for the observed stereoselective differences in peptide yields (**Fig 4.10**). Here, two-point attachment of cysteine via the carboxylate and sulfur groups restricts the rotation of the amino acid at the mineral surface, and hence the positioning of the amino group for nucleophilic attack. While this effect would not be expected to change the equilibrium thermodynamics of amino acid/silicate association, it would affect the kinetic rate of downstream peptide synthesis due to differences in solvent accessibility for the nucleophilic amino group which may differ for bivalently adsorbed L-amino acids compared to D-amino acids.

Amino acid surface attachment results in changes in its electronic environment, which can be detected by NMR spectroscopy. Solid state ^{13}C cross polarisation spectra of silicate/cysteine complexes, prepared through aqueous mixing of cysteine and spec pure SiO_2 followed by drying to preserve solution state interaction complexes, revealed multiple resolvable signals attributable to silicate-bound cysteine (**Fig 4.11a**), which exhibited chemical shifts distinct from that of unbound cysteine or cystine (**Fig 4.11b-c**). For L-cysteine, three $^{13}C_\beta$ resonances with similar or identical chemical shifts were resolvable in samples prepared at both pH 1.5 and pH 10.5 (**Fig 4.11a**), indicating that the sulfur attachment occurs over a wide pH range. The two upfield shifted $^{13}C_\beta$ signals indicated the presence of noncovalent, electrostatically bound complexes of the cysteine sulfur group with the silicate surface. In contrast, the downfield shifted

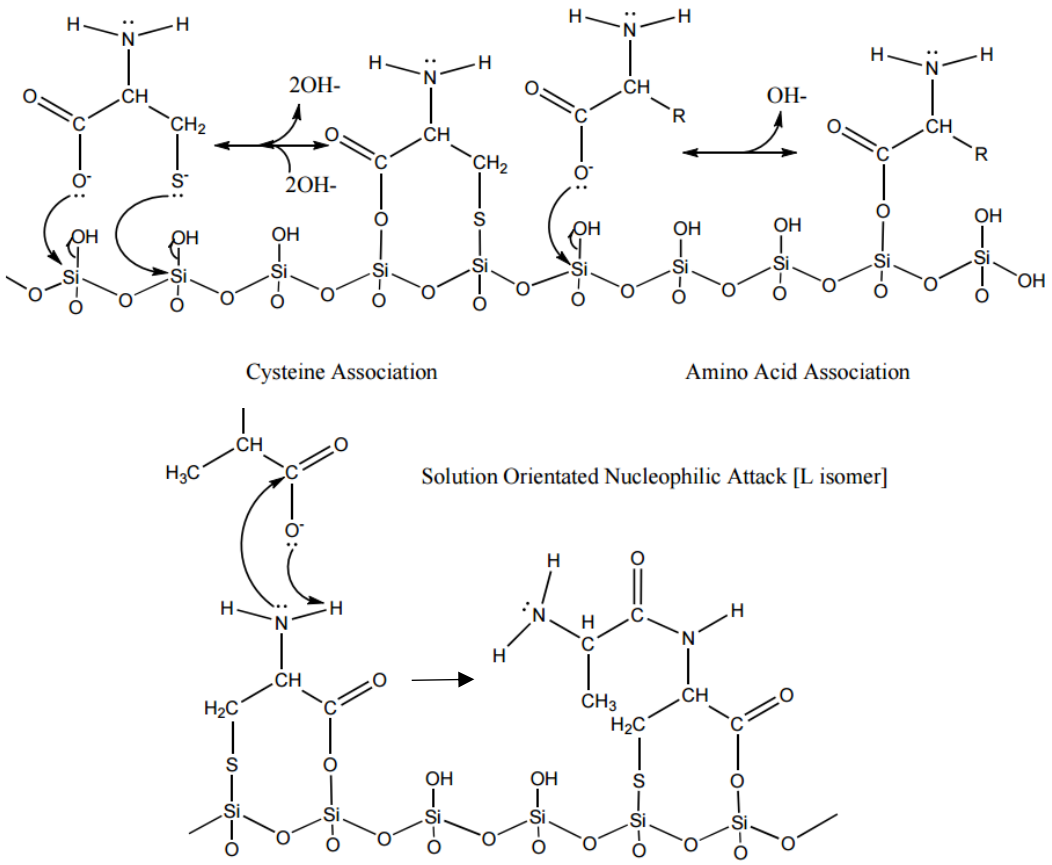
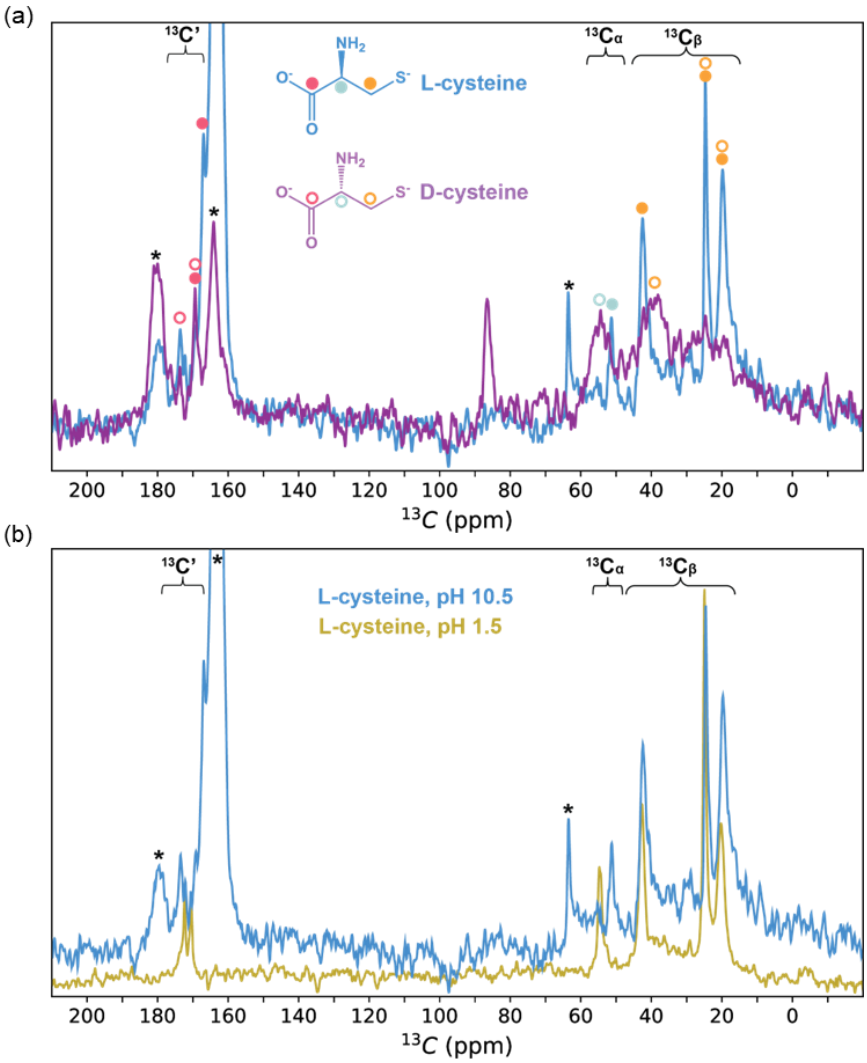


Figure 4.10 – Proposed cysteine catalysed mineral peptide synthesis mechanism.



(c)

	^{13}Cα (*ppm*)	^{13}Cβ (*ppm*)			^{13}COOH (*ppm*)	
L-cysteine, free	54.69	26.81			172.11	
L-cystine, free	52.36	34.01			173.74	
L-cysteine, bound, pH 1.5, monovalent	54.65	42.76	24.91	20.26	172.37	170.54
L-cysteine, bound, pH 10.5, bivalent	51.58	42.74	24.95	20.37	170.82	167.02
D-cysteine, bound, pH 10.5, bivalent	*60.3-51.2 (broad)*	*45.1-37.8 (broad)*	*26.9-24.1 (broad)*	*23.4-17.3 (broad)*	*171.28*	

Figure 4.11 – D-cysteine and L-cysteine association with silicates is bivalent at alkaline pH.

*(**a**) Overlay of solid-state NMR ^{13}C cross polarisation spectra of L-cysteine (blue) and D-cysteine (purple) absorbed to silicate at pH 10.5. The ^{13}C resonances are indicated by coloured circles for L-cysteine (closed) and D-cysteine (open). Chemical shift positions of the enantiomers are similar for carboxylate ^{13}C' signals and differ for the ^{13}Cα and ^{13}Cβ signals. Wash buffer signals at 163.2ppm (bicarbonate), 169.3ppm (carbonate) 179.7ppm (carbonate salt), and 63.7ppm (bicarbonate sideband) are indicated (*). (**b**) Solid-state NMR ^{13}C cross polarisation spectra of L-cysteine absorbed to silicates at pH 10.5 (blue) and pH 1.5 (yellow). The ^{13}Cβ signals are independent of pH, whereas the ^{13}Cα and carboxylate ^{13}C' signals are perturbed at low pH consistent with de/protonation of the carboxylate. (**c**) ^{13}C chemical shifts determined from solid-state spectra. Asterisks indicate the observed chemical shifts consistent with covalent thiolate attachment.*

$^{13}C_{\beta}$ signal at 43ppm agreed with the chemical shift reported for the $^{13}C_{\beta}$ of cysteine bound to gold and likely reflects a covalent silicothioether bond (–Si-S-C(H_2)–)(Abraham *et al.*, 2010). The presence of this covalent bond was confirmed by Polarisation Resolved Infrared Reflection Absorption Spectroscopy (PR-IRRAS) on samples of dissolved cysteine adsorbed to silicate films, which showed an absorption band for p polarised light at 1407cm^{-1} (**Fig 4.12**), consistent with the previously reported IR absorption of the Si-S-CH_2 in silicon bound ethanethiol (1406cm^{-1})(Flach *et al.*, 1997)(Blyholder and Bowen, 1962).

In contrast to the $^{13}C_{\beta}$ signals, the $^{13}C_{\alpha}$ of L-cysteine exhibited a single resonance in all samples but with stereospecific chemical shifts and line-broadening (**Fig 4.11c**). At pH 10.5, the $^{13}C_{\alpha}$ chemical shift (51.58ppm) was shifted upfield relative to the silicate-free cysteine $^{13}C_{\alpha}$ (54.69ppm), consistent with a significant electronic change which is attributed to bivalent binding to silicate. The pH 1.5 chemical shifts of the L-cysteine $^{13}C_{\alpha}$ and carboxylate ^{13}C were consistent with monovalent association and the release of the carboxylate/silicate interaction: the cysteine $^{13}C_{\alpha}$ chemical shift at low pH (54.65ppm) was similar to that of silicate-free cysteine (54.69ppm), and the intense, upfield shifted carboxylate ^{13}C resonance observed at pH 10.5 (170.82 and 167.02ppm) was replaced at pH 1.5 by two downfield shifted carboxylate resonances (172.37ppm and 170.54ppm). Given that the cysteine carboxylate has a pK_a of 1.92, I attributed the signal at 172.37ppm to the protonated carboxylate, similar to that of the free cysteine (172.11ppm), and the signal at 170.54 was attributed to the deprotonated carboxylate (**Fig 4.13**).

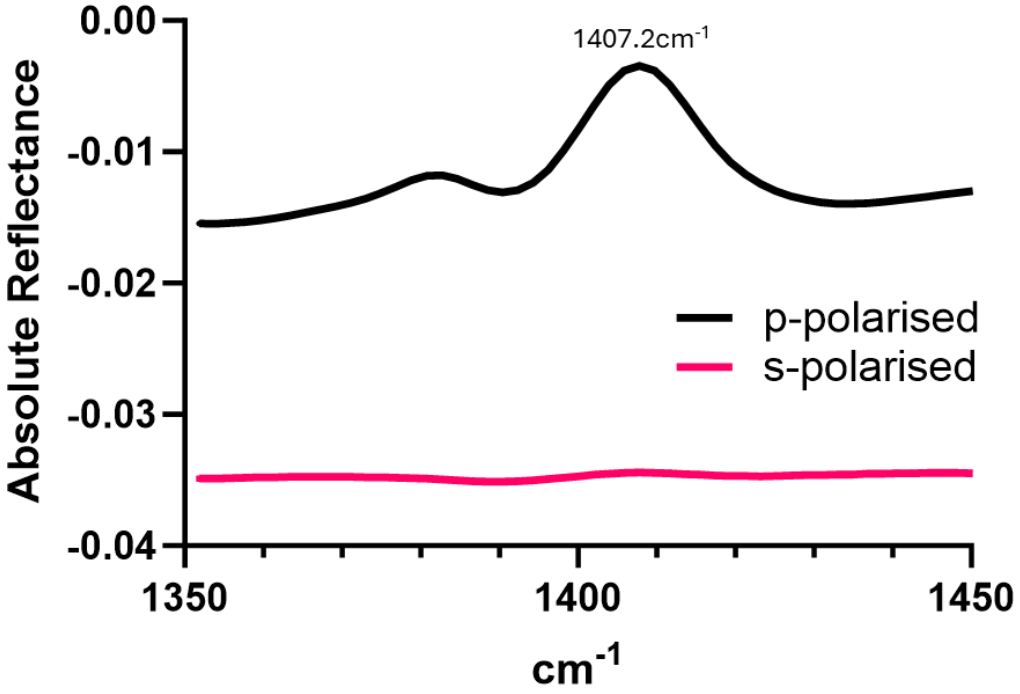


Figure 4.12 – A covalent bond between the cysteine thiolate and silicate is detected by PR-IRRAS.

PR- IRRAS spectra of L-cysteine attached to silicate produces a reflectance band for p-polarised light at 1407.2 cm^{-1}, consistent with a silicothioether bond. No reflectance is observed for s-polarised light indicating that the reflectance for p-polarised light arises from molecules adsorbed at the silicate surface.

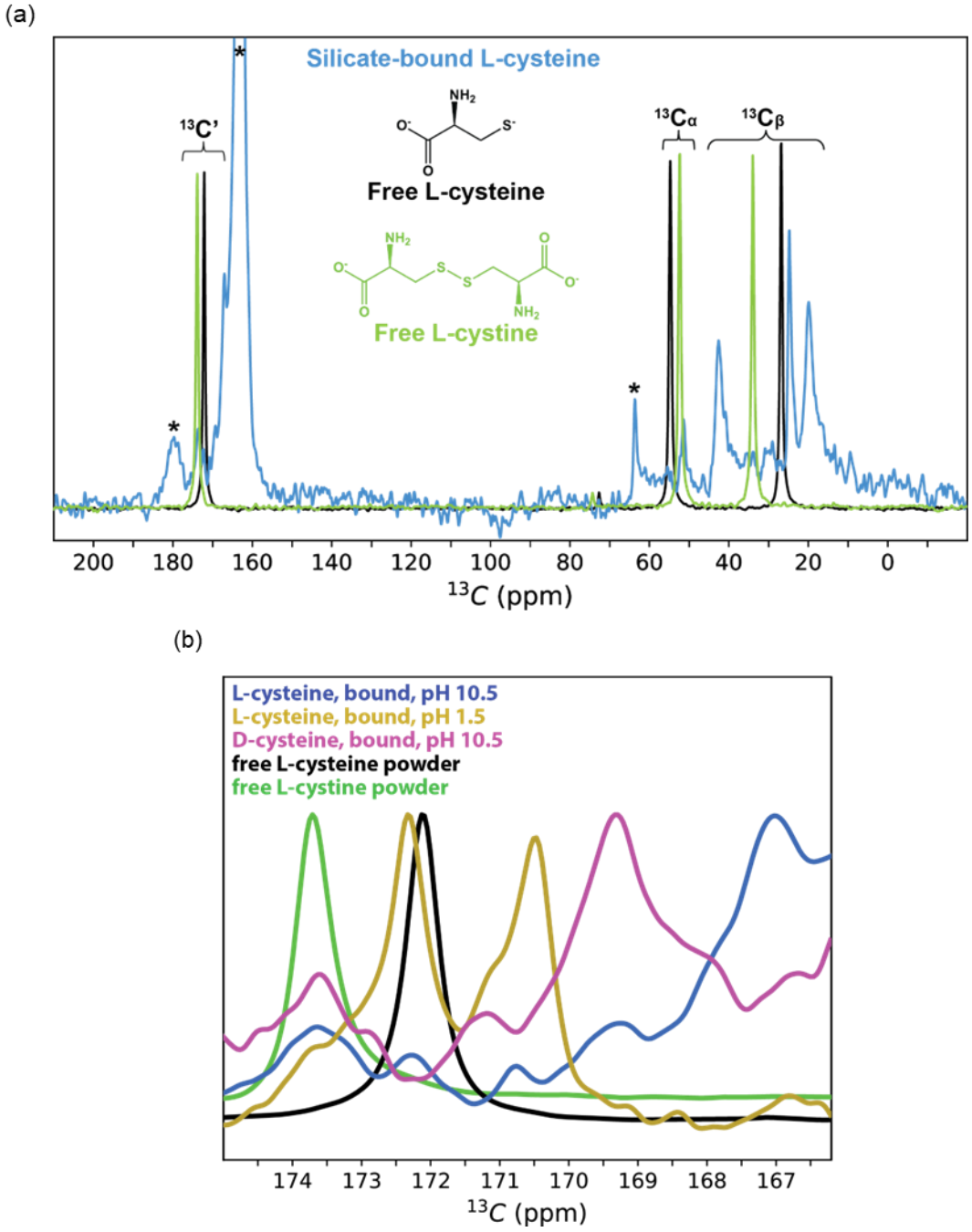


Figure 4.13 – Comparison of the ^{13}C solid-state NMR spectra of silicate-bound L-cysteine and free L-cysteine and free L-cystine.

*(**a**) Solid state NMR ^{13}C cross polarisation spectra of L-cysteine attached to silicate (blue) compared with free L-cysteine powder (black) and free L-cystine powder (green) within the carboxylate region. The regions corresponding to the carboxylate ^{13}C', ^{13}Cα, and ^{13}Cβ signals are indicated. Wash buffer signals at 163.2ppm (bicarbonate), 169.3ppm (carbonate) 179.7ppm (carbonate salt), and 63.7ppm (bicarbonate sideband) are indicated (*) (**b**) The carboxylate region of the solid-state NMR ^{13}C cross polarisation spectra.*

The pH 9.5 NMR spectra of D-cysteine bound to silicate indicated a similar number of $^{13}C_{\beta}$ signals and with similar chemical shifts to the pH 9.5 L-cysteine spectra, consistent with sulfur-driven attachment (**Fig 4.11a)**. These observations confirm that L- and D- enantiomers of cysteine have stereospecific physical topologies when bound to mineral surfaces. These differences were assigned to differences in the surface-bound topology of the enantiomer, resulting in the amino group had differing levels of solvent access and thus capability to engage in a nucleophilic attack on neighbouring soluble amino acids. Given that both species bind to the surface with the same number of interactions, this should not result in a thermodynamic different in silicate affinity.

Based on the solid state ^{13}C NMR results it is highly probable that cysteine attaches bivalently through both the sidechain sulfur and mainchain carboxylate at high pH but is attached predominantly monovalently via a sidechain sulfur at low pH. The association of cysteine with silicate via the thiol group at low pH demonstrates that the catalytic effect from increased residence time on the silicate surface would be applicable across a wide pH range. The stereoselectivity of this catalytic effect is expected to be possible at any pH greater than the pK_a of the cysteine carboxylate (>pH 2.0); to enable bivalent attachment, however detectable peptide synthesis yields are likely restricted to alkaline pH (>9.50), which is required for amino group deprotonation.

Stereospecific reorientation of the amino group in bivalently silicate adsorbed D-amino acids

The different physical properties that arise from bivalent surface attachment of enantiomers were investigated further by PR-IRRAS, which provides bond angle information for surface adsorbed molecules. A stereospecific reorientation of the amino group, which is the site of peptide bond formation, should be observed for amino acids capable of bivalent surface attachment but not for amino acids restricted to monovalent attachment at alkaline pH where the amino group is fully mobile if my hypothesis is correct (**Fig 4.14a**). Following incubation with SiO_2 under aqueous mixing conditions followed by washing and drying under vacuum, all samples including the L- and D-alanine controls showed reflectance in p polarised light but not s polarised light, consistent with the signals arising from surface-bound amino acids. A clear stereospecific inversion of the reflectance corresponding to the amino group (N-H) asymmetric bend was observed for bivalently bound cysteine and aspartate amino acids (**Fig 4.14b-c)** but was not observed in the alanine control (**Fig 4.14d**).

Cysteines reactivity makes the abiotic synthesis of CXXC motif bearing peptides extremely likely.

Monte Carlo simulations were performed to evaluate the impact of cysteine's increased reactivity on the formation of CXXC motifs in abiotically generated 10-mer peptides. In simulations not accounting for cysteine's reactivity, CXXC motifs were not identified if cysteine was present at 1% concentration among the 250 million peptide sequences generated. At the upper bound

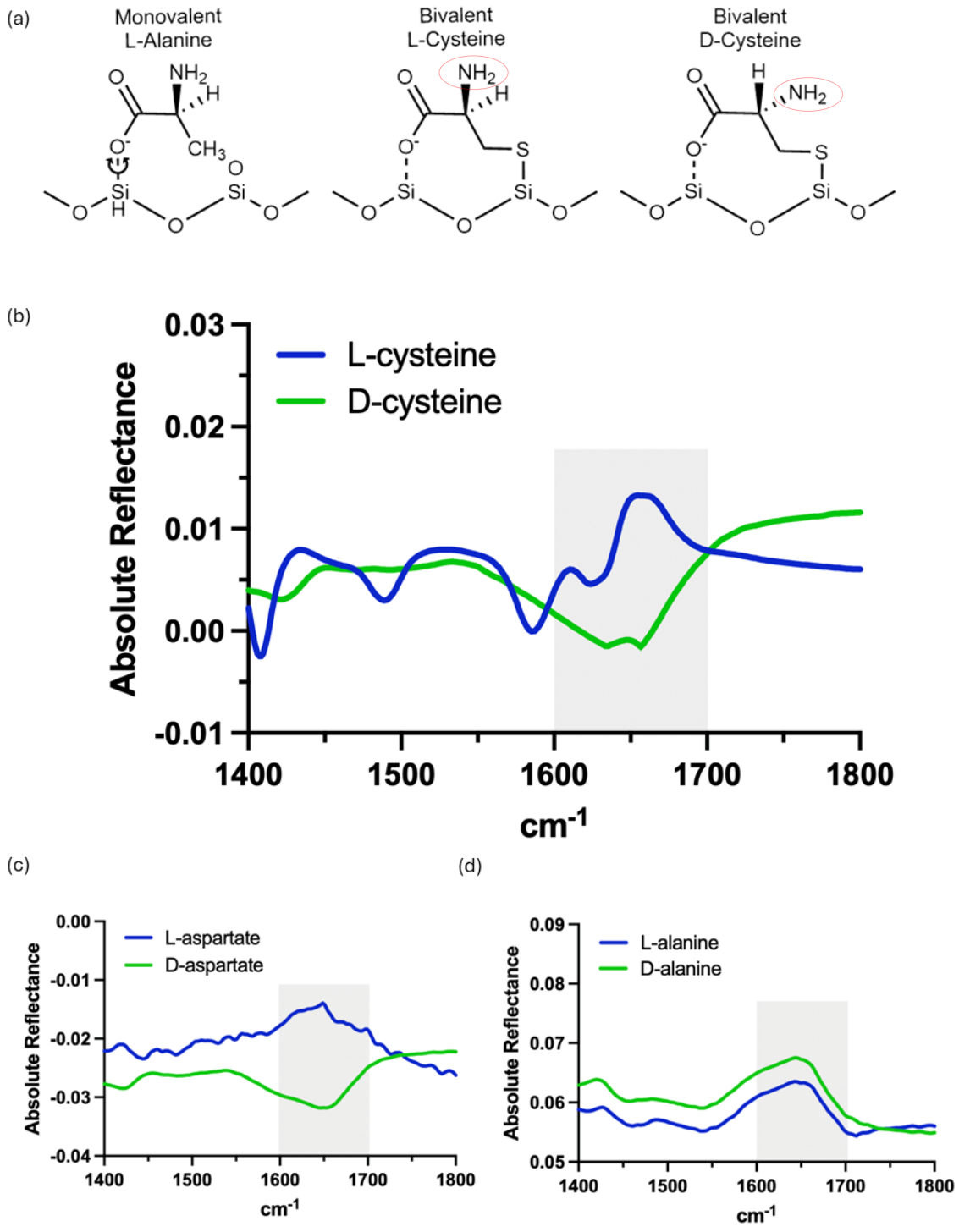


Figure 4.14 – Bivalent surface attachment stereo-specifically orients the amino acid amino group.

*(**a**) Chemdraw model of alanine, and DL-cysteine association at a simplified mineral interface. Bivalent mineral association at alkaline pH prevents molecular rotation at the surface, causing fixed amino group orientations in bivalently adsorbed amino acids only. PR-IRRAS of the enantiomers of **(b)** cysteine, **(c)** aspartate, and **(d)** alanine absorbed to silicate. Stereospecific inversion is observed in the reflectance corresponding to the asymmetric bend of the amino group (~1600-1700 cm^{-1}) in samples containing cysteine and aspartate but not alanine.*

cysteine concentration, the frequency of CXXC motifs increased to 0.00847%. When cysteine's reactivity was incorporated, the frequency of CXXC motifs increased to 0.00794% at the lower bound and 25.354% at the upper bound cysteine concentration (**Appendix B - Table 06**).

Systematic variation of cysteine concentration across ten equally spaced intervals from 0.001 to 0.1 mol fraction revealed approximately linear increases in the probabilities of both CXXC motif formation and the inclusion of at least one cysteine in peptides (**Fig 4.15**). Interpolation of the data indicated that, when considering cysteine's reactivity, a cysteine mol fraction of 0.0094 was required to achieve a 5% probability of CXXC motif formation, compared to 0.0957 mol fraction when not considering reactivity.

The increased probability of CXXC motifs in peptides synthesised in the presence of cysteine, creates a novel pathway to overcome the inherent limitations of peptide synthesis on mineral rocks. Conventional peptide synthesis on minerals is limited to synthesis of peptides typically no greater than 6-8mer due to reducing mineral association strength as the peptide grows. Thio-peptides could be uniquely capable of circumventing this limitation however due to their capability to spontaneously coordinate transition metals, which concentrates the peptides around soluble metal ions, or their much stronger interactions with the mineral itself via its thiol group. This potential method of ligation was explored further.

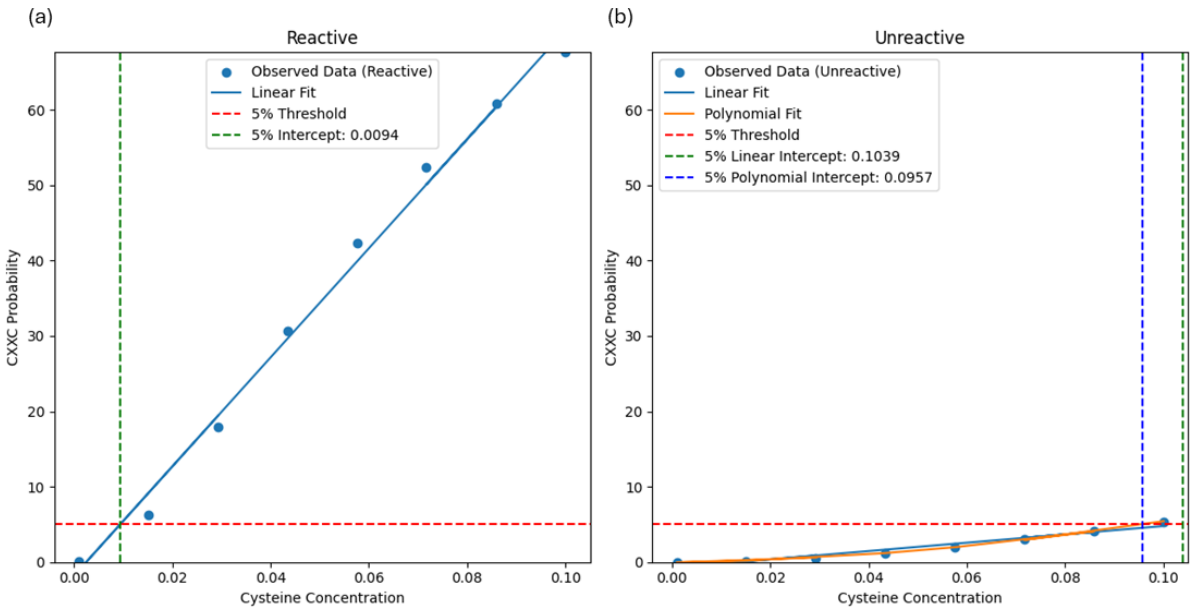


Figure 4.15 – Probability of CXXC Motif Formation in 10-mer Peptides as a Function of Cysteine Concentration, with and without Adjusting for Reactivity.

*Probability of spontaneous CXXC motif creation as a function of cysteine concentration was assessed using Monte Carlo simulations under two conditions: with cysteine reactivity incorporated and without reactivity adjustment. (**a**) When cysteine reactivity was included, simulations demonstrated a linear relationship between required cysteine concentration and CXXC motif probability (blue points and blue line). Interpolation revealed that a cysteine concentration of 0.0094 mole fraction was required to achieve a 5% probability of CXXC motif formation in a 10mer peptide synthesised randomly from abiotically available amino acids according to their relative concentration (dashed green line), with the 5% threshold indicated by the red dashed line. (**b**) Without accounting for cysteine's enhanced relative reactivity, simulations produced a polynomial trend (blue points and orange line). A linear fit of the data indicated that a cysteine concentration of 0.1039 mole fraction was required to achieve a 5% probability (dashed green line), while a polynomial fit reduced this to 0.0957 mole fraction (dashed blue line).*

Metal ions catalyse ligation of thiol heptapeptides.

To test whether metal ions catalyse the ligation of thiol-containing peptides, a peptide based on one identified in **Appendix B - Table 05a**, designated Arc4, was synthesised and mixed with $FeCl_3$ and $CoCl_2$. Analysis by UV-Vis demonstrated that Arc4 was capable of spontaneous Fe^{3+} and Co^{2+} coordination in solution under alkaline conditions (**Fig 4.16a,d**). Deconvoluted difference spectra were acquired (**Fig 4.16b,e),** clearly showing the acquisition of absorbances consistent with Ligand Metal Charge Transfer (LMCT) at 271 and 280nm for a thiol metal coordination complex (**Fig 4.16c,f**). Additional d-d transition peaks were observed caused by increased splitting of the metal ion d orbital upon coordination by a sulfur bearing group. More and better resolved peaks were observed in the peptide/Fe^{3+} samples consistent with a more tightly bound interaction complex compared to the peptide/Co^{2+} complexes. peptide/Zn^{2+} complexes could not be monitored through this technique as thio-Zn^{2+} complexes are spectroscopically silent and do not generate UV-Vis peaks. This result was particularly surprising given that Fe^{3+} and Co^{2+} typically form insoluble hydroxides at this pH range, suggesting that coordination by Arc4 was sufficient to prevent precipitation.

Arc4 was investigated further for its ability to ligate in the presence of a metal ion via TSPL (**Fig 4.01**) and was heated at 90°C in the presence of iron, zinc, or cobalt in silicate glassware in a 5% H_2 95% N_2 atmosphere. Notably, when the experiment was repeated without prior equilibration in 5% H_2, extensive oxidation of the thiol group was observed in the peptide which prevented ligation

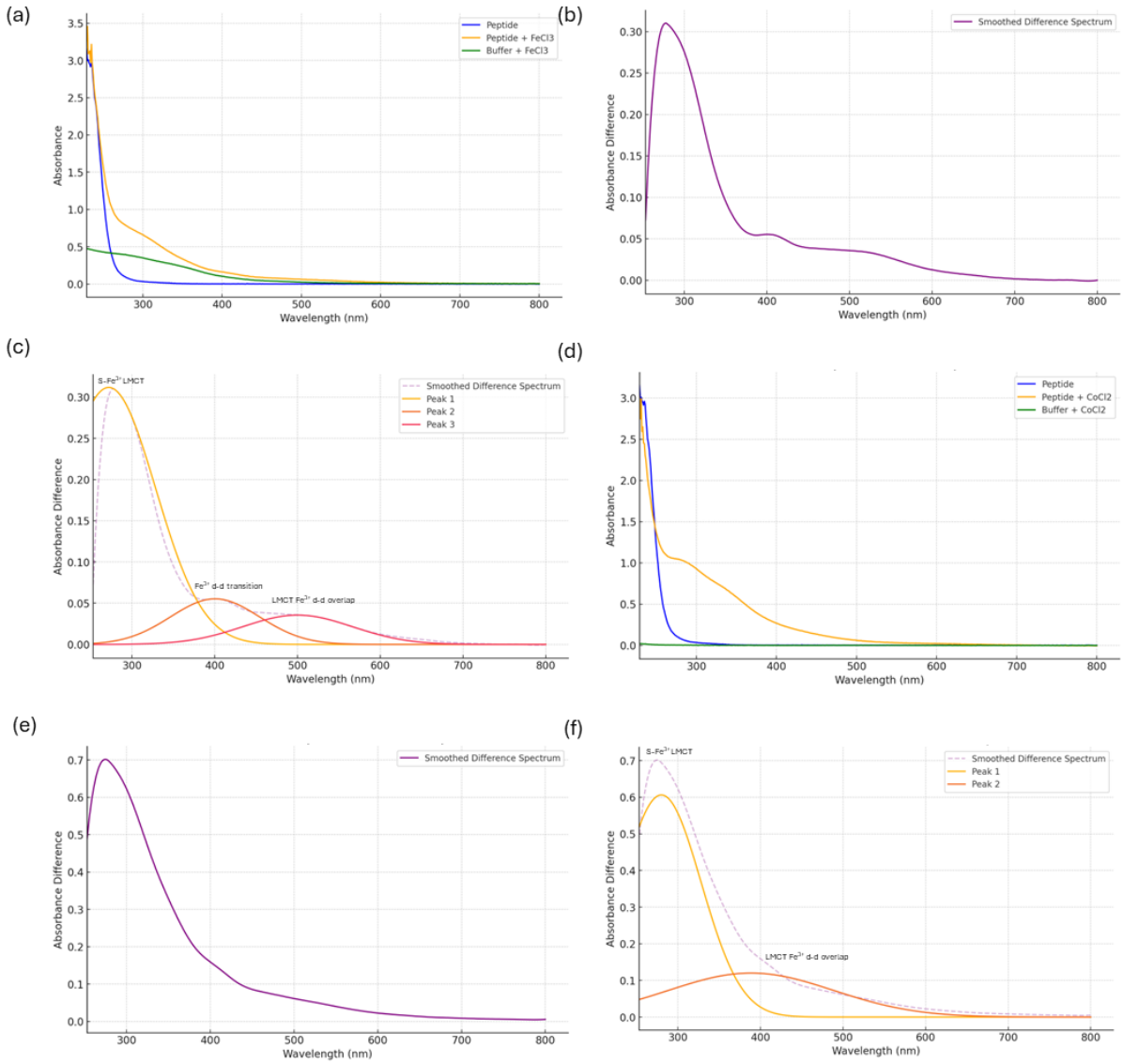


Figure 4.16 – Arc4 spontaneously coordinates Fe^{3+} and Co^{2+} under alkaline conditions.

*Arc4 (CGSCSST) was resuspended into 0.08M bicarbonate buffer (pH 9.5) to 1mg/ml and analysed by UV-vis before and after mixture with $FeCl_3$ (**a-c**) and CoCl2 (**d-f**) to 0.13mM. Difference spectra were acquired by subtraction of peptide only, and metal only spectra and smoothed using the Savitsky-Golay function, and deconvoluted using the gaussian fitting functions in Python. Identified peaks were centred at (**c**) 271, 400 and 500nm and (**f**) 280 and 388. Peaks identified are consistent with spontaneous Fe^{3+} and Co^{2+} coordination by Arc4 with the large peak system at 271, and 280nm respectively likely representative of ligand to metal charge transfer absorbances and the remaining peaks coming from d-d transitions that become visible in UV when iron or cobalt is coordinated by sulfur. The sharper peaks observed for the Arc4-Fe^{3+} spectra is consistent with a more tightly bound coordination sphere, resulting in improved peak resolution.*

occurring (**Appendix A - Figure 05**). Intriguingly, fully oxidised cysteine, (cysteine sulfenic acid) was not observed in our peptide synthesis experiments previously, suggesting that soluble metal ions catalyse thiol oxidation if the solution is not sufficiently reducing. These results suggest that a reducing atmosphere is essential to any thiol/metal ion-based ligation reaction pathway and that ligation depends on the presence of the thiol group.

Mass spectrometry analysis of Arc4 before (**Fig 4.17**) and after (**Fig 4.18**) heating in the presence of Zn^{2+}, and Fe^{2+} showed a mass detection of a 14mer product of Arc4 within 5 ppm of the predicted mass. This was detected only in reactions containing iron or zinc and was not observed in reactions containing Co^{2+} or containing only silicate material.

^{1}H NMR analysis of the non-heated and Zn^{2+} samples showed the appearance of new $^{1}H_{N}$-$^{1}H_{\alpha}$ and $^{1}H_{N}$-$^{1}H_{\beta}$ resonances proximal to those of the unreacted monomers (**Fig 4.19**). The acquisition of these new resonances is consistent with the generation of new peptide bond containing organic species that is similar to the unreacted monomer but in a different chemical environment, as would be expected upon chemical ligation of the original 7mer peptide. These results demonstrate that peptide ligation of transition metal binding thiol-bearing peptides occurs in these reaction conditions and that transition metals facilitated the ligation process. Notably the Co^{2+} samples did not show any evidence of the ligation product despite being capable of metal coordination by UV-Vis. This was assigned to a looser and less well-defined coordination complex, as suggested by the UV-Vis results, which may prevent peptide ligation from proceeding. This

would also suggest that tight metal ion coordination complexes are required for peptide ligation through this mechanism.

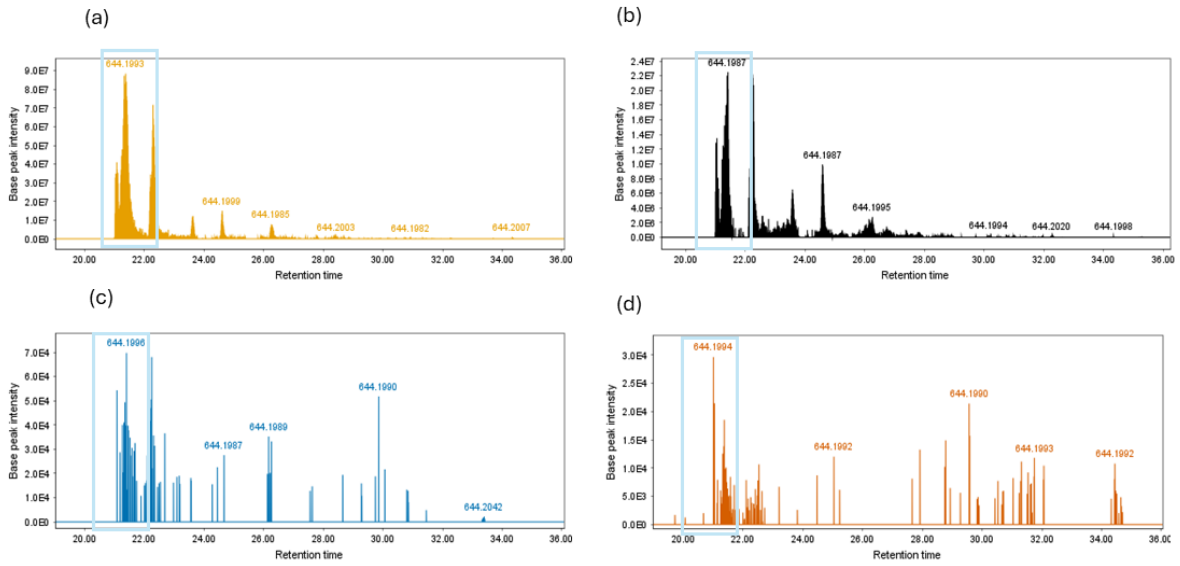


Figure 4.17 – Arc4 is stable at 90℃ and pH 9.5 for up to 36 hours.

*The Arc4 peptide (644.201 ± 5ppm) was observed by extracted ion chromatogram following Hybrid Quadrupole-Orbitrap Mass Spectroscopy after peptide heating at 90°C and pH 9.5 for 36 hours in the presence of $ZnCl_2$ (a), $FeCl_2$ (**b**), $CoCl_2$ (**c**), and no added metal (negative control) (**d**). After heating, peptides were lyophilized, resuspended in 0.1% formic acid, and injected onto the spectrometer. These results confirm that Arc4 is stable under alkaline and high-temperature conditions for at least 36 hours in the presence of a reducing atmosphere. The increased signal to noise observed in (**a**) and (**b**) indicate that presence of Zn^{2+} and Fe^{2+} stabilised the peptide against thermal degradation.*

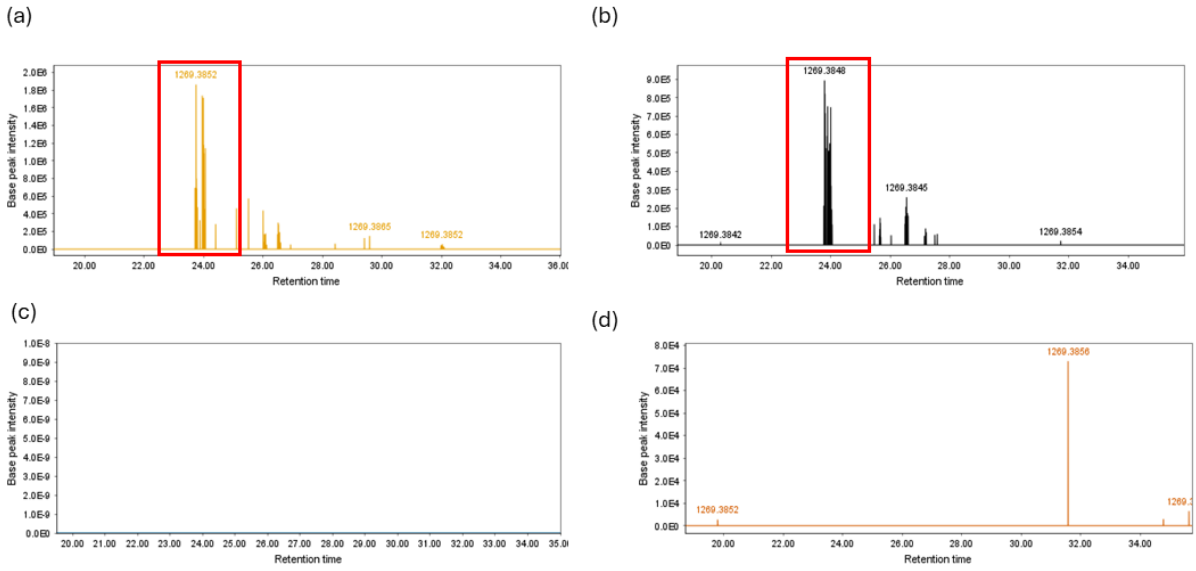
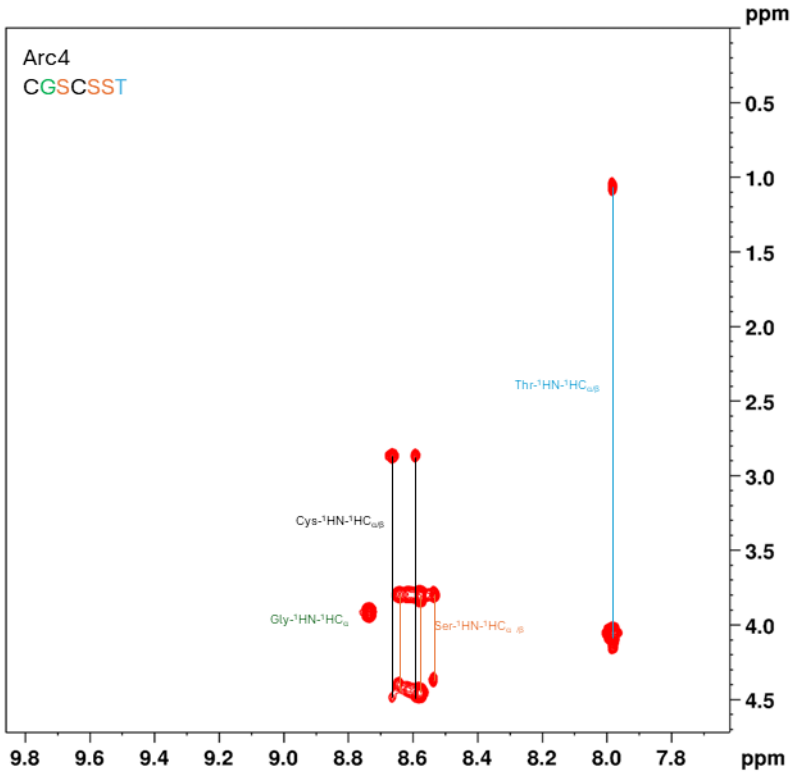


Figure 4.18 – Arc4 is ligated in the presence of Zn^{2+} and Fe^{2+} under reducing conditions.

*The ligated Arc4 peptide (1269.385 ± 5ppm) was detected by extracted ion chromatogram following Hybrid Quadrupole-Orbitrap Mass Spectroscopy after heating at 90°C and pH 9.5 for 36 hours in the presence of $ZnCl_2$ (**a**), $FeCl_2$ (**b**), $CoCl_2$ (**c**), and no added metal (negative control) (d). Following heating, peptides were lyophilized, resuspended in 0.1% formic acid, and injected onto the spectrometer. The ligated Arc was observed at 24 minutes retention time and was only seen when ligation occurred in the presence of $ZnCl_2$ (**a**) and $FeCl_2$ (**b**), consistent with the predicted ligated Arc4 peptide, while no peaks were detected in $CoCl_2$ (**c**) or negative control samples (**d**), indicating that ligation did not occur under these conditions.*

(a)



(b)

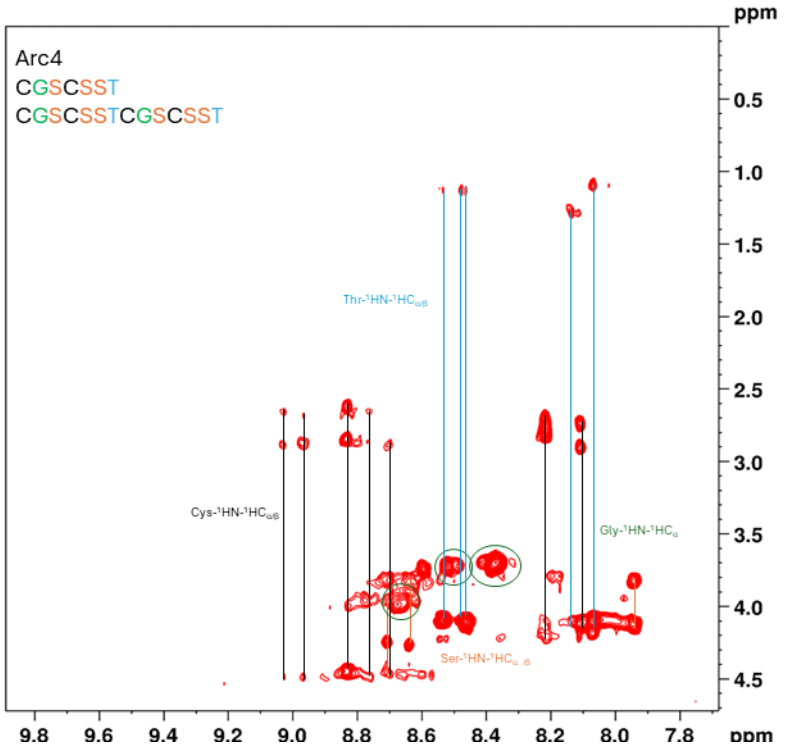


Figure 4.19 – New ^{1}HN-^{1}HCα and ^{1}HN-^{1}HCβ resonances following heating in the presence of Zn^{2+} are consistent with peptide ligation.

*2D ^{1}H-^{1}H TOCSY spectra were conducted before (**a**) and after (**b**) Arc-4 peptide ligation in the presence of $ZnCl_2$ Significant numbers of new crosspeaks were observed. These results were consisted with the hypothesis that ligation of Arc4 occurs only in the presence of Zn^{2+} or Fe^{2+} under reducing conditions and requires heating at alkaline pH.*

Discussion

Cysteine as the enabler of high yield abiotic peptide synthesis in aqueous conditions

This study investigated whether the inclusion of cysteine enhances abiotic peptide synthesis under conditions analogous to white smoker hydrothermal vents, which are alkaline, moderately saline, and reducing environments.

The results demonstrate that the addition of cysteine to a conventional set of ten abiotically available amino acids significantly increases peptide yields when heated in the presence of an oxide mineral under alkaline conditions. Initial observations by UV-Vis were corroborated by ^{1}H, ^{1}H-^{13}C NMR and TIMS-TOF MS. High peptide yields depended on the presence of a mineral substrate and a thiol bearing amino acid, indicating that peptide formation occurs predominantly through surface catalysis via a rate enhancing silico-thiol interaction.

Stereoselective peptide synthesis by bivalent mineral association

The stereoselectivity of the yield was 40% greater than observed in previous salt-induced peptide synthesis experiments (10% L vs D synthesis bias) and achieved in amorphous borosilicate glass and on pyroxene without reliance on chiral templating minerals (Plankensteiner *et al.*, 2004)(Hazen *et al.*, 2001). The enhanced peptide yields with L-cysteine was predicted to be attributed to a stereospecific orientation of the amino group when bivalently bound to a silicate surface. The amino group is the site of peptide bond formation, and its reorientation may change solvent accessibility, which in turn affects the rate of peptide bond formation. As such this effect is not driven by thermodynamic

differences in amino acid surface adsorption but rather differences in solvent access for the amino group between the different adsorbed species.

Solid-state ^{13}C NMR and polarisation-modulation infrared reflection-absorption spectroscopy (PM-IRRAS) confirmed that cysteine binds bivalently to silicate surfaces via a covalent Si-S bond, involving both the thiol and carboxylate groups. The association of cysteine with silicate was notably strong, and it remains bound after multiple washes, in contrast to glycine, which adsorbs weakly to silicate surfaces and is almost completely removed after two wash cycles (Lopes *et al.*, 2009). The size of the ppm shifts observed by solid state NMR, particularly for silicate-bound cysteine $^{13}C\alpha$ and $^{13}C\beta$ resonances, are too large to have occurred from contaminants within the sample and were consistent with the creation of new covalent bonds proximal to these groups. Although the cysteine thiol interacts with silicates across a wide pH range, stereoselective and high peptide yields can only be achieved at alkaline pH because of the requirement for amino group deprotonation to achieve high peptide yields (Rodriguez-Garcia *et al.*, 2015). The cysteine mineral attachment is reversible in peptides, as evidenced by the presence of readily purified cysteine-bearing peptides in the supernatant.

There is also evidence of bivalent attachment of aspartate to silicate, as previously suggested by Erastova *et al,* which results in a corresponding stereospecific orientation of the aspartate amino group by PR-IRRAS (Erastova *et al.*, 2017). However, the much weaker nucleophilicity of the carboxylate, and the correspondingly weaker silicate association, prevented the observation of any stereoselective effect by UV-Vis due to the much lower peptide yields.

Synthesised peptide lengths are inherently limited in mineral catalysed peptide synthesis.

The presence of L-cysteine increased the lengths of peptides that were produced compared with either D-cysteine or L-alanine. Our ability to detect longer peptides is likely due in part to an overall increase in peptide yields, which suggests that cysteine incorporated into peptides retains an increased affinity for silicate surfaces.

Although the longest L-cysteine bearing peptides observed were longer than those detected here for L-alanine or reported elsewhere from oxide mineral catalysed syntheses (Cleaves *et al.*, 2009), these lengths, are much smaller than a modern, protein domain, which is typically hundreds of amino acids long. Indeed, model systems that are restricted to meteoritic amino acids have provided only low yields of peptides of typically no more than four amino acids (Cleaves *et al.*, 2009), irrespective of concentration. Although it has been proposed that a combination of surface-catalysed peptide elongation, and peptide ligation in solution, facilitated by wet/dry cycles plausibly in sub-aerial puddles, may overcome the apparent barrier to peptide length in surface catalysed reactions (Rode *et al.*, 1999), this has not been experimentally observed with the proteinogenic amino acid set. However, the enhanced reactivity of cysteine, its spontaneous metal binding capability and strong affinity for silicates could provide a mechanism to overcome these length limitations. Monte Carlo simulations weighted using cysteines increased reactivity, show that if cysteine makes up only 1% of total abiotic amino acids there is a 5% probability that a 10mer peptide generated abiotically will contain a CXXC motif.

Reducing conditions enabled thiol-peptide selective peptide ligation.

Cysteine-enriched peptides exhibited an ability to facilitate metal ion-catalysed ligation. A thiol-containing heptapeptide spontaneously coordinates Co^{2+} and Fe^{3+} in solution, preventing the metals precipitation at pH 9.5. When heated in the presence of Fe^{3+} and Zn^{2+}, a thiol peptide was shown to undergo ligation to form a 14mer peptide. Notably the mass of the 14mer generated through condensation polymerisation of a Arc4 was distinct (by ≈-14 Daltons) from a disulfide-bonded 7mer dimer, which was made even more unlikely by the presence of 5mM TCEP in the injected sample. ^{1}H NMR and mass spectrometry confirmed the formation of the 14-mer under these conditions. Metal ion-catalysed ligation did not occur in the absence of Fe^{2+} or Zn^{2+}, or when the thiol peptide had been fully oxidised, which indicates that the metal-thiol coordination plays a crucial role in this process. Notably, despite successful coordination of the thiol bearing peptide in solution, peptide ligation did not proceed in solutions with Co^{2+}, suggesting that the coordination complex itself influences peptide ligation. The coordination of thiol peptides with Fe^{2+} and Zn^{2+} likely involves higher affinity coordination complexes that better stabilise peptide termini in close proximity and, hence enabling ligation. In contrast, the looser, more dynamic coordination observed with Co2+, may not maintain the spatial alignment required for peptide bond formation.

The data also indicates that this ligation mechanism is restricted to thiol-bearing peptides, particularly those containing a compatible coordination motif. This is because only unoxidized cysteine side chains are capable of creating the

coordination complexes required for bringing the separate peptide termini into proximity.

This mechanism is reminiscent of the salt-induced peptide formation (SIPF) route in which amino acids chelate around Cu^{3+} at extremely high (>3M NaCl or KCl) salt concentrations (Plankensteiner *et al.*, 2004). Notably the TSPL reaction took place with 0.1M Na^{+} present from the bicarbonate buffer, which was not sufficient to cause ligation in the controls, indicating that transition metals are essential to this process. As well as being metal selective, this pathway is redox sensitive, owing to the oxidation of cysteine, which prevents the thiol peptide coordinating metal ions. As such, peptide synthesis and ligation are restricted to redox potentials of lower than -220mV locally to the site of peptide synthesis, conditions that are most plausibly found in ancient off-axis, so called white smoke alkaline hydrothermal vents.

There is no reason that the TSPL mechanism proposed here could not be repeated over multiple cycles and hence generate longer proteins relevant to the emergence of biology. The TSPL mechanism therefore addresses a critical challenge in previous abiotic protein synthesis studies by providing a method to overcome the inherent limitations of mineral based pathways as well as constraining peptide construction to alkaline hydrothermal vents, rather than transient sub-aerial pools.

Limitations

While the data presented here is compelling, it is sensible to acknowledge limitations in its applicability. The experimental design aims to constrain the free

parameters present in any natural system, and to explore the various physicochemical effects on the efficacy of peptide synthesis. By virtue, the study focuses upon silicate and oxide minerals, controlled pH, and synthetic amino acid mixtures to simulate prebiotic conditions, which, while experimentally defensible, cannot fully replicate the complexity of a natural largely unknown ancient environment. The inevitable gradients present primarily in temperature, E_h and pH, competing chemical species and, to a lesser extent, pressure present in hydrothermal systems may influence both cysteine availability and peptide synthesis kinetics. Additional competing non-peptide productive reaction products were not explored in depth here due to the complexity of the analysis which was beyond the scope of my analysis but likely would play an important role in a real-world system. Future work incorporating more realistic environmental fluctuations, alternative mineral systems, and more abiotic amino acid candidates would enhance the robustness of the conclusions drawn here.

An additional consideration to the claim of increased peptide synthesis yield made here is the comparatively high amino acid concentrations used (mM rather than μM). Experimentally, the increased concentrations, although perhaps not relevant to a naturally system, were necessitated by the requirement to accurately identify run-products and to expedite the detection of peptide synthesis. However, unless there is a significant non-ideal behaviour of amino-acid interaction at these elevated concentrations, which is perhaps unlikely, this work should extrapolate to lower concentration scenarios with correspondingly increased reaction times. To check that peptide yields are similarly increased in lower concentration systems, future experiments should be conducted on this

system at more realistic local amino acid concentrations at longer durations. Supporting the notion that concentration alone does not invalidate these results, I do note that stereospecific effects on yield were also observed to at least 55μM when heated for up to 6 weeks.

I also note that this work is highly dependent on the work of Foden *et al,* who propose that their abiotic cysteine synthesis mechanism is widely generalisable (Foden *et al.*, 2020). This seems eminently plausible but given the importance of their work on this study, and the significant consequences for the emergence of life that arises from their cysteine synthesis mechanism, it would be prudent to confirm its applicability under alkaline hydrothermal conditions.

Due to time constraints, my observations of metal ion-catalysed ligation, while compelling, is based on a single peptide. Future experiments should therefore be conducted with different peptides of varying length and sequences to determine if this mechanism is viable for abiotically preferential CXXC peptide elongation. I was unfortunately unable to evaluate whether this process occurs in aspartate or non-anionic peptides. While it's unlikely this pathway would proceed at mild salinities for such non-anionic amino acids peptides, given the high salinities needed for SIPF to proceed for even amino acids to spontaneously coordinate transition metal ions, aspartate bearing peptides might feasibly form similar structures around magnesium or calcium ions in water and undergo ligation as observed here for thiol peptides. Given the bioavailability of both Ca and Mg cations in seawater, and their use by modern life (e.g. magnesium role in *oxidative phosphorylation, and glycolysis*), this is probable. However, given the lower nucleophilicity of aspartate compared to cysteine, and the results presented in

Chapter 2-3, I predict that this is less probable in the nascent early oceans. Thus, future studies should investigate a much larger library of peptides and how variations in environmental factors, such as pH or ionic strength, influence the efficiency of thiol-driven ligation. Additionally, exploring the impact of cysteine placement within peptide sequences (e.g., CXXC vs. CXC motifs) could provide insights into the sequence-specific constraints of this pathway.

Conclusion

In conclusion, my findings support the hypothesis that cysteine enhances abiotic peptide synthesis by increasing peptide yields and facilitating metal ion-catalysed ligation. The inclusion of cysteine in surface-catalysed peptide synthesis under alkaline, moderately saline, and reducing conditions provides a plausible mechanism for the emergence of functional peptides and contributes to our understanding of the minimal amino acid set required for early protein function. These results reinforce the potential significance of cysteine in the origin of life and suggest once again that environments capable of supporting cysteine availability are critical for abiogenesis.

The robust enhancement of peptide yields under alkaline, reducing conditions featuring cysteine, and the stereoselective bias toward L-cysteine, confirms Hypothesis 6 (that “warm, reducing, alkaline and moderately saline conditions support abiotic protein synthesis and function”). In addition, the need for a reducing environment to maintain cysteine’s thiol group supports Hypothesis 5, which predicted that reducing conditions are required for incorporating transition metals. My observation that Fe^{2+} and Zn^{2+} facilitate peptide ligation in

cysteine-containing peptides directly ties back to Aim 4, demonstrating that cysteine's unique capabilities can enable metal-driven polymerisation.

While these findings highlight the importance of cysteine's thiol chemistry in peptide ligation and metal coordination, it should be acknowledged that aspartate, due to its greater abundance in prebiotic environments, may also support metal-catalysed peptide ligation through its carboxylate sidechain. Although the peptides tested here in the ligation experiments were limited to thiol-bearing sequences, this mechanism could be equally relevant to aspartate-rich peptides. Aspartate-based peptides, while not redox active, could facilitate alternative abiotic processes and offer a parallel route for peptide elongation and function under early Earth conditions. This represents a significant future avenue for investigation, as aspartate-rich peptides may have supported abiotic chemical pathways distinct from those enabled by cysteine.

Collectively, these findings complete Aim 1 by further refining the solvent constraints, pH ≈9.5, moderate temperature (≈90°C), reducing atmosphere, to favour peptide bond formation. They also support Aim 3 (the minimal amino acid set) and Aim 2 (demonstrating plausible abiotic-like proteins), showing that adding cysteine can surmount the short-chain limitation typical of purely "abiotic amino acid" reactions.

Chapter 5: Evaluation of the functional planetary habitability of Earth, Mars, and Enceladus

Recap

In Chapter 1, I set out a series of predictions regarding the behaviour and solvent requirements of abiotic proteins and established overarching aims to:

1. Define the solvent environments conducive to abiotic protein folding and function.
2. Model abiotic-like proteins using exclusively abiotic amino acids, assessing their stability and folding behaviour.
3. Propose the minimal amino acid set required for protein folding and activity in abiotic conditions.
4. Consider the impact of including or excluding cysteine in abiotic protein systems, given recent evidence on its abiotic synthesis.

These aims were based on the likely fact that early proteins may have solvent specific restrictions on their folding and function (Di Mauro *et al.*, 2012). By modelling abiotic-like proteins, I predicted that this would create a protein function 'solvent window' to determine which solvents, and which organic species are needed for the function of these proteins. Based on *a priori* assessments, I made the following predictions which related to each aim as indicated

- Hypothesis 1: Cysteine is essential to enable stable transition metal coordination in abiotic proteins. (Aim 3,4)
- Hypothesis 2: Acidic pH at or near pH 4 to 5 will cause protonation of acidic residues and induce unfolding in abiotic peptides and proteins. (Aim 1,2)
- Hypothesis 3: Abiotic peptides and proteins adopt molten globule states rather than fully folded tertiary structures. (Aim 1,2)

- Hypothesis 4: Increasing salinity does not significantly alter abiotic peptide and protein conformation due to an absence of tertiary folding interactions. (Aim 1,2)
- Hypothesis 5: Reducing conditions are required for abiotic peptides and proteins to incorporate transition metals. (Aim 1)
- Hypothesis 6: Warm, reducing, alkaline and moderately saline conditions provide suitable conditions for abiotic protein synthesis, folding and function. (Aim 1)

I predicted that the absence of thiol-bearing amino acids, namely cysteine, would likely preclude the formation of stable transition metalloprotein architectures that is highly likely required for early life to conduct redox chemistry within the first protocells (Shalayel *et al.*, 2020). Conversely, I predicted that if cysteine is available locally during abiogenesis than this would enable early life to coordinate transition metals to enable abiotic protein folding and catalysis to bring about an RNA world. Work by Vallee *et al* has also predicted that thiol rich peptides could be precursors to the RNA world, although their reaction schemes in support of this theory are always shown independently of mineral based interactions (Vallee *et al.*, 2017).

Each of these hypotheses was addressed through a combination of computational analyses, experimental characterisations, and peptide synthesis experiments. From this, I am now able to integrate my results with my predictions on abiotic protein behaviour as well as the proposed conditions on the Archean Earth, Noachian Mars, and modern Enceladus, to attempt to describe their planetary habitability through the lens of abiotic protein function. The subsequent

sections will examine how the constraints revealed in Chapters 2, 3, and 4 compare against the theoretical expectations set out in Chapter 1.

Experimentally derived abiotic protein folding constraints.

Study 1

The investigations of Chapter 2 were designed to disprove my hypothesis on cysteine's criticality to abiotic transition metal complex formation and test whether the absence of cysteine could be circumvented by other abiotic amino acids. Within this study, rubredoxin served as a model ancient metalloprotein scaffold. By substituting rubredoxins cysteines with aspartate and serine, which are abundant abiotic amino acids, the goal was to ascertain whether early proteins might have relied on these non-thiol ligands to achieve stable transition metal binding.

Following the introduction of aspartate and serine within the metal binding site in the rubredoxin model system, no transition metal coordination was detected under the tested conditions. At best, transient, and nonspecific interactions occurred, but these were inadequate to induce or maintain a functional metal coordination sphere. These results were consistent with those of Jordan *et al,* and I observed that even in the monomeric state, aspartic acid could not substitute for cysteines thiol chemistry (Jordan *et al.*, 2021).

These findings are consistent with my initial prediction that thiol chemistry may be the most effective route for early metalloprotein architectures, at least under these acidic and neutral conditions. Prior to this study, one might have envisioned that other carboxylate-bearing amino acids, notably aspartate or

glutamate, could substitute for cysteine in a proto-metalloprotein if suitably positioned within a folded protein. The data provided evidence to suggest that such a scenario is not feasible under the conditions tested. Instead, softer metals like iron and zinc likely demand correspondingly soft ligands, and in the absence of cysteine or functionally equivalent thiol or imidazole bearing amino acids, early proteins would likely have a limited ability to harness the benefits of transition metal coordination to deliver redox function within the expected condition ranges of the Archaeon ocean (Pearson, 1963)(Harel *et al.*, 2014)(Catling and Zahnle, 2020).

These results suggest that the environments most conducive to early functional protein evolution are the same conditions that support the abiotic synthesis of cysteine or related thiolate amino acids, namely reducing neutral or alkaline water in the presence of ferrocyanide (Foden *et al.*, 2020). Without thiol-bearing residues, even optimal sequences of abiotic amino acids would likely create abiotic proteins that remain functionally limited to the chance generation of autocatalytic structures (Rout *et al.*, 2022). On the other hand, as soon as cysteine became either abiotically available, through the pathways noted by recent studies, or evolutionarily incorporable, it would dramatically expand the functional inventory of emerging proteins (Vallee *et al.*, 2017)(Shalayel *et al.*, 2020). This shift could open the door to redox chemistry and, by extension, to more advanced protometabolic networks that could drive the progression from simple chemical systems toward an RNA world (Martin *et al.*, 2008)(Fried *et al.*, 2022). This would support the findings of recent studies suggesting that protometabolism were organo-sulfur based (Goldford *et al.*, 2019). These

findings also support Hypothesis 1, which posits that cysteine is essential for stable transition metal coordination in abiotic proteins. By showing that aspartate and serine could not compensate for thiol chemistry, Chapter 2 also addresses Aim 3 and 4, demonstrating the unique impact of cysteine's thiol group on abiotic metalloprotein assembly.

Study 2

The second study, described in Chapter 3, sought to broaden the investigation beyond a single model protein to a larger landscape of naturally occurring proteins mirroring the abiotic amino acid composition predicted on early Earth. By selecting multiple candidate proteins and characterising their structural stability profiles, this approach addressed the central question posed in the introduction: how well could abiotically derived amino acids, lacking many modern residues essential to stable folding, fold into functional or even partially ordered structures?

Although the identified abiotic-like proteins varied in sequence and predicted secondary structure, ranging from helical, which dominated the dataset, and β-solenoid motifs to more mixed α/β architectures, all purified proteins shared the hallmark characteristics of molten globularity as predicted (Di Mauro *et al.*, 2012). Their folding states showed only secondary structure formation, and their stabilities were strongly influenced by pH. None achieved the kind of cooperatively folded tertiary structure seen in contemporary enzymes.

The observation that introducing cysteine improved stability, albeit modestly, suggests that the chemical toolkit of amino acids exerts an influence

on the acquisition of tertiary ‘foldability.’ Yet even the presence of cysteine did not vault these proteins into the realm of robust, modern-style tertiary folding. Instead, it imparted an incremental benefit that allowed them to maintain partial structure under a slightly broader range of solvent conditions. While it is plausible that thiol-bearing proteins lacking stable tertiary structure are capable of some level of protometabolic function, it suggests that the early evolutionary path toward fully folded proteins was neither rapid nor all-encompassing. Instead, it likely involved a series of small chemical and structural steps, each one slightly extending the range of conditions under which partial folds could persist.

The data appears to a relatively broad protein folding window: neutral to mildly alkaline pH levels (pH 6-10) that stabilise deprotonated functional groups, moderate ionic strengths (0-1000mM), and a comparatively broad temperature range up to 90°C, with thermal heating seen to increase disorder but not precipitate the majority of tested proteins. These results align with Hypothesis 2 (abiotic proteins will not fold at acidic pH) and Hypothesis 3 (abiotic proteins adopt molten globule states), and also support predictions made by Di Mauro *et al* that “proteins made from just the earlier amino acids can likely form molten globules but not tightly packed structures” (Di Mauro *et al.*, 2012). Additionally, the general insensitivity to moderate changes in ionic strength supports Hypothesis 4, indicating that salinity alone does not fully disrupt the molten globular structure. Chapter 3 thus provides evidence for the required solvent conditions and structural behaviours of abiotic-like proteins as required by Aims 1 and 2. While this study is the first to systematically test three 95% abiotic proteins, further work

is needed to confirm this folding window, particularly in relation to the protein to the limited number of proteins tested within this study.

Study 3

Chapter 4 shifted attention to the processes of peptide synthesis, focusing on the potential role of cysteine. Here I focused on peptide bond formation under conditions reflective of an alkaline hydrothermal vent based on our observations in Chapter 2 and 3 that such conditions would support abiotic protein function, and that these conditions are highly conducive to peptide bond synthesis. These experiments built on the work of Vallee and Foden *et al* by integrating cysteine peptide synthesis experiments with silicate surfaces to make the experiments more representative of environments where life is believed to have emerged (Vallee *et al.*, 2017)(Foden *et al.*, 2020)(Martin and Russell, 2003)(Takahagi *et al.*, 2019).

Introducing cysteine into an otherwise standard set of abiotic amino acids led to a dramatic increase in peptide yields in the presence of silicate minerals. Moreover, once these cysteine, or related thiols such as homocysteine, containing peptides were formed, their thiol group enabled metal ion coordination and ligation, enabling the synthesis of longer, more compositionally diverse peptide chains which is likely to be a process specific to peptides enriched in thiols under moderate salinities. Notably this process had a strict reducing environment requirement of at or below the redox potential where cysteine is oxidised to cystine of -220mV (Jocelyn, 1967). While this is mildly reducing, it is less reduced than the anoxic, H_2 enriched conditions expected on early Earth (Catling and Zahnle, 2020).

The earlier chapters had shown how challenging it was for abiotic proteins alone to yield stable or functionally relevant structures. While cysteine is never a critical amino acid for protein folding in modern proteins it can stabilise structures through disulfide bonding and metal-binding (Pace *et al.*, 2011). Consistent with this modern observation, cysteine only offered mild stabilisation in the results of Chapter 3 yet in Chapter 4, its presence has a profound impact on functional potential. As a result, what might be random peptide assembly under abiotic conditions becomes more productive, more directional, and more conducive to generating proteins potentially capable of rudimentary catalytic functions when cysteine is present, While proteins assembled through TSPL will always be uncoded yet compositionally biased depending on amino acid availability and reactivity, its potential selectivity for specific cysteine motifs, and the near guaranteed transition metal binding capacity of the long chain peptides and proteins generated makes it much more likely that function could emerge spontaneously compared to synthesis pathways without cysteine.

These experiments support the predictions of Hypothesis 5 and 6 that warm, alkaline, and reducing conditions, and the presence of thiols, constitute a feasible environment for abiotic protein folding and peptide assembly.

Summary

From the perspective of planetary habitability, these findings indicate that environments capable of supporting thiol-containing amino acid synthesis are well suited for both the synthesis and function of abiotic proteins. This “Thiolpeptide Goldilocks zone,” defined by high dissolved sulfur content, neutral to mildly alkaline pH, reducing redox potential (≤-220 mV), elevated temperature

(60-120°C), and moderate salinity (0-1000 mM), supports peptide synthesis, ligation, metal coordination, and stabilisation of molten globular protein structures (**Table 5.01**). Under these specific conditions, cysteine plays a critical functional role in both promoting peptide assembly and enabling redox-active metal centres, features likely essential for the emergence of primitive metabolic networks.

However, it is important to recognise that this scenario represents only one possible window for abiotic peptide and protein evolution. Earlier chapters in this thesis demonstrated that peptides rich in aspartate or other carboxylate residues, the presence of which observably increased peptide yield compared to alanine, may also support peptide metal binding and ligation under a wider redox range even in the absence of thiols.

While these results demonstrate the significance of sulfur-rich, reducing, and alkaline conditions in facilitating both peptide synthesis and the emergence of redox-capable peptide chemistry via cysteine, they do not exclude the possibility that other combinations of aspartate dominated assemblies could also support abiotic protein assembly and function, albeit with a much less clear route to catalytic redox activity than thiol-based chemistries. The diversity of plausible prebiotic settings must be considered in any search for the chemical origins of protein evolution.

Together, this data suggests that eleven amino acids, namely the ten-consensus abiotic amino acids plus cysteine, form a “functional set” required for peptide synthesis, peptide ligation and proto-metalloprotein formation and function of thiol bearing abiotic peptides and proteins. An abiotic mechanism

giving rise to redox-active centres appears much more difficult in the absence of cysteine (Shalayel *et al.*, 2020).

	Cysteine synthesis window	Peptide Synthesis Window	Protein Folding Window	Thiopeptide 'Goldilocks' Protein Function Window	Archean Earth (range)	Noachian Mars (range)	Enceladus (range)
pH	≈pH 7-11	9.5	6-10	7-9.5	6-11	5-6.5	8.5-10.5
Temperature (°C)	60-115 (NaSH boiling temperature)	60 – 120 *(cyclisation becomes favourable)*	20 - 90	60 - 90	40 – 90 *(Ocean-White Smoker)*	-20 - 20	0-90 *(Ocean-H. Vents)*
Salinity (mM)	Not reported	450	0-1000	0-1000	0.5 *(ocean)*	0.5-3	0.1
redox potential (mV)	Not reported	≤-220	≤-220 *(if metal guided)*	≤-220	≈-330	0-300	≈-350
Transition metal availability	High *(30mM)*	Medium *(0.13mM)*	High	High	High	Poor	Poor

Table 5.01 – The thioprotein function 'goldilocks' zone

The proposed thioprotein goldilocks functional zone that provides conditions compatible with the synthesis, polymerisation and metal binding activity of thiopeptides. Cells highlighted green showing alignment between widespread predicted planetary conditions for locations of interest against this standard are shown. Red is used to indicate poor alignment, and amber indicates an overlap for relatively restricted areas of that planet. Data ranges for each environment are based on values presented in ***Table 1.01***

Habitability of plausible abiotic solvents

Earth

Given the antiquity of life in Earth, local environmental niches and/or bulk solvent conditions must have permitted the emergence of functional proto-metalloproteins (Catling and Zahnle, 2020). Although bulk oceanic pH is predicted to have been mildly acidic to neutral due to the elevated CO_2 content of the atmosphere, localised environments, such as alkaline hydrothermal vents, would have provided both solvent and mineral conditions suitable for cysteine catalysed peptide synthesis (Martin *et al.*, 2008)(Krissansen-Totton Joshua *et al.*, 2022). Here, the interfaces between the alkaline and acidic conditions could have provided mixing conditions for the delivery of dissolved transition metals that could be acquired by thiol peptides to bring about selective ligation via TSPL (Grotzinger and Kasting, 1993). Such conditions for synthesis dovetail with the requirements previously identified in Chapters 2 and 3 for abiotic protein folding. Thus, sulfur-rich niches at these vents were likely capable of producing cysteine during hydrothermal circulation, and this could have been central to the genesis of proto-metalloproteins, as supported by the necessity of thiol-bearing amino acids for stable metal coordination both monomerically and within a peptide (Foden *et al.*, 2020)(Jordan *et al.*, 2021). The alignment between the conditions used in these experiments with the Archean Earth, suggests that this early terrestrial setting presented a uniquely promising backdrop for the evolution of functional peptides and proteins (National Research Council, 2007).

In comparison to previous origin of life studies on Earth, the data presented here support existing conclusions, which have proposed that abiotic peptides and proteins likely provide a suitable architecture to enable abiogenesis on Earth (Tagami and Li, 2023)(Maury, 2018). My findings, that these proteins do not explicitly require high salinity or need to be amyloidal to form potentially functional structures expands the sites of plausible abiotic protein function away from hypersaline environments as previously suggested (Longo and Blaber, 2014) . In addition, my propensity scaling data grounds previous predictions of abiotic protein structures to the observed abiotic amino acid radiometric abundances to reveal that these proteins had limited secondary with a preference to helical rather than β-structures and were molten globular and acid sensitive. In terms of assessing planetary habitability, this supports previous work suggesting acidic ocean interfacing alkaline hydrothermal vents provide suitable conditions for abiogenesis (Barge *et al.*, 2019)(Martin and Russell, 2006).

Mars

The conditions of early Mars would be less favourable for the synthesis of metal-binding thiol-bearing peptides. Modelled as more acidic, increasingly oxidising, and significantly colder that Earth, the ancient Martian oceans, and lakes appear likely to have introduced substantial hurdles to the formation of stable, abiotic proteins (Fairén *et al.*, 2004). The inherently lower solubility of iron as Fe^{3+}, along with the scarcity of reducing conditions, would have limited the availability of transition metal ions to support metal-guided folding (Liu, J. *et al.*, 2024). Meanwhile, progressively more oxidising conditions on the surface as Mars evolved may deny the advantages of thiol-bearing amino acids to the

synthesis and function of proto-metalloproteins and peptides. Additionally, cooler average temperatures would have hindered peptide synthesis, which requires temperatures in excess of 65°C (Shock, 1992), While brief, localised hydrothermal systems on Mars might momentarily have offered more supportive conditions, the global trends do not match the chemical and physical criteria that favoured the proposed cysteine-driven complexity on Earth (Ojha *et al.*, 2021).

While alkaline hydrothermal systems have been proposed on early Mars, direct evidence supporting their existence is limited since serpentine deposits, which form from alkaline hydrothermal activity, are missing (Ehlmann *et al.*, 2010). However, based on the data from Mars so far, I cannot rule out that subsurface conditions, or alkaline hydrothermal vents existed on Mars interfacing with an acidic ocean. This would establish transient reducing environments that may be much more favourable to organic accumulation. However, I do propose that on the whole, early Mars, lacking strong evidence for alkaline hydrothermal vents and an ever-increasing salinity and redox potential and was less poised to progress from short functionless abiotic peptides towards thiol-peptide-based chemistries.

Comparing my results with other contemporary studies, my results are broadly consistent with previous descriptions of Mars, which is often characterised as poorly suited to the genesis or survival of life. Indeed, with over 6 decades of exploration, no biosignatures have been acquired on Mars. It has been proposed that one of the primary drivers of this was a high proportion of iron in the mantle of Mars that prevented a functioning planetary water cycle (Tian *et al.*, 2021). This subsequently limited any long-term standing water from forming,

and thus prevented the chemistry needed for life from emerging (Wade *et al.*, 2021). While the consistency between my findings comparing Mars' conditions against those needed for abiotic proteins is interesting it is important to acknowledge that a number of geological niches may exist on Mars that do not align with the conditions I have compared against here. While no single study could ever capture and compare against the environmental diversity across a planet as large as Mars, my data does at minimum suggest that it appears less compatible with functional protein synthesis accumulation than other sites within the solar system that may be better suited to further exploration.

Enceladus

In contrast, Enceladus emerges as perhaps a more compelling candidate for abiogenesis. The Cassini mission and subsequent analyses characterised Enceladus's subsurface ocean as alkaline, reducing, and moderate in salinity (Glein and Waite, 2020). These environments parallel the conditions shown here for cysteine-mediated peptide synthesis and thiol-peptide metal coordination and ligation. Two challenges to this however are the weak evidence for sulfur detections in the Enceladian ejecta and the low availability of transition metals compared to early Earth due to the bulk alkaline ocean where iron, cobalt, and zinc are poorly soluble (Peter *et al.*, 2024)(Hao *et al.*, 2022). However, I note that I observed that the Arc4 peptide readily acquired Fe^{3+} and Co^{2+} under alkaline conditions, suggesting that the strong metal binding capability of cysteine bearing peptides can retain metal ions in solution even in otherwise precipitating conditions. Given the parallels of this environment with early Earth, and assuming that cysteine peptides can successfully acquire and solubilise iron and zinc,

Enceladus stands out as an extraterrestrial location where the processes described in previous chapters might reasonably occur or at least have once occurred. This conclusion would elevate its planetary habitability status as a location more suitable for abiotic peptide synthesis and function compared to early Mars, although less so than early Earth. Given the recent discovery of phosphates on Enceladus, these results establish a need for further investigation at this site, particularly looking for evidence of sulfur bearing amino acids (Hao *et al.*, 2022).

In the context of previous investigations of Enceladus, my results suggest that the alkaline hydrothermal conditions reported there are compatible with the cysteine catalysed processes reported here. While this supports recent assessments of Enceladus habitability, I note that some recent reports may have overstated its habitability potential without addressing the critical gaps in our understanding of its organic inventory or its solvent niches (Tenelanda-Osorio *et al.*, 2021)(Hendrix and House, 2023). As such, it is important to add a caveat to my assessment of what cannot be declared about the habitability of Enceladus given the *in-situ* data available.

As of the end of 2024, the data on Enceladus Ocean confirms that it contains an alkaline carbonate-rich ocean and survivable conditions for modern methanogens, and it possesses the same chemical ingredients (CHNOPS) found in terrestrial life but is heavily depleted in soluble transition metals due to the oceans alkalinity (Khawaja *et al.*, 2019)(Taubner *et al.*, 2018)(Glein and Waite, 2020)(Hao *et al.*, 2022). While the presence of CHNOPS elements and hydrothermal activity are indicators of potential habitability, this alone does not

establish the conditions necessary for abiogenesis (Cleaves and Chalmers, 2004).

The conditions reported for Enceladus are thus far short from confirmation that Enceladus has the full required library of organic species and conditions to enable the processes needed for abiogenesis to proceed from monomeric organics to a complex information-transfer capable system of life dependent on abiotically derived polymers. Indeed, many organic compounds that drive many amino acid reaction schemes, such as HCN, is reportedly very unlikely to form under Enceladus' current conditions (Hao *et al.*, 2022). As such, and as noted by Cable *et al,* it is difficult to truly describe Enceladus as habitable without further characterisation and without at least partially addressing how the conditions on Enceladus could enable the synthesis and accumulation of ribonucleotides and RNA, which degrades 5600x faster under Enceladus' conditions (Cable *et al.*, 2020). This challenge for alkaline degradation of RNA is less applicable for the Archean Earth due to the acidic ocean/alkaline hydrothermal mixing systems which provide interfaces where both RNA and peptides can be synthesised and delivered to temperate niches to enable abiogenesis although within a relatively small radius of their synthesis locations (Barge *et al.*, 2019). To fully address this question, stronger constraints are needed for the Enceladus' Ocean composition, and its potential for hosting microenvironments where pH/eH gradients may be more favourable than the bulk ocean to drive accumulation and diversification of an organic inventory. Realisation of this will almost certainly require another exploration mission to Enceladus within the next decades.

Comparison with recent works

Top-down vs bottom-up approaches to assessing planetary habitability.

The majority of recent works investigating the habitability of Mars and Enceladus have focused on the presence of conditions that could support life, whether in the terms of possessing exploitable redox gradients to support methanogenic life or conditions that support life's survival (a top-down approach) (Taubner *et al.*, 2018)(Sauterey *et al.*, 2022)(Cockell, 2014). This approach assumes that compatible conditions for life to survive makes it more likely that an organism will emerge there and can be thought of as a 'top-down' approach to habitability. In essence, if modern life occupies conditions thought to be common at the dawn of life, then life must have occurred (initiated) in these environments. The wide tolerances of Earth's extremophiles are often cited as supporting evidence in this research. However, it is important to recognise that extremophiles have adapted to survive in extreme environments rather than spontaneously emerged with a capacity to thrive in extreme conditions (Merino *et al.*, 2019)(Reed *et al.*, 2013)(Coker, 2019). This assessment is supported by investigations of acidophiles, alkaliphiles and halophiles, which, in nearly all cases, maintain a mesophilic-like cytoplasmic pH, eH and water activities. Indeed, the remarkable feature of extremophiles is *how* they overcome hostile exterior conditions to maintain a functionally mesophilic-like intracellular environment.

While it is entirely possible early life likely had some preference to one condition or another, indeed my own data points to conditions favouring modern alkaliphiles, it is unlikely that life can spontaneously emerge with resistances to

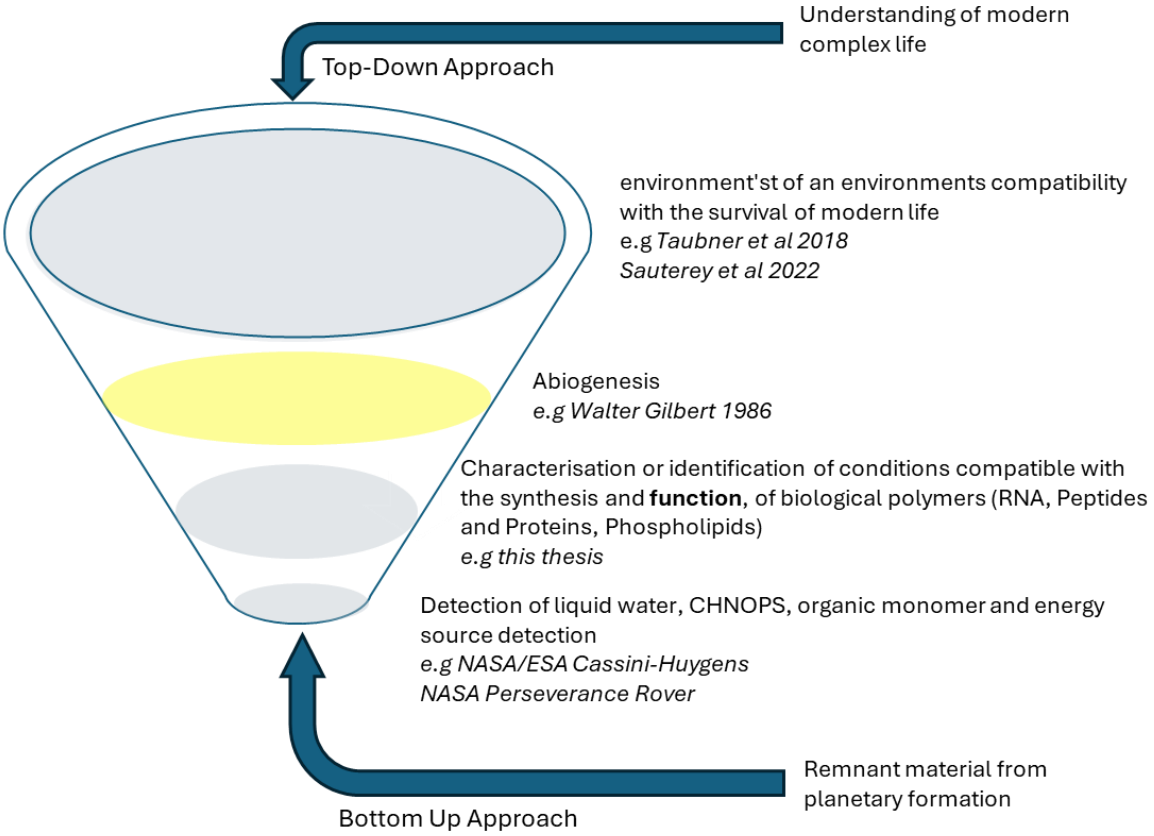


Figure 5.01 – Conceptual approaches for assessing the habitability of an aqueous environment.

highly extreme conditions beyond those discussed here (Cleaves and Chalmers, 2004). These considerations make these types of top-down approaches to assessing habitability inherently less robust, and arguably weaker than other methodologies. Retorts against this counterargument typically focus on theories such as panspermia, where life is proposed to have emerged on one planet, which subsequently seeded another, therefor making the assessment of survival conditions relevant. However, the theory of panspermia is almost entirely speculative and untestable.

The fundamental issue with the top-down approaches and indeed approaches simply looking for an environment's possession of CHNOPS material is that there is a vast jump in complexity from extremophiles and simple molecules to abiogenesis. The work of this thesis takes a different approach by asking if these locations can support the emergence and foldability of abiotic proteins, which many have suggested to be the progenitors for enabling the dawn of the first information transfer pathways that enable life to begin (**Fig 5.01**) (Maury, 2018)(Piette and Heddle, 2020)(van der Gulik and Speijer, 2015). I argue that this style of bottom-up approach to assessing planetary habitably demonstrated here is a stronger framework to assessing habitability than top-down approaches, which are more likely to result in false positive claims of survivable niches being habitable but where the genesis of life cannot proceed (Cleaves and Chalmers, 2004).

Conclusion

In summary, the experimental and theoretical frameworks developed in this thesis provide a novel qualitative yardstick against which to measure the

potential of different planetary bodies to support the earliest stages of life's protein architectures. Archean Earth conditions appear well aligned with the observed requirements for abiotic protein synthesis and folding, while early Mars, with its harsher, low temperature, oxidising, and acidic conditions, seems less supportive. Enceladus, which mirrors some of the parameters of early Earth's alkaline hydrothermal vents, emerges as a promising candidate for further exploration (**Table 5.01**). The approach used here contributes to existing methodologies and provides supporting evidence for the potential habitability of Enceladus, although I note that significant further work is needed before Mars or Enceladus can be conclusively declared habitable or not.

A central implication of these findings is that *in situ* measurements of local solvent parameters, namely pH, redox states, temperature profiles, and mineral compositions, are essential to interpret the functional potential of any organic molecules identified beyond Earth. Without this contextual data, even the discovery of amino acids or short peptides remains ambiguous with respect to their potential for function. Identifying alkaline, sulfur rich and reducing niches, for instance, would strengthen the case that complex, cysteine-enriched peptides could form and begin to potentially enable prebiotic enzymatic capabilities.

In the broader perspective, the refined understanding of abiotic protein function provides a functional window that can now be compared against the conditions reported in upcoming missions to Europa and Enceladus. These conditions also serve to enable better targeting of where and how to search for evidence of life's earliest molecular steps as future programs to Mars and Venus progress through the next century.

Chapter 6: Synthesis and future directions

This thesis has provided evidence that suggests that abiotic protein folding, and function are fundamentally constrained by specific environmental conditions. The presence of cysteine and transition metals in alkaline, reducing environments may have been a key determinant of early biochemical evolution. While also suggested by previous works, the novel lens through which this was determined within a 'bottom up' framework adds further evidence to the suitability of these niches to support abiogenesis (Martin and Russell, 2006). This work builds heavily on the results of Foden and Vallee *et al,* by proposing an updated set of abiotic amino acids to now include cysteine, which I have dubbed the 'abiotic functional amino acid set' (Foden *et al.*, 2020)(Vallee *et al.*, 2017). In addition, I provide additional evidence to support alkaline hydrothermal conditions as being best suited to support the full cascade of protein synthesis, folding and function needed to enable an RNA world (Martin and Russell, 2006)(Goldford *et al.*, 2019)(Fujikawa *et al.*, 2024)(Moody *et al.*, 2024). These findings challenge the assumption that early peptides lacked functional potential in the absence of genetic encoding (Bernhardt, 2012). While previous work suggested that abiotic peptides were functionally limited due to sequence disorder, this study demonstrates that metal coordination can enable folding even in partially sequence-agnostic peptides (Di Mauro *et al.*, 2012). This supports a 'peptide-RNA model' in which peptides played a major role in establishing the conditions needed for a later RNA world (van der Gulik and Speijer, 2015).

Enabling the genesis of information transfer.

Taken together, my results suggest that alkaline hydrothermal conditions are restricted to yielding short peptides and without a feasible route to

functionality. But in the presence of sulfur, these can then significantly accelerate stereoselective peptide synthesis to create thiol peptides (Foden *et al.*, 2020). Subsequently, thiol-bearing peptides have the capacity to coordinate and ligate around metal ions (Shalayel *et al.*, 2020). Given enough cycles, this process could conceivably create proto-functional thiol metalloproteins. These metalloproteins facilitate transition metal coordination, creating nucleation sites to provide more structure to molten globular abiotic proteins and enable redox reactions to occur. Although each protein sequence will differ, given enough peptides and sequence diversity it is likely that some of these will have some activity, particularly in the presence of transition metals.

it may be possible, that these abiotic thiol-bearing, molten globular proteins could enable RNA nucleotide synthesis to bring about RNA world (**Fig 6.01**). While I accept that this theory is speculative, this work provides stronger support for the possibility of generating semi-functional proteins is more likely if cysteine were present. Future work is obviously required, particularly on the latter stages of this pathway to convert this from speculation into a testable hypothesis, but I hope that this thesis provides a solid foundation for investigating both the importance of cysteine facilitated peptide synthesis and the origins of life's stereoselective bias.

While the conditions needed for the cysteine pathway may initially appear to be contradictory to those required for the long-term stability of RNA in an 'RNA world', it has been noted that no single location is likely to possess all the environmental requirements needed for abiogenesis to occur (Stüeken *et al.*, 2013). Indeed, this very reason is why hydrothermal systems are so often

favoured as sites of abiogenesis; they provide multiple discrete solvent niches for different reactions to occur and for subsequent reactants to combine within mixing zones in the system (Martin and Russell, 2003). It is easy to imagine a sulfur-rich hydrothermal vent outflow system whereby thiol-proteins are synthesised on the silicate rich chimney walls and then delivered into a more acidic, or temperature crevice or nodule at or near the acidic seawater interface whereby RNA reactions can proceed, catalysed by thiol-proteins. While this type of discussion may always be speculative in regard to how life formed on Earth, refining conditions where such reactions are possible will enable better targeting of future exploration missions of other extraterrestrial bodies.

Future planetary missions should prioritize environments with high soluble Fe^{2+} and HS^- availability, as these conditions most strongly correlate with cysteine-driven peptide formation. If abiotic metallopeptides played a role in early catalysis, biosignature searches should be extended to include organometallic complexes and short thiol-containing peptides.

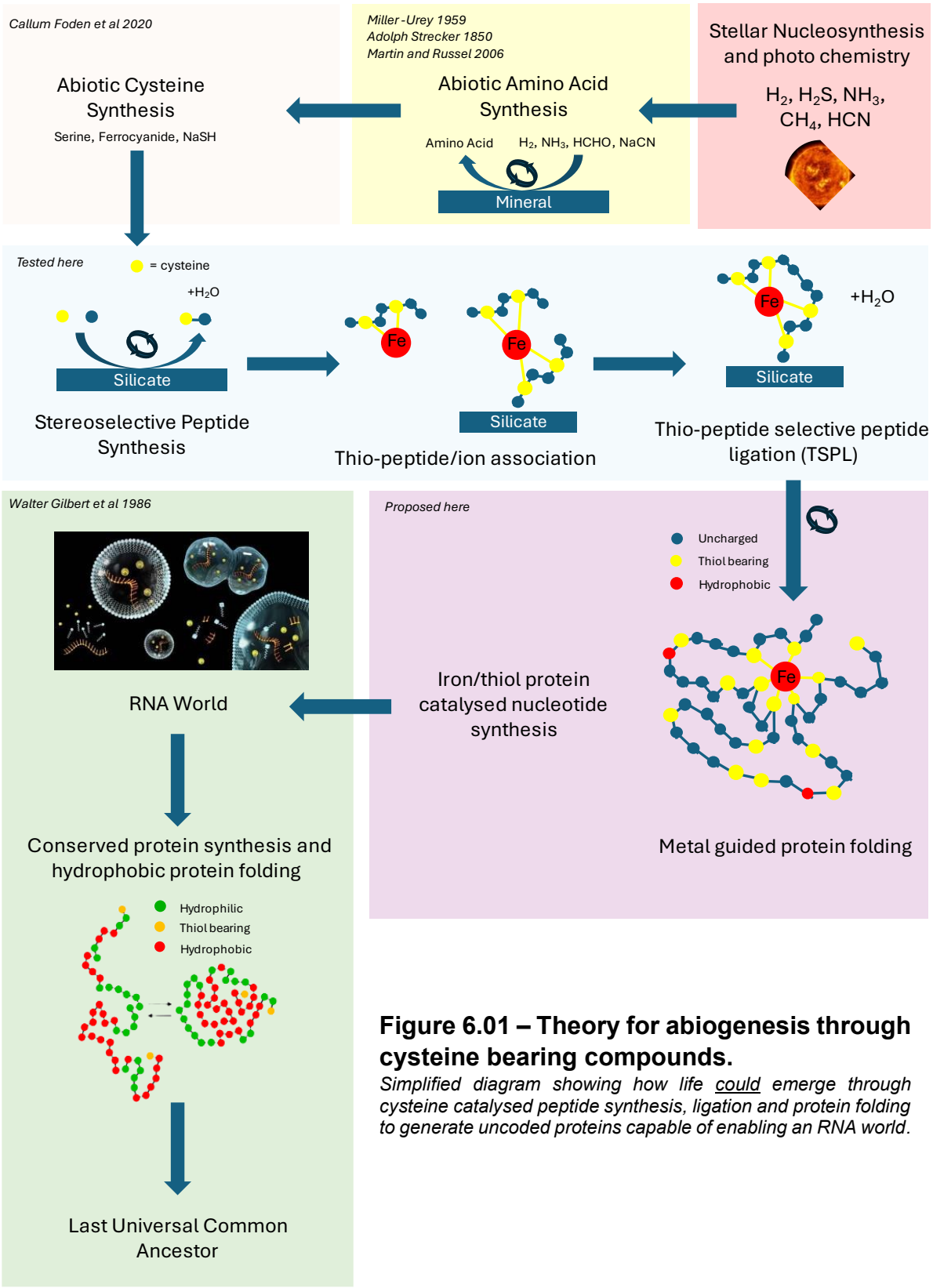


Figure 6.01 – Theory for abiogenesis through cysteine bearing compounds.

Simplified diagram showing how life could emerge through cysteine catalysed peptide synthesis, ligation and protein folding to generate uncoded proteins capable of enabling an RNA world.

Future directions

The results of the thesis set up several new experimental campaigns that could test my predictions on cysteine enabled abiogenesis. Firstly, the scale of cysteine-driven peptide elongation and the breadth of peptide sequences capable of engaging in this pathway is only partly explored here. More diverse tests of cysteine-containing peptides, or ones containing related thiols such as homocysteine, especially those that vary in length or distribution of cysteines, are needed to see how effectively they can engage in metal binding and self-ligation. In addition, further investigation into the ability of L-stereoselectivity to propagate from sequences enriched in L cysteine will better identify if this process could have constituted the first selection pressure, potentially of many, towards homochirality.

Secondly, investigating if the thiol-peptide synthesised are capable of catalytic function will identify whether molten globule abiotic proteins generated from these are capable of catalysing bond-forming events needed for the formation of proto-metabolisms or nucleotide synthesis. Introducing small pools of nucleotide precursors to mineral-based ligation experiments may also reveal whether thiol-enriched, metal-coordinating peptides can indeed protect or engage with these precursors in ways conducive to RNA synthesis.

Thirdly, the theoretical pathway for metal guide protein folding of abiotic proteins was not directly testable in this work. Future experiments using *de novo* designed protein sequences based on the high cysteine incorporation ratios predicted here could design and express an array of cysteine-enriched proteins to identify if metal folding can drive protein folding in abiotic proteins.

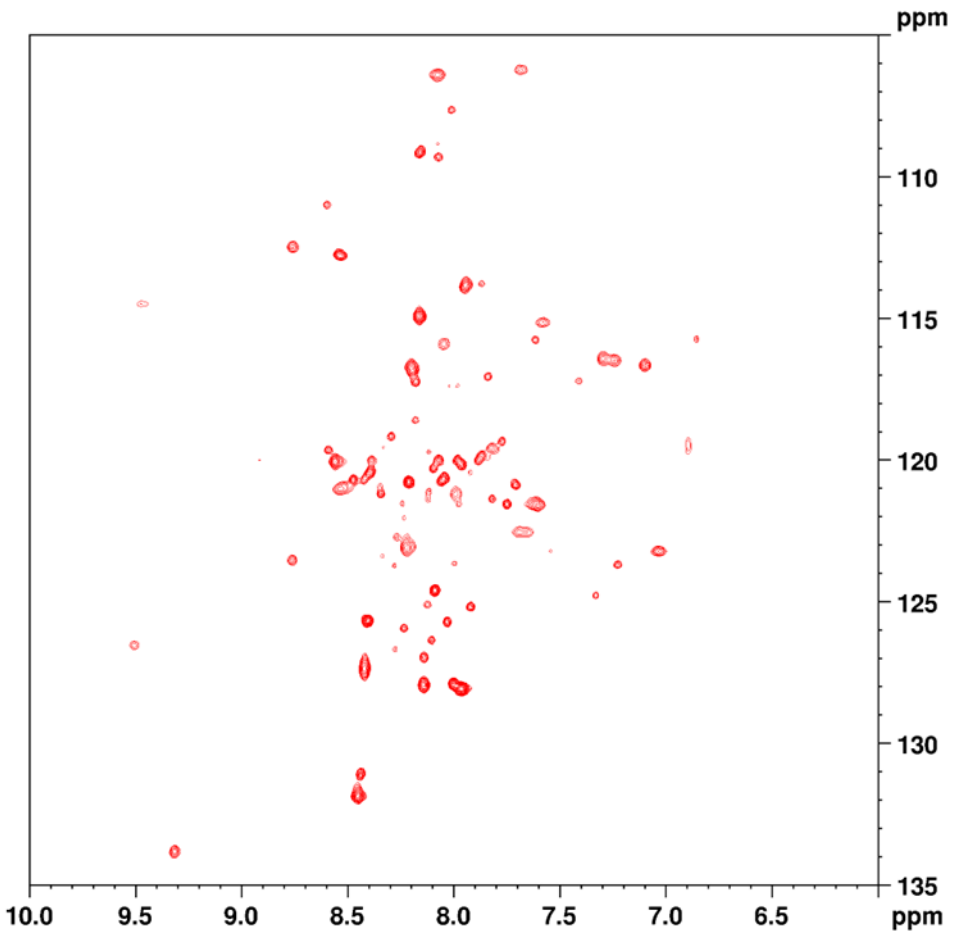
Lastly, future synthesis experiments could use more realistic hydrothermal conditions to investigate how environmental cycling can influence peptide synthesis. Shifting pH or salinity levels, as well as transient hot-cold or acidic/basic cycles, might diversify the types and lengths of peptide synthesised or ligated, enabling more routes for function to emerge. In addition, my conclusions here rest heavily on the possibility of abiotic cysteine synthesis. While the conditions reported by Foden *et al.* are applicable to alkaline hydrothermal vents, direct testing of this synthesis in hydrothermal model conditions would add confidence to this read across (Foden *et al.*, 2020).

Conclusion

In summary, the work presented in this thesis supports the central hypothesis that stable, functionally competent proteins did not arise fully formed but emerged gradually and were enabled by particular planetary solvent conditions. By identifying the potential role of cysteine and defining the environmental parameters necessary for it to deliver rate enhancing effect on peptide synthesis and enable metal coordination, my experiments have enhanced the planetary habitability assessment toolkit for assessing the possibility of life on other worlds. As exploration missions probe deeper into the solar system's icy oceans, the criteria established here can guide future interpretations of habitability in a way that is grounded in protein biochemical realities, which undoubtedly would have played key roles in supporting the genesis of life on Earth, and, perhaps, elsewhere.

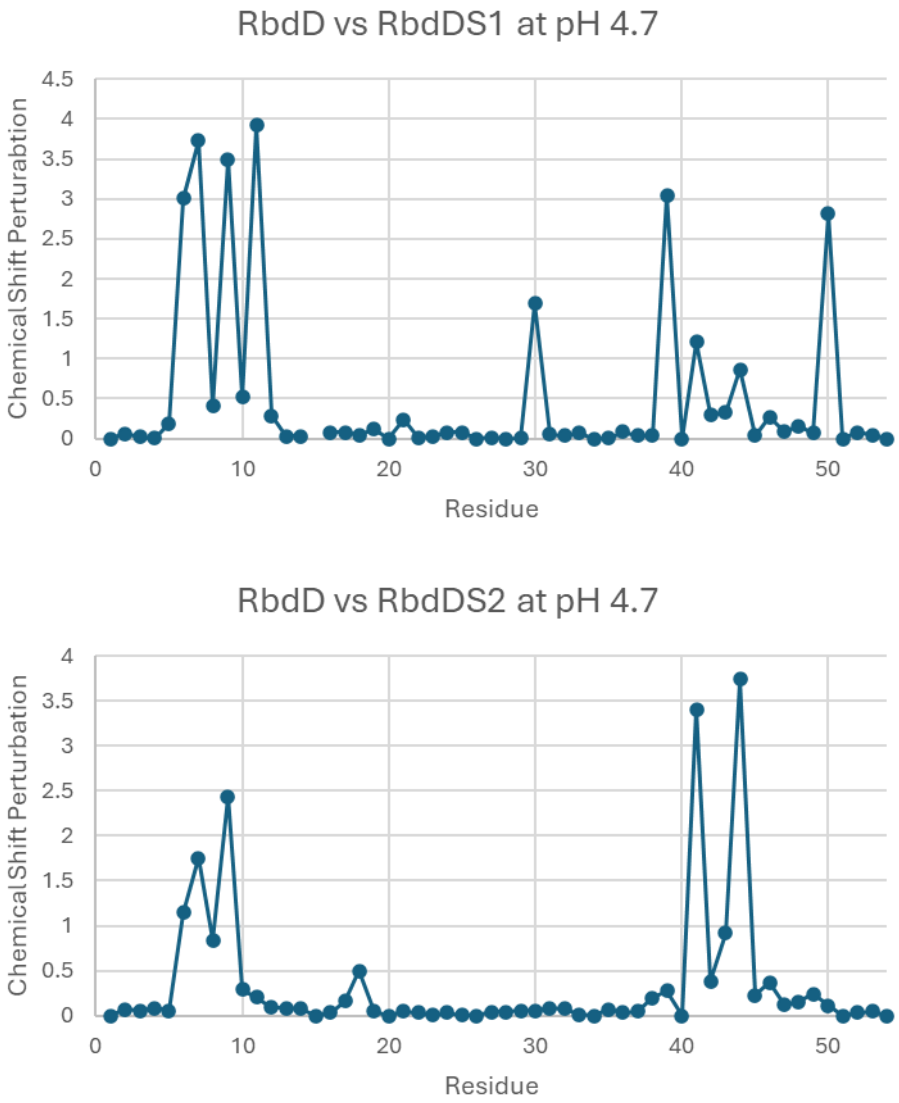
Appendices

Appendix A – Figures



Appendix A - Figure 01 – Metal addition at pH 6.2 does not rescue RbdD folding.

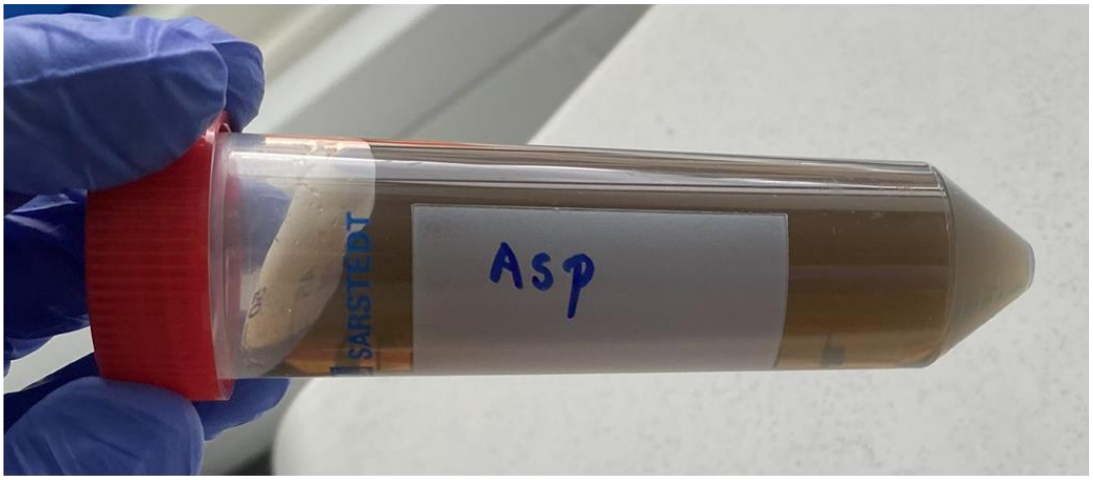
2D ^{1}H-^{15}N HSQC spectrum of 1mM RbdD with 1mM ZnCl2 at pH 6.4 in unbuffered water. Data was acquired with 32 scans per increment, 2048 total points (t_1), and 15.95ppm sweep width in the direct (^{1}H) dimension and 256 increments with a sweep width of 32.07ppm in the indirect (^{15}N) dimension. Metal addition of ZnCl2 did not result in significant changes in observed peak resonances that would have been indicative of protein folding.



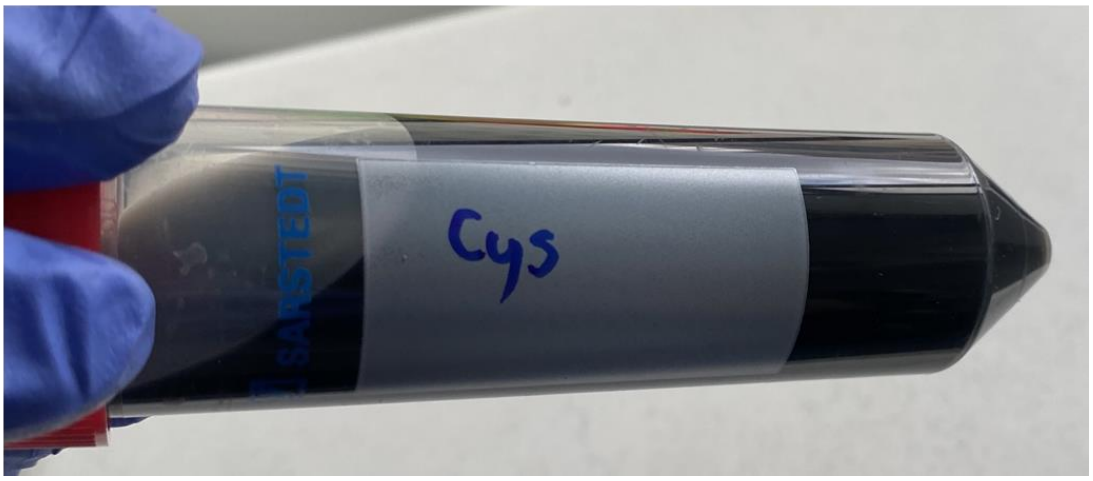
Appendix A - Figure 02 – RbdDS1 and RbdDS2 aspartates remain deprotonated at pH 4.7.

*2 D ^{1}H-^{15}N HSQC spectrum of 635μM RbdD, and 1004μM RbdDS1 (a) and (**b**) 1812μM RbdDS1, equilibrated in 100% N_2 overnight, were recorded at 20°C with a final pH of 4.7±0.15 in unbuffered water. Data was acquired with 32 scans per increment, 2048 total points (t_1), and 15.95ppm sweep width in the direct (^{1}H) dimension and 256 increments with a sweep width of 32.07ppm in the indirect (^{15}N) dimension. Chemical shift perturbations were calculated from the formula proposed by* (Williamson, 2013) *and plotted in Excel.*

(a)

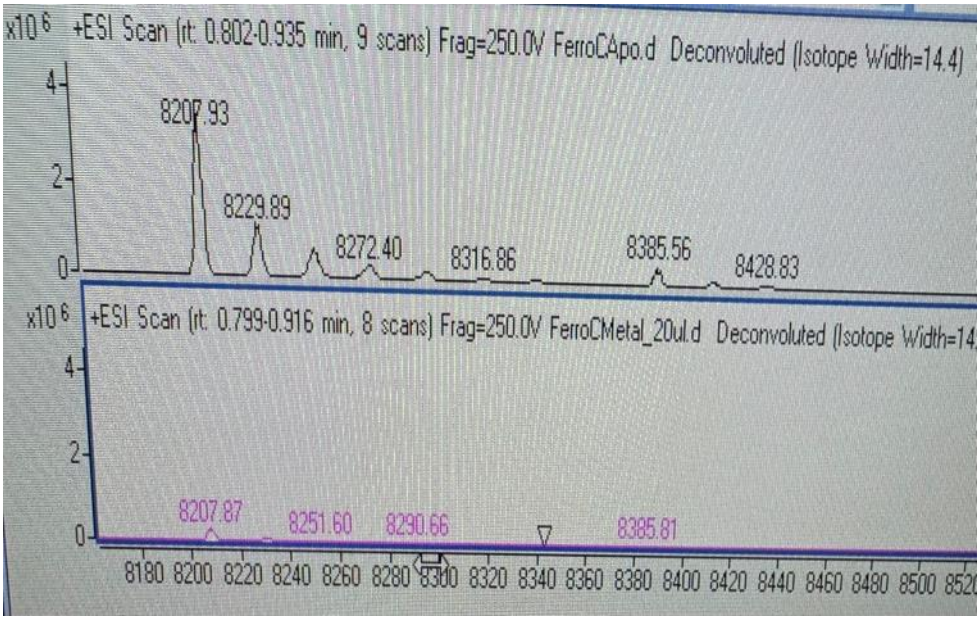


(b)



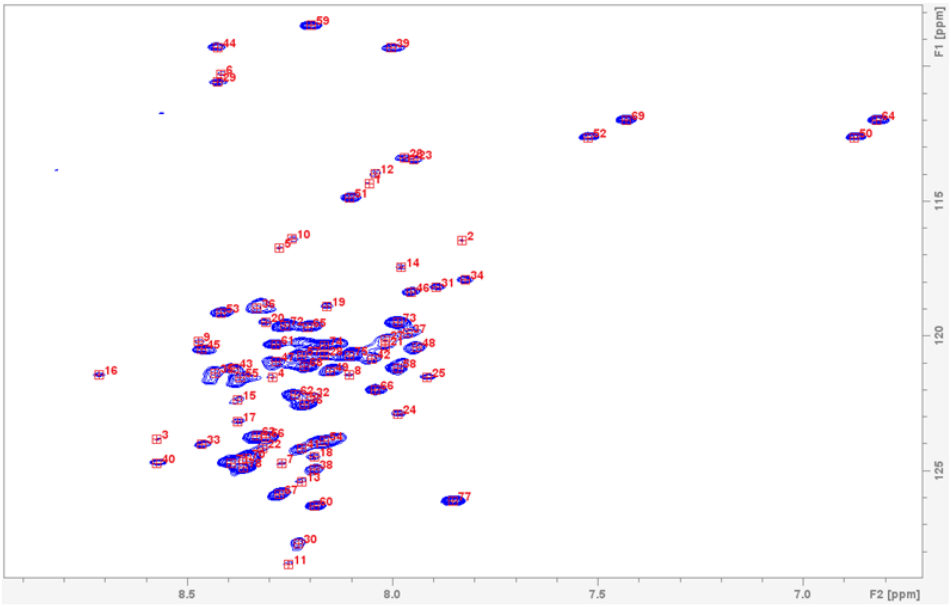
Appendix A - Figure 03 – Aspartate destabilises soluble FeS clusters.

*The methodology reported by Jordan et al 2021 was replicated using 5mM L-aspartate (**a**) or L-cysteine (**b**), 1 mM $FeCl_3$ and 1 mM Na_2S at pH 9 in 0.1M bicarbonate buffer in an anoxic hood (5ppm O_2). Successful creation of iron sulfur clusters was monitored visually for synthesis in the presence of aspartate or cysteine, with precipitation of iron sulfide observed in the presence of L aspartate meanwhile cysteine solutions retained a black colouration with no precipitation observed in agreement with Jordan et al.*



Appendix A - Figure 04 – Intact mass of ferredoxin

Successful purification of Ferredoxin was confirmed by intact qTOF mass spectroscopy with identification of a peak at 8207.93 compared to a predicted mass of 8207.90.

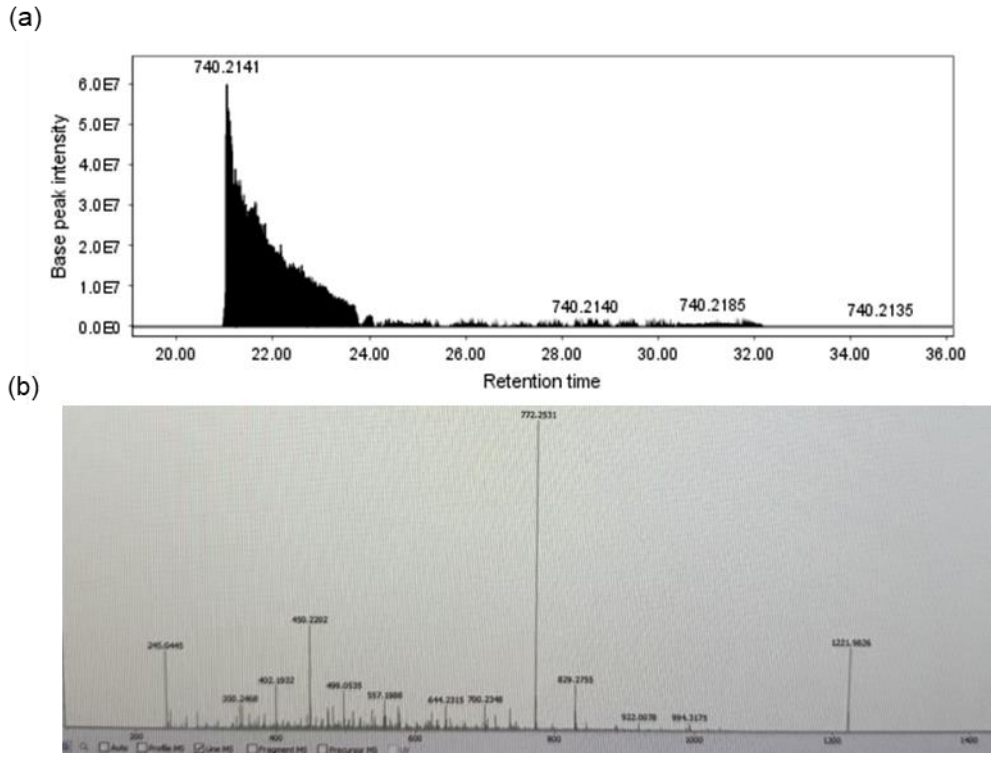


^{15}N

^{1}H

Appendix A - Figure 05 – Purified ferredoxin purifies as an apo-protein

Automated peak picking of ^{13}C-^{15}N labelled ferredoxin was performed using the Topspin automated peak picking tool. 78 peaks were identified against 79 predicted peaks (including the Ferredoxin Histag) Data was recorded at 20°C and pH 6.0 in 0.1M phosphate buffer with 32 scans per increment, 1276 total points (t_1), and 16.03ppm sweep width in the direct (^{1}H) dimension and 320 increments with a sweep width of 30.00ppm in the indirect (^{15}N) dimension.



Appendix A - Figure 06 – Thiol peptides oxidise under heating at 90℃ under non reducing conditions.

*The Arc3 peptide (740.216) was detected by extracted ion chromatogram following TIMS TOF MS after heating at 90°C and pH 9.5 for 36 hours in the presence of $ZnCl_2$ both with (**a**) and without (**b**) sparging in H_2. Following heating, peptides were lyophilized, resuspended in 0.1% formic acid, and injected onto the spectrometer. Failure to sparge in H_2 however instead generated a mass detection 772.253, consistent with the complete oxidation of two cysteines to cysteine-sulfenic acid. No ligation products were identified in the oxidised samples but were identified in the reduced samples.*

Appendix B – Tables

Cys -																										
Accession	Amino Acids	Mass	Organism	Primary Structural Feature	Protein Name	Avg pLDDT	Relative Abiotic Score	Isolated B bridge	B ladder	3.10 Helix	A Helix	Pi Helix	Hydrogen bonded turn	Bend	Un-structured	Tandem Repeat Percentage	A	G	V	P	L	I	D	E	S	T	Relative Complexity
A0A059FV71	158	14821	Hyphomonas johnsonii	Helical	Lipoprotein	70.98	0.10	0.00	0.00	0.00	73.30	0.00	2.84	0.00	23.86	16.46	0.47	0.01	0.08	0.01	0.03	0.01	0.09	0.18	0.04	0.04	10
A0A077WY34	119	11376	Lichtheimia ramosa.	Helical	ABC Transmembrane type 1 domain containing protein	81.96	0.12	0.00	0.00	0.00	85.40	0.00	0.00	0.00	14.60	16.81	0.31	0.00	0.24	0.00	0.03	0.28	0.00	0.00	0.03	0.08	6
A0A081QLU6	404	40564	Streptococcus mitis.	Beta Solenoid	Fibrinogen Binding Family Protein	77.87	0.11	0.47	60.66	0.00	0.00	0.00	6.64	16.82	15.41	3.96	0.25	0.00	0.12	0.00	0.14	0.00	0.25	0.10	0.12	0.00	6
A0A081QLV2	148	15258	Streptococcus mitis.	Beta Solenoid	Surface Protein SdrI	95.63	0.10	0.00	63.25	0.00	0.00	0.00	1.20	21.08	14.47	0.00	0.25	0.00	0.12	0.00	0.19	0.00	0.28	0.03	0.08	0.00	6
A0A0B5E2J9	244	23857	Celeribacter indicus.	Helical	Collagen Alpha Chain	78.24	0.09	0.00	0.00	3.44	73.28	13.74	0.38	0.38	8.78	4.10	0.04	0.17	0.12	0.01	0.25	0.01	0.14	0.02	0.05	0.16	10
A0A6G2FVF1	59	6209	Halogeometricum	Beta Barrel	Lysine Biosynthesis Protein	89.77	0.09	0.00	32.47	0.00	0.00	0.00	18.18	1.30	48.05	0.00	0.12	0.07	0.10	0.08	0.10	0.05	0.10	0.22	0.03	0.08	10
A0A0U5FDN5	152	15619	Limosilactobacillus reuteri	Helical	Surface Protein SdrI	89.68	0.10	0.00	0.00	0.00	87.65	0.00	0.59	0.00	11.76	26.32	0.19	0.00	0.10	0.00	0.28	0.00	0.11	0.12	0.11	0.07	7
A0A0W1SS17	59	6311	Haloferax profundi.	Beta Barrel	Acetylornithine deacetylase (structurally identical to LysW)	89.47	0.10	0.00	32.47	0.00	0.00	0.00	18.18	1.30	48.05	0.00	0.07	0.07	0.08	0.08	0.10	0.08	0.10	0.24	0.08	0.05	10
A0A161HIE3	135	13730	Sugiyamaella lignohabitans.	Helical	Putative MAP7 domain-containing protein 3	83.76	0.10	0.00	0.00	0.00	80.39	3.27	0.00	0.00	16.34	17.78	0.36	0.01	0.04	0.04	0.13	0.02	0.04	0.29	0.00	0.02	9
A0A194W441	140	14893	Valsa mali.	Helical	A-kinase anchor protein 5	71.50	0.10	0.00	7.59	1.90	54.43	6.96	1.27	2.53	25.32	0.00	0.13	0.04	0.22	0.04	0.08	0.11	0.04	0.28	0.02	0.03	10
A0A1H2YLX6	43	4108	Paracoccus sanguinis.	Helical	Sulfite exporter TauE/SafE family protein	75.39	0.08	0.00	0.00	0.00	47.54	0.00	3.28	1.64	47.54	0.00	0.07	0.23	0.12	0.02	0.19	0.19	0.00	0.05	0.05	0.05	9
A0A1L3LMW8	118	10331	Sinorhizobium americanum.	Helical	Membrane Glycoprotein	83.72	0.11	0.00	0.00	0.00	72.06	0.00	5.15	4.41	18.38	35.59	0.45	0.04	0.19	0.01	0.00	0.06	0.00	0.01	0.01	0.19	8
A0A1S3KBP0	281	26572	Lingula unguis.	Beta Solenoid	Integumentary mucin C.1-like	74.00	0.10	1.00	54.52	0.00	0.00	0.00	1.00	17.73	25.75	38.43	0.30	0.03	0.10	0.01	0.02	0.01	0.01	0.12	0.10	0.24	10
A0A1S3QEG0	385	39206	Salmo salar (Atlantic salmon).	Beta Solenoid	Bile salt-activated lipase-like	90.93	0.09	0.00	59.31	0.00	0.00	0.00	0.00	16.13	24.56	1.04	0.10	0.11	0.01	0.06	0.17	0.01	0.08	0.17	0.03	0.24	10
A0A1V3JF38	278	27532	Rodentibacter myodis.	Helical	Methyl-accepting transducer domain-containing protein	88.31	0.10	0.00	0.00	0.00	92.23	0.00	0.68	0.00	7.09	10.07	0.39	0.00	0.10	0.01	0.14	0.00	0.10	0.18	0.02	0.01	8
A0A210R5G6	504	46312	Mizuhopecten yessoensis	Beta Solenoid	Mucin-17	83.92	0.10	0.00	79.89	0.00	0.00	0.00	0.00	7.85	12.26	31.75	0.00	0.25	0.20	0.00	0.01	0.21	0.03	0.00	0.19	0.08	7
A0A285VWK1	140	13351	Chromohalobacter canadensis.	Helical	Uncharacterized protein	79.8	0.08	0.00	0.00	1.90	41.14	0.00	18.99	16.46	21.51	20.00	0.03	0.26	0.08	0.01	0.24	0.00	0.18	0.04	0.06	0.06	9
A0A2R6IN80	61	5718	Halobacteriales archaeon	Helical	Membrane Transporter	75.37	0.08	0.00	0.00	3.80	53.16	0.00	3.80	5.06	34.18	22.95	0.15	0.16	0.18	0.08	0.20	0.00	0.02	0.00	0.02	0.15	8
A0A2T2YP25	222	21004	Adhaeribacter arboris.	Helical	Uncharacterized protein	92.78	0.10	0.00	0.00	0.00	90.83	0.00	0.00	0.00	9.17	8.11	0.38	0.00	0.23	0.03	0.13	0.00	0.02	0.10	0.06	0.00	8
A0A3B3HT88	584	54747	Oryzias latipes	Beta Solenoid	Uncharacterized protein	77.56	0.10	0.00	82.89	0.00	0.00	0.00	0.00	3.82	13.29	12.33	0.24	0.03	0.07	0.02	0.00	0.06	0.01	0.07	0.19	0.27	10
A0A3R9LQN5	188	18953	Streptococcus oralis.	Helical	Cell envelope integrity inner membrane protein TolA	75.72	0.10	0.00	0.00	0.00	89.81	0.00	0.00	0.00	10.19	36.17	0.32	0.01	0.12	0.00	0.19	0.00	0.16	0.15	0.02	0.01	8
A0A402B4U1	82	7580	Dictyobacter alpinus.	Beta Solenoid	Uncharacterized protein	74.29	0.08	1.00	60.00	0.00	0.00	0.00	2.00	9.00	28.00	4.88	0.06	0.22	0.23	0.01	0.06	0.01	0.10	0.02	0.04	0.20	10
A0A6J1EF88	130	12290	Cucurbita moschata	Beta Solenoid	Uncharacterized protein LOC111432757	83.41	0.10	0.68	66.22	0.00	0.00	0.00	0.00	1.35	31.75	0.00	0.15	0.07	0.08	0.08	0.07	0.08	0.00	0.08	0.27	0.08	9

A0A6J2KDU0	276	24220	Bombyx mandarina	Beta Solenoid	Mucin-5AC-like isoform X2	79.82	0.10	0.00	35.37	0.00	0.00	0.00	18.37	2.38	43.88	33.33	0.34	0.08	0.05	0.13	0.01	0.03	0.02	0.00	0.09	0.20	9
A0A6P6JE53	513	45870	Carassius auratus	Beta Solenoid	Mucin-22-like	81.95	0.08	0.00	64.78	0.00	0.00	0.00	0.75	15.44	19.03	29.24	0.27	0.14	0.02	0.01	0.10	0.01	0.02	0.07	0.22	0.09	10
A0A6P6QXS3	582	50121	Carassius auratus	Beta Solenoid	Uncharacterized protein LOC113114559	73.29	0.05	0.33	72.33	0.00	0.00	0.00	0.00	20.33	7.01	0.00	0.18	0.27	0.09	0.00	0.00	0.00	0.09	0.07	0.18	0.09	8
A0A6P6S2C3	61	4865	Cyclospora cayetanensis.	Helical	Ice-structuring protein 4-like	71.29	0.10	0.00	0.00	0.00	65.82	0.00	0.00	1.27	32.91	32.79	0.67	0.07	0.03	0.07	0.02	0.00	0.02	0.02	0.07	0.00	8
A0A7C6N7M8	41	3297	Clostridiales bacterium.	Helical	Glycine zipper domain-containing protein	80.49	0.05	0.00	0.00	0.00	61.02	0.00	0.00	1.69	37.29	29.27	0.24	0.32	0.22	0.02	0.02	0.00	0.02	0.00	0.05	0.07	8
A0A7J3WM24	34	3686	Methanomicrobia archaeon.	Helical	50S ribosomal protein P1	76.34	0.09	0.00	0.00	0.00	55.77	0.00	0.00	0.00	44.23	23.53	0.15	0.09	0.00	0.03	0.09	0.00	0.03	0.47	0.06	0.06	8
A0A7R8GZ83	639	63951	Lepeophtheirus salmonis	Beta Solenoid	(salmon louse) hypothetical protein	85.27	0.10	0.00	85.54	0.00	0.00	0.00	0.30	4.87	9.29	31.30	0.09	0.06	0.40	0.00	0.01	0.10	0.14	0.08	0.08	0.00	10
A0A7T2ZWV9	232	24594	Streptococcus oralis.	Beta Solenoid	Uncharacterized protein	78.55	0.10	0.40	65.20	0.00	0.00	0.00	0.00	12.80	21.60	3.45	0.22	0.00	0.03	0.00	0.19	0.00	0.23	0.11	0.06	0.10	7
A0A7W1B186	42	3812	Geodermatophilaceae bacterium.	Helical	Mycofactocin biosynthesis peptidyl-dipeptidase MftE	83.79	0.06	0.00	0.00	10.00	41.67	0.00	0.00	0.00	48.33	19.05	0.26	0.21	0.07	0.07	0.19	0.00	0.07	0.02	0.02	0.02	9
A0A7W5FMM2	94	10103	Paenibacillus phyllosphaerae.	Helical	Phage/plasmid-associated DNA primase	73.56	0.10	0.00	0.00	0.00	63.39	0.00	2.68	4.46	29.47	19.15	0.24	0.00	0.10	0.02	0.01	0.01	0.05	0.45	0.05	0.02	9
A0A803K9S1	1043	105418	Xenopus tropicalis	Beta Solenoid	Uncharacterized protein	77.28	0.11	0.57	67.95	0.00	0.00	0.00	0.00	14.89	16.59	0.00	0.12	0.00	0.01	0.05	0.12	0.06	0.05	0.10	0.21	0.25	10
A0A815YBE8	132	11909	Rotaria magnacalcarata.	Helical	Antifreeze glycoprotein	82.06	0.10	0.00	0.00	0.00	70.67	0.00	2.67	1.33	25.33	30.30	0.29	0.04	0.03	0.02	0.00	0.02	0.04	0.00	0.13	0.42	8
A0A822CNU0	99	9489	Rotaria sp. Silwood1.	Helical	Uncharacterized protein	87.75	0.10	0.00	0.00	0.00	78.63	0.00	0.00	0.85	20.52	4.04	0.23	0.01	0.37	0.07	0.07	0.01	0.08	0.00	0.03	0.10	9
A8WRF4	192	18775	Caenorhabditis briggsae.	Beta Solenoid	Protein CBG01852	84.51	0.11	1.90	74.29	0.00	0.00	0.00	0.00	0.48	23.33	5.21	0.15	0.00	0.05	0.04	0.01	0.05	0.01	0.10	0.20	0.38	9
E6QKN5	50	4507	mine drainage metagenome.	Helical	Mucin-associated surface protein	83.80	0.09	0.00	0.00	0.00	66.18	0.00	2.94	0.00	30.88	8.00	0.30	0.10	0.12	0.00	0.00	0.00	0.00	0.12	0.16	0.16	6
Average	228	21974.55	N/A	N/A	N/A	81.16	0.09	0.17	28.02	0.55	38.01	0.63	2.94	5.46	24.21	15.14	0.23	0.08	0.12	0.03	0.09	0.04	0.07	0.11	0.09	0.11	8.58

Appendix B - Table 01 – Proteins identified from the Cys- and Cys+ run.

*Proteins identified from the Trembl abiotic-like protein identification search are presented from the Cys- (**a**) and Cys+ (**b**) run. Identified proteins represent all remaining proteins after screening through the following criteria: an abiotic index ≥95%, sequence length >30 residues, relative abiotic score ≤0.12, pLDDT score ≥70%, tandem repeats <40%, and unstructured regions ≤50%. For each protein, accession number, structural classification, DSSP structural analysis, individual amino acid content percentage, abiotic index, pLDDT score, tandem repeat percentage, sequence length, source organism are reported.*

Cys +

Accession	Amino Acids	Mass	Organism	Primary Structural Feature	Protein Name	Avg pLDDT	Relative Abiotic Score	Isolated B bridge	B ladder	3,10 Helix	A Helix	Pi Helix	Hydrogen bonded turn	Bend	Un-structured	Tandem Repeat Percentage	A	G	V	P	L	I	D	E	S	T	Relative Complexity
A0A7C6N7M8	41	3297	Clostridiales bacterium	Helical	Glycine zipper domain-containing protein	80.49	0.06	0.00	0.00	0.00	61.02	0.00	0.00	1.69	37.29	29.27	0.24	0.00	0.32	0.22	0.02	0.02	0.00	0.02	0.00	0.05	0.07
A0A7W1B186	42	3812	Geodermatophilaceae bacterium	Helical	Mycofactocin biosynthesis peptidyl-dipeptidase MftE	83.79	0.06	0.00	0.00	10.00	41.67	0.00	0.00	0.00	48.33	19.05	0.26	0.00	0.21	0.07	0.07	0.19	0.00	0.07	0.02	0.02	0.02
A0A6P6JE53	513	45870	Carassius auratus	Beta Solenoid	Mucin-22-like	81.95	0.08	0.00	64.78	0.00	0.00	0.00	0.75	15.44	19.03	29.24	0.27	0.01	0.14	0.02	0.01	0.10	0.01	0.02	0.07	0.22	0.09
A0A7S3VAI7	100	9301	Chaetoceros debilis	Helical	ABC transmembrane type-1 domain-containing protein	73.70	0.09	0.00	5.08	0.00	60.17	0.00	2.54	3.39	28.82	34.00	0.25	0.00	0.11	0.24	0.01	0.00	0.18	0.06	0.01	0.05	0.05
A0A100JL95	71	6747	Streptomyces scabiei	Helical	Secreted protein	71.26	0.09	0.00	0.00	11.24	34.83	0.00	8.99	4.49	40.45	5.63	0.24	0.03	0.07	0.13	0.13	0.13	0.01	0.04	0.03	0.03	0.13
A0A059FV71	158	14821	Hyphomonas johnsonii MHS-2	Helical	Lipoprotein	70.98	0.09	0.00	0.00	0.00	73.30	0.00	2.84	0.00	23.86	16.46	0.47	0.01	0.01	0.08	0.01	0.03	0.01	0.09	0.18	0.04	0.04
A0A402B4U1	82	7580	Dictyobacter alpinus	Beta Solenoid	Uncharacterized protein	74.29	0.09	1.00	60.00	0.00	0.00	0.00	2.00	9.00	28.00	34.15	0.06	0.01	0.22	0.23	0.01	0.06	0.01	0.10	0.02	0.04	0.20
A0A285VWK1	140	13351	Chromohalobacter canadensis	Helical	Uncharacterized protein	79.80	0.09	0.00	0.00	1.90	41.14	0.00	18.99	16.46	21.51	20.00	0.03	0.00	0.26	0.08	0.01	0.24	0.00	0.18	0.04	0.06	0.06
A0A7V2QI93	55	5799	Thermoflexia bacterium	Beta Barrel	LysWC	91.26	0.11	0.00	28.77	0.00	0.00	0.00	21.92	0.00	49.31	0.00	0.09	0.07	0.07	0.20	0.09	0.09	0.02	0.07	0.22	0.00	0.04
A0A6J7GWF5	43	4172	freshwater metagenome	Helical	Unannotated protein	77.46	0.09	0.00	0.00	0.00	45.90	0.00	3.28	4.92	45.90	9.30	0.21	0.02	0.07	0.26	0.00	0.14	0.02	0.12	0.02	0.02	0.07
E6QKN5	50	4507	mine drainage metagenome	Helical	Mucin-associated surface protein	83.80	0.09	0.00	0.00	0.00	66.18	0.00	2.94	0.00	30.88	8.00	0.30	0.00	0.10	0.12	0.00	0.00	0.00	0.00	0.12	0.16	0.16
A0A1V3JF38	278	27532	Rodentibacter myodis	Helical	Methyl-accepting transducer domain-containing protein	88.31	0.09	0.00	0.00	0.00	92.23	0.00	0.68	0.00	7.09	10.07	0.39	0.00	0.00	0.10	0.01	0.14	0.00	0.10	0.18	0.02	0.01
A0A6J2KDU0	276	24220	Bombyx mandarina	Beta Solenoid	Mucin-5AC-like isoform X2	79.82	0.10	0.00	35.37	0.00	0.00	0.00	18.37	2.38	43.88	33.33	0.34	0.00	0.08	0.05	0.13	0.01	0.03	0.02	0.00	0.09	0.20
A0A161HIE3	135	13730	Sugiyamaella lignohabitans	Helical	Putative MAP7 domain-containing protein 3	83.76	0.10	0.00	0.00	0.00	80.39	3.27	0.00	0.00	16.34	35.56	0.36	0.00	0.01	0.04	0.04	0.13	0.02	0.04	0.29	0.00	0.02
A0A7W5FMM2	94	10103	Paenibacillus phyllosphaerae	Helical	Phage/plasmid-associated DNA primase	73.56	0.10	0.00	0.00	0.00	63.39	0.00	2.68	4.46	29.47	19.15	0.24	0.00	0.00	0.10	0.02	0.01	0.01	0.05	0.45	0.05	0.02
A0A3R9LQN5	188	18953	Streptococcus oralis	Helical	Cell envelope integrity inner membrane protein TolA	75.72	0.10	0.00	0.00	0.00	89.81	0.00	0.00	0.00	10.19	36.17	0.32	0.00	0.01	0.12	0.00	0.19	0.00	0.16	0.15	0.02	0.01
A0A1H2YLX6	43	4108	Paracoccus sanguinis	Helical	Sulfite exporter TauE/SafE family protein	75.39	0.10	0.00	0.00	0.00	47.54	0.00	3.28	1.64	47.54	0.00	0.07	0.00	0.23	0.12	0.02	0.19	0.19	0.00	0.05	0.05	0.05
A0A1S3KBP0	281	26572	Lingula unguis	Beta Solenoid	Integumentary mucin C.1-like	74.00	0.10	1.00	54.52	0.00	0.00	0.00	1.00	17.73	25.75	38.43	0.30	0.01	0.03	0.10	0.01	0.02	0.01	0.01	0.12	0.10	0.24
A0A6P6S2C3	61	4865	Cyclospora cayetanensis	Helical	Ice-structuring protein 4-like	71.29	0.10	0.00	0.00	0.00	65.82	0.00	0.00	1.27	32.91	32.79	0.67	0.00	0.07	0.03	0.07	0.02	0.00	0.02	0.02	0.07	0.00
A0A674A7C7	108	9890	Salmo trutta	Beta Solenoid	Uncharacterized protein	78.67	0.10	0.79	39.68	0.00	0.00	0.00	19.05	1.59	38.89	14.81	0.06	0.14	0.22	0.04	0.00	0.11	0.02	0.00	0.06	0.23	0.07
A0A081QLU6	404	40564	Streptococcus mitis	Beta Solenoid	Fibrinogen binding family protein	77.87	0.10	0.47	60.66	0.00	0.00	0.00	6.64	16.82	15.41	3.96	0.25	0.00	0.00	0.12	0.00	0.14	0.00	0.25	0.10	0.12	0.00
A0A7T2ZWV9	232	24594	Streptococcus oralis	Beta Solenoid	Uncharacterized protein	78.55	0.10	0.40	65.20	0.00	0.00	0.00	0.00	12.80	21.60	3.45	0.22	0.00	0.00	0.03	0.00	0.19	0.00	0.23	0.11	0.06	0.10
A0A081QLV2	148	15258	Streptococcus mitis	Beta Solenoid	Surface protein SdrI	95.63	0.10	0.00	63.25	0.00	0.00	0.00	1.20	21.08	14.47	0.00	0.25	0.00	0.00	0.12	0.00	0.19	0.00	0.28	0.03	0.08	0.00
A0A3B3HT88	584	54747	Oryzias latipes	Beta Solenoid	Uncharacterized protein	77.56	0.10	0.00	82.89	0.00	0.00	0.00	0.00	3.82	13.29	12.33	0.24	0.00	0.03	0.07	0.02	0.00	0.06	0.01	0.07	0.19	0.27
A0A5R8Y1S9	79	7506	Arcobacter arenosus	Helical	Uncharacterized protein	85.10	0.10	0.00	0.00	0.00	64.95	0.00	0.00	1.03	34.02	5.06	0.04	0.04	0.20	0.24	0.01	0.14	0.08	0.11	0.00	0.05	0.05
A0A1H5WW24	59	6225	Halobellus limi	Beta Barrel	LysW	90.36	0.10	0.00	32.47	0.00	0.00	0.00	20.78	0.00	46.75	6.78	0.12	0.00	0.07	0.08	0.08	0.08	0.07	0.08	0.24	0.03	0.10

A0A2R6IN80	61	5718	Halobacteriales archaeon	Helical	Transporter	75.37	0.10	0.00	0.00	3.80	53.16	0.00	3.80	5.06	34.18	22.95	0.15	0.00	0.16	0.18	0.08	0.20	0.00	0.02	0.00	0.02	0.15
A0A2T2YP25	222	21004	Adhaeribacter arboris	Helical	Uncharacterized protein	92.78	0.10	0.00	0.00	0.00	90.83	0.00	0.00	0.00	9.17	8.11	0.38	0.00	0.00	0.23	0.03	0.13	0.00	0.02	0.10	0.06	0.00
A0A822CNU0	99	9489	Rotaria sp	Helical	Uncharacterized protein	87.75	0.10	0.00	0.00	0.00	78.63	0.00	0.00	0.85	20.52	4.04	0.23	0.00	0.01	0.37	0.07	0.07	0.01	0.08	0.00	0.03	0.10
A0A815YBE8	132	11909	Rotaria magnacalcarata	Helical	Antifreeze glycoprotein	82.06	0.11	0.00	0.00	0.00	70.67	0.00	2.67	1.33	25.33	30.30	0.29	0.00	0.04	0.03	0.02	0.00	0.02	0.04	0.00	0.13	0.42
A0A0B5E2J9	244	23857	Celeribacter indicus	Helical	Collagen alpha chain	78.24	0.11	0.00	0.00	3.44	73.28	13.74	0.38	0.38	8.78	4.10	0.04	0.00	0.17	0.12	0.01	0.25	0.01	0.14	0.02	0.05	0.16
A0A6J1EF88	130	12290	Cucurbita moschata	Beta Solenoid	Uncharacterized protein	83.41	0.11	0.68	66.22	0.00	0.00	0.00	0.00	1.35	31.75	0.00	0.15	0.00	0.07	0.08	0.08	0.07	0.08	0.00	0.08	0.27	0.08
A0A1L3LMW8	118	10331	Sinorhizobium americanum	Helical	Membrane Glycoprotein	83.72	0.11	0.00	0.00	0.00	72.06	0.00	5.15	4.41	18.38	35.59	0.45	0.00	0.04	0.19	0.01	0.00	0.06	0.00	0.01	0.01	0.19
A0A7J3WM24	34	3686	Methanomicrobia archaeon	Helical	50S ribosomal protein P1	76.34	0.11	0.00	0.00	0.00	55.77	0.00	0.00	0.00	44.23	23.53	0.15	0.00	0.09	0.00	0.03	0.09	0.00	0.03	0.47	0.06	0.06
A0A7R8GZ83	639	63951	Lepeophtheirus salmonis	Beta Solenoid	(salmon louse) hypothetical protein	85.27	0.11	0.00	85.54	0.00	0.00	0.00	0.30	4.87	9.29	34.74	0.09	0.01	0.06	0.40	0.00	0.01	0.10	0.14	0.08	0.08	0.00
A0A210R5G6	504	46312	Mizuhopecten yessoensis	Beta Solenoid	Mucin-17	83.92	0.11	0.00	79.89	0.00	0.00	0.00	0.00	7.85	12.26	31.75	0.00	0.00	0.25	0.20	0.00	0.01	0.21	0.03	0.00	0.19	0.08
A0A194W441	140	14893	Valsa mali	Helical	A-kinase anchor protein 5	71.50	0.11	0.00	7.59	1.90	54.43	6.96	1.27	2.53	25.32	0.00	0.13	0.00	0.04	0.22	0.04	0.08	0.11	0.04	0.28	0.02	0.03
E9PWP7	4499	462664	Mus musculus (Mouse)	Helical/Beta Sheet	Mucin-17	84.37	0.12	0.76	22.22	3.28	24.75	0.00	13.13	8.08	27.78	18.89	0.07	0.00	0.03	0.11	0.02	0.02	0.05	0.11	0.07	0.05	0.43
I3DU15	80	7998	Bacillus methanolicus	Beta Barrel	Putative secreted protein	85.37	0.12	0.00	48.98	0.00	0.00	0.00	9.18	6.12	35.72	5.00	0.06	0.03	0.08	0.16	0.14	0.14	0.08	0.05	0.03	0.04	0.16
A0A1G2ZSS2	58	6014	Planctomycetes bacterium RBG_13_63_9	Helical/Beta Sheet	Ferrodoxin	90.62	0.10	0.00	21.05	0.00	13.16	0.00	15.79	0.00	0.00	50.00	0.10	0.14	0.05	0.16	0.03	0.03	0.05	0.16	0.14	0.02	0.09
Average	**280.63**	**27706**	**N/A**	**N/A**	**N/A**	**80.88**	**0.10**	**0.13**	**23.10**	**0.89**	**37.88**	**0.60**	**4.74**	**4.57**	**26.84**	**18.15**	**0.22**	**0.01**	**0.09**	**0.14**	**0.03**	**0.09**	**0.04**	**0.07**	**0.10**	**0.07**	**0.10**

Statistical Test: One Way ANOVA with Tukey Post Hoc Analysis			
F			29.05
P value			<0.0001
R squared			0.8994
Tukey's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Adjusted P Value
L-cys vs. D-cys	8.260	3.405 to 13.12	0.0006
L-cys vs. L-ala	12.67	8.041 to 17.30	<0.0001
L-cys vs. D-ala	15.02	9.350 to 20.69	<0.0001
L-cys vs. L-met	15.44	8.893 to 21.99	<0.0001
D-cys vs. L-ala	4.410	-0.4452 to 9.265	0.0851
D-cys vs. D-ala	6.760	0.9044 to 12.62	0.0196
D-cys vs. L-met	7.180	0.4716 to 13.89	0.0328
L-ala vs. D-ala	2.350	-3.320 to 8.020	0.7172
L-ala vs. L-met	2.770	-3.777 to 9.317	0.7021
D-ala vs. L-met	0.4200	-6.899 to 7.739	0.9998

Appendix B - Table 02 – Statistical analysis of data presented in Figure 4.03a.

Statistical Test: Unpaired T Test	
P value	0.0015
P value summary	**
One- or two-tailed P value?	two-tailed
t, df	t=7.696, df=4

Appendix B - Table 03 – Statistical analysis of data presented in Figure 4.03b

(a)

Detected peptide composition	Detected mass	Predicted mass	Detectionppm error
ADG	262.1023	262.1034	3.5553
TTT	322.1614	322.1609	1.0486
CDS	324.0844	324.0860	4.5309
IIS	332.2191	332.2180	2.6076
ACEG	379.1295	379.1282	2.8087
CCEI	467.1649	467.1629	3.8654
DEEE	521.1702	521.1726	4.0416
AVVVV	486.3274	486.3286	1.8722
AADDT	492.1918	492.1936	3.1741
ACCDT	512.1476	512.1479	0.0558
ADDPT	518.2080	518.2093	1.8409
EPSST	520.2253	520.2249	0.1170
CCCTT	530.1401	530.1408	0.7533
AEIIP	542.3197	542.3185	1.8176
CEEGI	550.2181	550.2178	0.1540
EIIPP	568.3325	568.3341	2.3743
AAADGV	503.2440	503.2460	3.3884
ACCGGP	507.1698	507.1690	0.9598
ADGGPT	517.2250	517.2253	0.0396
AAACPT	533.2375	533.2388	1.9510
CGPSSS	537.1978	537.1973	0.2772
AGPPVV	539.3169	539.3188	3.0230
AAADPP	541.2613	541.2616	0.1872
AEEGGS	549.2152	549.2151	0.4262
CEEGGG	551.1766	551.1766	0.5475
CCGSST	557.1700	557.1694	0.4887
CEGPSS	579.2082	579.2079	0.0711
APPTVV	583.3427	583.3450	3.4133
CCCCGP	585.1266	585.1288	3.1920
PPPSST	585.2880	585.2879	0.3962
CCCPSS	599.1640	599.1622	2.4009
STTTTV	609.3088	609.3090	0.1610
CIPPSV	615.3191	615.3171	2.7594
CCDSTV	627.2083	627.2113	4.1426
EIPPSS	629.3145	629.3141	0.0348
CCCIPP	635.2375	635.2350	3.3243
CCCDIP	653.2115	653.2092	2.9513
CDGGGGP	562.1912	562.1926	1.8791
AACCCCD	688.1540	688.1557	1.9967
CCCCCTT	736.1630	736.1591	4.7475
CCIIIVV	762.4263	762.4252	0.7770
DDIIPPP	766.3981	766.3982	0.4534
AAAACGGG	577.2370	577.2399	4.4542
GGISSSST	695.3234	695.3206	3.4266
CEEGGGPP	745.2834	745.2821	1.1371
CDDGSSST	771.2436	771.2461	2.7157
EIIPPSSS	829.4303	829.4302	0.4318
CIIIIPPT	869.5151	869.5165	1.0804
GGGGGPPPT	696.3274	696.3311	4.8569
AAAAAPVVV	768.4579	768.4614	4.0371
AAGVVVVVV	812.5210	812.5240	3.1914
CCCCCGGPS	832.1876	832.1915	4.1571
AADPVVVVV	868.5109	868.5138	2.8593
CCCDIIIII	1008.4930	1008.4930	0.3972
AAGGGGGGPP	697.3237	697.3264	3.3004
CCCGGGGGGV	769.2440	769.2426	1.3260
AAAAAAADVV	829.4410	829.4414	0.1075
AAADGTTVVV	903.4771	903.4782	0.6410
CCCCCCGGGP	905.1857	905.1901	4.2720
AAAAAAAAAT	830.4337	830.4366	3.0011
GGGGPPPPPPP	926.4718	926.4731	0.7798
ACCCCCCCGGT	1026.2060	1026.2100	2.8299
AACDDVVVVVV	1088.5640	1088.5660	0.8864
AAACGGGGGPPP	911.4030	911.4040	0.4974
AADDGTTTTTTV	1153.5220	1153.5220	0.2206

(b)

Detected peptide composition	Detected mass	Predicted mass	Detectionppm error
ET	249.1091	249.1081	3.4402
ADG	262.1022	262.1034	3.7842
CGPP	373.1536	373.1540	0.6898
ADPP	399.1858	399.1874	3.6510
EPPT	443.2113	443.2137	4.7296
DIIP	457.2654	457.2657	0.1695
AAPPP	452.2494	452.2504	1.6921
AVVVV	486.3268	486.3286	3.2293
PPPPT	508.2764	508.2766	-0.2608
ACCDE	540.1402	540.1429	4.4103
DIIVV	558.3488	558.3498	1.1156
CDDEP	578.1792	578.1763	4.5148
GGGGIT	461.2361	461.2354	0.8354
GGGPPS	471.2175	471.2198	4.2758
ADGGPT	517.2247	517.2253	0.5791
AAACCD	553.1744	553.1745	-0.4700
PPSSSS	561.2517	561.2515	-0.1669
STTTTV	609.3079	609.3090	1.1848
CDPSVV	619.2725	619.2756	4.4908
GGGGGPS	488.2126	488.2100	4.7694
GGGGGII	530.2954	530.2933	3.4365
CIISSSS	696.3269	696.3233	4.5976
GGISSSST	695.3238	695.3206	4.0306
CGISSSSS	727.2951	727.2927	2.7380
GGIIISSS	733.4131	733.4091	4.9353
GGGGIIPST	758.4069	758.4043	2.8677
AAAAAPVVV	768.4581	768.4614	3.6987
AAGVVVVVV	812.5204	812.5240	3.8806
AGTVVVVVV	842.5317	842.5346	2.9122
AAGGGGGGPP	697.3231	697.3264	4.0891
CDGGGGPPSS	833.3118	833.3094	2.2909
ACCGGGGGGPP	832.3084	832.3076	0.4193

(c)

Detected peptide composition	Detected mass	Predicted mass	Detectionppm error
DIII	473.2954	473.2970	2.7714
AGIIT	474.2923	474.2922	0.5142
DGVVV	488.2697	488.2715	3.1303
DSVVV	518.2808	518.2821	1.9660
TTTVV	520.2976	520.2977	0.3136
IISTT	534.3142	534.3134	1.0652
EPPPP	536.2692	536.2715	3.6941
EIITT	576.3223	576.3239	2.3720
GGPPPS	511.2512	511.2511	0.3309
GPPPPV	563.3165	563.3188	3.5258
DDDGGP	575.1943	575.1944	0.4335
ADPPTV	599.3032	599.3035	0.0323
ISSTTT	609.3093	609.3090	0.0658
IITPPP	637.3899	637.3920	2.7604
DDIPVV	657.3454	657.3454	0.5074
AAAPVVV	626.3852	626.3872	2.5679
AAAAAAAV	615.3436	615.3460	3.3372
GGGGPSST	619.2716	619.2682	4.8790

Appendix B - Table 04 – Compositions of peptides detected by TIMS-TOF MS.

Peptides were synthesised in the presence of excess L-cysteine or L-alanine and heated at 90˚C for 42 days, C18 purified, and analysed by TIMS-TOF MS. Amino acid compositions (alphabetical) and theoretical masses of peptides are indicated. Cysteine bearing sequences are in red text.

Statistical Test: Mann Whitney Test	
L-cysteine vs L-alanine	
P value	0.0188
P value summary	*
T Test Type	Two-tailed
Sum of Ranks	111.50, 59.50
Mann Whitney U	14.50
L-cysteine vs D-cysteine	
P value	0.0946
P value summary	ns
T Test Type	Two-tailed
Sum of Ranks	104.50, 66.50
Mann Whitney U	21.50
L-alanine vs D-cysteine	
P value	0.1563
P value summary	ns
T Test Type	Two-tailed
Sum of Ranks	69.50, 101.50
Mann Whitney U	24.50

Appendix B - Table 05 – Statistical analysis of data presented in Figure 4.08

Molar Cysteine Ratio	Reactive (CXXC Probability)	No Reactivity (CXXC Probability)
1%	2.73% (0.0273)	0.0686% (0.000686)
3.5%	19.6% (0.196)	0.812% (0.00812)

Appendix B - Table 06 – Upper and lower bound cysteine CXXC probability

250 million 10mer peptide sequences were randomly generated based on the abiotic amino acid abundances with cysteine making up 1% or 3.5% of the amino acid content, with its likelihood of incorporation unmodified (no reactivity) or multiplied 7.8x based on its observed synthesis effect, and the total number of peptides identified containing CXXC was reported.

Appendix C – Code

Note: The following scripts were designed by the author with support from the Open AI ChatGPT 3.5 and 4o models for execution in Python. All script outputs were manually verified by the author to ensure scripting consistency and effectiveness, and all scripts were extensively tested on synthetic data before use to ensure consistent script writing.

Appendix C.1 – Pymol CSP mapping script

```
# PyMOL script to color protein based on CSP values with blue to red gradient

# Define a color spectrum (blue for low, red for high)
# Adjust the maximum value to 1 for better contrast
spectrum b, blue_red, minimum=0.0, maximum=1.0

# Define CSP values for each residue
alter resi 1, b=0.001183216
alter resi 2, b=0.119636533
alter resi 3, b=0.116825083
alter resi 4, b=0.240765654
alter resi 5, b=0.060832557
alter resi 6, b=1.0  # Cap at 1
alter resi 7, b=0.667918857
alter resi 8, b=1.0  # Cap at 1
alter resi 9, b=1.0  # Cap at 1
alter resi 10, b=1.0  # Cap at 1
alter resi 11, b=1.0  # Cap at 1
alter resi 12, b=1.0  # Cap at 1
alter resi 13, b=1.0  # Cap at 1
alter resi 14, b=0
alter resi 15, b=0
alter resi 16, b=0.043988635
alter resi 17, b=0.056960513
alter resi 18, b=0.046909487
alter resi 19, b=0.089503631
alter resi 20, b=0
alter resi 21, b=0.121505555
alter resi 22, b=0.00464758
alter resi 23, b=0.082553013
alter resi 24, b=0.024143322
alter resi 25, b=0.083611602
alter resi 26, b=0
alter resi 27, b=0.042244526
alter resi 28, b=0.041867649
alter resi 29, b=0.0056921
alter resi 30, b=0.094068592
alter resi 31, b=0.15779924
alter resi 32, b=0.085560505
alter resi 33, b=0.042374521
alter resi 34, b=0
alter resi 35, b=0.010904128
alter resi 36, b=0.080699442
alter resi 37, b=0.203333224
alter resi 38, b=1.0  # Cap at 1
alter resi 39, b=0.434399586
alter resi 40, b=0
alter resi 41, b=1.0  # Cap at 1
alter resi 42, b=1.0  # Cap at 1
alter resi 43, b=1.0  # Cap at 1
alter resi 44, b=1.0  # Cap at 1
```

```
alter resi 45, b=1.0  # Cap at 1
alter resi 46, b=0.40748865
alter resi 47, b=0.041891527
alter resi 48, b=0.256579228
alter resi 49, b=1.0  # Cap at 1
alter resi 50, b=0.100799802
alter resi 51, b=0
alter resi 52, b=0.125140321
alter resi 53, b=0.006640783
alter resi 54, b=0.036470536

# Apply the spectrum to the protein based on the CSP values
spectrum b, blue_red, minimum=0.0, maximum=1.0

# Set additional PyMOL visualisation settings
show cartoon
set cartoon_transparency, 0.2
set cartoon_smooth_loops, 1

# Optional: Add a legend for the color scale with adjusted max
ramp_new CSP_ramp, none, [0.0, 1.0], [blue, red]

# Zoom into the protein
zoom

# Automatically use the loaded structure instead of specifying file name
# The loaded file (whether cif or pdb) will automatically be used
```

Appendix C.2 – Abiotic protein bioinformatic scan

splitfasta.py

```
import sys
import os


def split_fasta_file(input_file, chunk_size=5e9):
    with open(input_file, "r") as f:
        current_chunk = ""
        current_size = 0
        chunk_num = 1
        for line in f:
            current_chunk += line
            current_size += len(line.encode('utf-8'))
            if current_size >= chunk_size and ">" in line:
                output_file = f"{os.path.splitext(input_file)[0]}_{chunk_num}.fasta"
                with open(output_file, "w") as out:
                    out.write(current_chunk)
                current_chunk = ""
                current_size = 0
                chunk_num += 1
        if current_chunk:
            output_file = f"{os.path.splitext(input_file)[0]}_{chunk_num}.fasta"
            with open(output_file, "w") as out:
                out.write(current_chunk)


def extract_accession_numbers(input_file, output_file, batch_size=500000000):
    with open(input_file, 'r') as in_file, open(output_file, 'w') as out_file:
        while True:
            lines = in_file.readlines(batch_size)
            if not lines:
                break
            accession_numbers = (line.split(',')[0] for line in lines)
            for accession in accession_numbers:
                out_file.write(accession + '\n')
            del lines, accession_numbers
```

```
if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: python extract_accession_numbers.py <input_file> <output_file>")
        sys.exit(1)
    else:
        input_file = sys.argv[1]
        output_file = sys.argv[2]

        # Split input file into smaller chunks
        split_fasta_file(input_file, chunk_size=5e9)

        # Process each chunk separately
        input_file_base = os.path.splitext(input_file)[0]
        for chunk_file in os.listdir():
            if chunk_file.startswith(input_file_base) and chunk_file.endswith(".fasta"):
                chunk_path = os.path.abspath(chunk_file)
                chunk_output_file = os.path.splitext(chunk_file)[0] + ".txt"
                extract_accession_numbers(chunk_path, chunk_output_file, batch_size=500000000)
```

abiotic_index_calculator_multithreadversion.py

```
import sys
from collections import Counter
from multiprocessing import Pool, cpu_count
from datetime import datetime

def read_fasta(file):
    with open(file, 'r') as f:
        sequences = {}
        current_seq = ''
        current_id = ''

        for line in f:
            if line.startswith('>'):
                if current_id:
                    sequences[current_id] = current_seq
```

```
                current_id = line.strip()[1:]
                current_seq = ''
            else:
                current_seq += line.strip()

        if current_id:
            sequences[current_id] = current_seq

    return sequences

def abiotic_index(args):
    seq_id, seq, amino_acids = args
    sequence = seq.upper()
    counts = Counter(sequence)
    total_count = sum(counts.values())
    abiotic_count = sum(counts[a] for a in amino_acids)
    index = (abiotic_count / total_count) * 100
    return seq_id, index, total_count

def main(file, cutoff, min_length):
    amino_acids = {'A', 'G', 'D', 'E', 'V', 'P', 'L', 'I', 'S', 'T', 'C'}
    sequences = read_fasta(file)

    num_cores = min(cpu_count(), 6)
    with Pool(num_cores) as pool:
        args = [(seq_id, seq, amino_acids) for seq_id, seq in sequences.items()]
        results = pool.map(abiotic_index, args)

    now = datetime.now()
    timestamp = now.strftime("%Y%m%d_%H%M%S")
    output_file = f'output_{timestamp}_cutoff_{cutoff}_min_length_{min_length}.txt'
    with open(output_file, 'w') as output:
        for seq_id, index, total_count in results:
            if index >= cutoff and total_count >= min_length:
                output.write(f'{seq_id}: {index:.2f}, Length: {total_count}\n')
                print(f'{seq_id}: {index:.2f}, Length: {total_count}')
```

```
if __name__ == '__main__':
    if len(sys.argv) != 4:
        print('Usage: python abiotic_index_calculator_multithreadversion.py <input_file> <cutoff> <min_length>')
    else:
        input_file = sys.argv[1]
        cutoff = float(sys.argv[2])
        min_length = int(sys.argv[3])
        main(input_file, cutoff, min_length)
```

extractacsimple.py

```
import sys


def extract_accessions(input_file, output_file):
    with open(input_file, 'r') as input_handle, open(output_file, 'w') as output_handle:
        for line in input_handle:
            accession = line.split('|')[1]
            output_handle.write(accession + '\n')


if __name__ == '__main__':
    if len(sys.argv) != 3:
        print('Usage: python extractacsimple.py <input_file> <output_file>')
    else:
        input_file = sys.argv[1]
        output_file = sys.argv[2]
        extract_accessions(input_file, output_file)
        print(f"Accessions extracted from {input_file} and saved to {output_file}")
```

experimentstructurefetcher.py

```
import sys
import os
import time
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import WebDriverWait
```

```
from selenium.webdriver.support import expected_conditions as EC

def setup_driver(download_folder):
    # Path to the chromedriver executable
    chrome_driver_path = "D:/chromedriver.exe"

    options = webdriver.ChromeOptions()
    # Set the default download directory to the specified folder
    prefs = {"download.default_directory": download_folder}
    options.add_experimental_option("prefs", prefs)

    # Instantiate a Chrome session
    driver = webdriver.Chrome(service=Service(chrome_driver_path), options=options)
    return driver

def download_pdb_files(accession, driver):
    url = f"https://www.uniprot.org/uniprotkb/{accession}/entry#structure"
    driver.get(url)
    try:
        # Scroll down to ensure all elements are loaded, including those that load dynamically
        driver.find_element(By.TAG_NAME, 'body').send_keys(Keys.PAGE_DOWN)
        WebDriverWait(driver, 15).until(
            EC.presence_of_element_located((By.CSS_SELECTOR, "a[href*='ebi.ac.uk/pdbe/entry-files/download']"))
        )
        time.sleep(5)  # Additional wait to let everything load

        # Find all links that contain PDB downloads
        pdb_links = driver.find_elements(By.CSS_SELECTOR, "a[href*='ebi.ac.uk/pdbe/entry-files/download']")
        for link in pdb_links:
            pdb_url = link.get_attribute('href')
            driver.execute_script("window.open(arguments[0]);", pdb_url)
            print(f"Downloading PDB file from {pdb_url}")
            time.sleep(2)  # Wait a couple of seconds to ensure file downloads
    except Exception as e:
        print(f"Failed to download PDB files for {accession}: {e}")

def main(input_file, output_folder):
```

```
    driver = setup_driver(output_folder)
    with open(input_file, "r") as file:
        accession_numbers = [line.split('|')[1].strip() for line in file if line.strip()]

    for accession in accession_numbers:
        download_pdb_files(accession, driver)

    driver.quit()

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: python script_name.py <input_data_file> <output_folder>")
        sys.exit(1)

    input_file = sys.argv[1]
    output_folder = sys.argv[2]
    os.makedirs(output_folder, exist_ok=True)
    main(input_file, output_folder)
```

basic_analysis_tool4.py

```
import requests
import sys
import csv
import re
from datetime import datetime

def download_entry(accession):
    url = f"https://www.uniprot.org/uniprot/{accession}.txt"
    try:
        response = requests.get(url)
        response.raise_for_status()  # Will raise an exception for HTTP codes not in 200-299 range
        return response.text
    except requests.RequestException as e:
        print(f"Error downloading data for accession {accession}: {e}")
        return None
```

```
def extract_data(entry_text, accession):
    data = {
        "accession": accession,  # Always use the provided accession number
        "amino_acids": "Not found",
        "mass": "Not found",
        "organism": "Not found",
        "subcellular_location": "Not found",
        "domain_keywords": "Not found",
        "protein_name": "Not found"
    }

    if entry_text is None:
        print("Error: No data to process.")
        return data  # Return the data dict with "Not found" for most fields but the accession

    patterns = {
        "amino_acids": re.compile(r"(\d+)\s+AA;"),
        "mass": re.compile(r"(\d+)\s+MW;"),
        "organism": re.compile(r"OS\s+(.+?)\."),
        "subcellular_location": re.compile(r"CC\s+-!- SUBCELLULAR LOCATION: (.+?)\."),
        "domain_keywords": re.compile(r"KW\s+(.+?)\."),
        "protein_name": re.compile(r"RecName: Full=(.+?);")
    }

    for key, pattern in patterns.items():
        match = pattern.search(entry_text)
        if match:
            data[key] = match.group(1).strip()

    return data

def main(input_file):
    output_file = f"basic_analysis_results_{datetime.now().strftime('%Y%m%d_%H%M%S')}.csv"

    with open(input_file, "r") as f:
        accession_numbers = [line.strip() for line in f if line.strip()]
```

```
    with open(output_file, "w", newline="") as csvfile:
        fieldnames = ["accession", "amino_acids", "mass", "organism", "subcellular_location", "domain_keywords",
"protein_name"]
        writer = csv.DictWriter(csvfile, fieldnames=fieldnames, delimiter='\t')
        writer.writeheader()

        for accession in accession_numbers:
            print(f"Processing accession {accession}")
            entry_text = download_entry(accession)
            data = extract_data(entry_text, accession)
            writer.writerow(data)
            if data['amino_acids'] == "Not found":
                print(f"No data found for {accession}")

if __name__ == "__main__":
    if len(sys.argv) != 2:
        print("Usage: python script_name.py <input_data_file>")
        sys.exit(1)

    input_file = sys.argv[1]
    main(input_file)
```

dssppipe.py

```
import os
import glob
import subprocess

def read_accession_numbers_from_file(file_path):
    with open(file_path, "r") as file:
        accession_numbers = [line.strip() for line in file]
    return accession_numbers

def process_pdb_files_with_dssp(accession_numbers, base_folder):
    for accession_number in accession_numbers:
        pdb_folder = os.path.join(base_folder, accession_number)

        if os.path.isdir(pdb_folder):
            pdb_files = glob.glob(os.path.join(pdb_folder, "*.pdb"))
```

```
        for pdb_file in pdb_files:
            file_name, _ = os.path.splitext(os.path.basename(pdb_file))
            dssp_file = os.path.join(pdb_folder, f"{file_name}.dssp")

            try:
                subprocess.run(["mkdssp", "-i", pdb_file, "-o", dssp_file], check=True)
                print(f"Processed PDB file {pdb_file} with DSSP.")
            except subprocess.CalledProcessError as e:
                print(f"Error running DSSP for {pdb_file}: {str(e)}")
    else:
        print(f"Subfolder not found: {pdb_folder}")

if __name__ == "__main__":
    import sys

    if len(sys.argv) != 3:
        print("Usage: python dssp_pipeline.py <input_file> <base_folder>")
        sys.exit(1)

    input_file = sys.argv[1]
    base_folder = sys.argv[2]

    accession_numbers = read_accession_numbers_from_file(input_file)
    process_pdb_files_with_dssp(accession_numbers, base_folder)
```

abioticscore.py

```
import argparse
import csv
from Bio import SeqIO

METEORIC_ABUNDANCE = {
    'A': 0.253,
    'C': 0.014,
    'G': 0.495,
    'V': 0.035,
```

```
    'P': 0.023,
    'L': 0.008,
    'I': 0.012,
    'D': 0.082,
    'E': 0.057,
    'S': 0.014,
    'T': 0.005
}

def clean_accession(accession):
    # Split the accession number and take the middle part if it contains '|'
    return accession.split('|')[1] if '|' in accession else accession

def relative_composition(sequence):
    sequence_length = len(sequence)
    composition = {}
    for aa in METEORIC_ABUNDANCE.keys():
        composition[aa] = sequence.count(aa) / sequence_length if sequence_length > 0 else 0
    return composition

def relative_abiotic_score(composition):
    deviations = [abs(composition.get(aa, 0) - METEORIC_ABUNDANCE[aa]) for aa in METEORIC_ABUNDANCE]
    return sum(deviations) / len(deviations)

def main(input_file, output_file):
    with open(input_file, "r") as fasta_file:
        sequences = SeqIO.parse(fasta_file, "fasta")
        with open(output_file, "w", newline='') as f:
            fieldnames = ['Accession'] + list(METEORIC_ABUNDANCE.keys()) + ['RelativeAbioticScore']
            writer = csv.DictWriter(f, fieldnames=fieldnames, delimiter='\t')
            writer.writeheader()

            for seq_record in sequences:
                accession = clean_accession(seq_record.id)
                composition = relative_composition(str(seq_record.seq))
                score = relative_abiotic_score(composition)
                composition['Accession'] = accession
```

```
            composition['RelativeAbioticScore'] = score
            writer.writerow(composition)

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="Calculate the relative composition of protein sequences and their deviation from meteoric abundance.")
    parser.add_argument("input_file", help="Input file containing FASTA formatted sequences.")
    parser.add_argument("output_file", help="Output file to store the results as a TSV file.")

    args = parser.parse_args()
    main(args.input_file, args.output_file)
```

extract_plddt.py

```
import os
import sys
import time  # Add this line to import the time module
from Bio.PDB import PDBParser
import numpy as np

def extract_plddt(input_folder):
    parser = PDBParser()
    results = []

    for root, dirs, files in os.walk(input_folder):
        for file in files:
            if file.endswith(".pdb"):
                accession = os.path.basename(root)
                pdb_file = os.path.join(root, file)
                structure = parser.get_structure(accession, pdb_file)
                model = structure[0]

                plddt_scores = []

                for atom in model.get_atoms():
                    plddt_scores.append(atom.get_bfactor())
```

```
            avg_plddt = np.mean(plddt_scores)
            results.append((accession, avg_plddt))

    return results

def save_tsv(results, output_file):
    with open(output_file, 'w') as f:
        f.write("Accession\tAvg_pLDDT\n")
        for accession, avg_plddt in results:
            f.write(f"{accession}\t{avg_plddt}\n")

if __name__ == "__main__":
    input_folder = sys.argv[1]
    output_file = f"output_{os.path.basename(input_folder)}_{time.strftime('%Y%m%d_%H%M%S')}.tsv"

    results = extract_plddt(input_folder)
    save_tsv(results, output_file)
```

tandemrepeatdetector.py

```
import sys
import time
from Bio import SeqIO

def read_fasta_sequences_from_file(input_file: str) -> dict:
    sequences = {}
    with open(input_file, "r") as f:
        for record in SeqIO.parse(f, "fasta"):
            # Split the record ID and take the second part, assuming 'tr|Accession|Description' format
            accession = record.id.split('|')[1] if '|' in record.id else record.id
            sequences[accession] = str(record.seq)
    return sequences

def tandem_repeat_percentage(sequence: str, min_repeat_length: int = 2, max_repeat_length: int = 12) -> float:
    tandem_repeats = []
    seq_length = len(sequence)
```

```
    for length in range(min_repeat_length, max_repeat_length + 1):
        for i in range(seq_length - 2 * length + 1):
            if sequence[i:i + length] == sequence[i + length:i + 2 * length]:
                repeat_start = i
                repeat_end = i + 2 * length
                while sequence[repeat_start:repeat_end] == sequence[repeat_end:repeat_end + length]:
                    repeat_end += length
                tandem_repeats.append((repeat_start, repeat_end))

    non_overlapping_repeats = []
    for repeat in sorted(tandem_repeats, key=lambda x: (x[0], x[1] - x[0])):
        if not non_overlapping_repeats or repeat[0] >= non_overlapping_repeats[-1][1]:
            non_overlapping_repeats.append(repeat)

    total_length = sum(repeat[1] - repeat[0] for repeat in non_overlapping_repeats)

    return total_length / seq_length * 100

def main():
    if len(sys.argv) != 2:
        print("Usage: python tandemrepeatdetector.py <input_fasta_file>")
        sys.exit(1)

    input_fasta_file = sys.argv[1]

    fasta_sequences = read_fasta_sequences_from_file(input_fasta_file)

    # Generate a timestamped output file name
    timestamp = time.strftime("%Y%m%d_%H%M%S")
    output_file = f"tandem_repeat_output_{timestamp}.tsv"

    # Write the results to the output file
    with open(output_file, "w") as f:
        f.write("Accession Number\tTandem Repeat Percentage\n")
        for accession, sequence in fasta_sequences.items():
            percentage = tandem_repeat_percentage(sequence)
            f.write(f"{accession}\t{percentage:.2f}\n")
```

```
    print(f"Results have been saved to {output_file}")

if __name__ == "__main__":
    main()
```

Appendix C.3 – Mass spec in vitro peptide identification scripts

peptidesequencepredictor.py

```
import sys
import time
from Bio import SeqIO

def read_fasta_sequences_from_file(input_file: str) -> dict:
    sequences = {}
    with open(input_file, "r") as f:
        for record in SeqIO.parse(f, "fasta"):
            # Split the record ID and take the second part, assuming 'tr|Accession|Description' format
            accession = record.id.split('|')[1] if '|' in record.id else record.id
            sequences[accession] = str(record.seq)
    return sequences

def tandem_repeat_percentage(sequence: str, min_repeat_length: int = 2, max_repeat_length: int = 12) -> float:
    tandem_repeats = []
    seq_length = len(sequence)

    for length in range(min_repeat_length, max_repeat_length + 1):
        for i in range(seq_length - 2 * length + 1):
            if sequence[i:i + length] == sequence[i + length:i + 2 * length]:
                repeat_start = i
                repeat_end = i + 2 * length
                while sequence[repeat_start:repeat_end] == sequence[repeat_end:repeat_end + length]:
                    repeat_end += length
                tandem_repeats.append((repeat_start, repeat_end))

    non_overlapping_repeats = []
    for repeat in sorted(tandem_repeats, key=lambda x: (x[0], x[1] - x[0])):
        if not non_overlapping_repeats or repeat[0] >= non_overlapping_repeats[-1][1]:
            non_overlapping_repeats.append(repeat)

    total_length = sum(repeat[1] - repeat[0] for repeat in non_overlapping_repeats)

    return total_length / seq_length * 100
```

```
def main():
    if len(sys.argv) != 2:
        print("Usage: python tandemrepeatdetector.py <input_fasta_file>")
        sys.exit(1)

    input_fasta_file = sys.argv[1]

    fasta_sequences = read_fasta_sequences_from_file(input_fasta_file)

    # Generate a timestamped output file name
    timestamp = time.strftime("%Y%m%d_%H%M%S")
    output_file = f"tandem_repeat_output_{timestamp}.tsv"

    # Write the results to the output file
    with open(output_file, "w") as f:
        f.write("Accession Number\tTandem Repeat Percentage\n")
        for accession, sequence in fasta_sequences.items():
            percentage = tandem_repeat_percentage(sequence)
            f.write(f"{accession}\t{percentage:.2f}\n")

    print(f"Results have been saved to {output_file}")

if __name__ == "__main__":
    main()
```

peptidemasspredictor.py

```
import sys
import time
from Bio import SeqIO

def read_fasta_sequences_from_file(input_file: str) -> dict:
    sequences = {}
    with open(input_file, "r") as f:
        for record in SeqIO.parse(f, "fasta"):
            # Split the record ID and take the second part, assuming 'tr|Accession|Description' format
```

```
            accession = record.id.split('|')[1] if '|' in record.id else record.id
            sequences[accession] = str(record.seq)
    return sequences

def tandem_repeat_percentage(sequence: str, min_repeat_length: int = 2, max_repeat_length: int = 12) -> float:
    tandem_repeats = []
    seq_length = len(sequence)

    for length in range(min_repeat_length, max_repeat_length + 1):
        for i in range(seq_length - 2 * length + 1):
            if sequence[i:i + length] == sequence[i + length:i + 2 * length]:
                repeat_start = i
                repeat_end = i + 2 * length
                while sequence[repeat_start:repeat_end] == sequence[repeat_end:repeat_end + length]:
                    repeat_end += length
                tandem_repeats.append((repeat_start, repeat_end))

    non_overlapping_repeats = []
    for repeat in sorted(tandem_repeats, key=lambda x: (x[0], x[1] - x[0])):
        if not non_overlapping_repeats or repeat[0] >= non_overlapping_repeats[-1][1]:
            non_overlapping_repeats.append(repeat)

    total_length = sum(repeat[1] - repeat[0] for repeat in non_overlapping_repeats)

    return total_length / seq_length * 100

def main():
    if len(sys.argv) != 2:
        print("Usage: python tandemrepeatdetector.py <input_fasta_file>")
        sys.exit(1)

    input_fasta_file = sys.argv[1]

    fasta_sequences = read_fasta_sequences_from_file(input_fasta_file)

    # Generate a timestamped output file name
    timestamp = time.strftime("%Y%m%d_%H%M%S")
```

```
    output_file = f"tandem_repeat_output_{timestamp}.tsv"

    # Write the results to the output file
    with open(output_file, "w") as f:
        f.write("Accession Number\tTandem Repeat Percentage\n")
        for accession, sequence in fasta_sequences.items():
            percentage = tandem_repeat_percentage(sequence)
            f.write(f"{accession}\t{percentage:.2f}\n")

    print(f"Results have been saved to {output_file}")

if __name__ == "__main__":
    main()
```

massmatcher.py

```
import sys
import time
from Bio import SeqIO

def read_fasta_sequences_from_file(input_file: str) -> dict:
    sequences = {}
    with open(input_file, "r") as f:
        for record in SeqIO.parse(f, "fasta"):
            # Split the record ID and take the second part, assuming 'tr|Accession|Description' format
            accession = record.id.split('|')[1] if '|' in record.id else record.id
            sequences[accession] = str(record.seq)
    return sequences

def tandem_repeat_percentage(sequence: str, min_repeat_length: int = 2, max_repeat_length: int = 12) -> float:
    tandem_repeats = []
    seq_length = len(sequence)

    for length in range(min_repeat_length, max_repeat_length + 1):
        for i in range(seq_length - 2 * length + 1):
            if sequence[i:i + length] == sequence[i + length:i + 2 * length]:
                repeat_start = i
```

```
                repeat_end = i + 2 * length
                while sequence[repeat_start:repeat_end] == sequence[repeat_end:repeat_end + length]:
                    repeat_end += length
                tandem_repeats.append((repeat_start, repeat_end))

    non_overlapping_repeats = []
    for repeat in sorted(tandem_repeats, key=lambda x: (x[0], x[1] - x[0])):
        if not non_overlapping_repeats or repeat[0] >= non_overlapping_repeats[-1][1]:
            non_overlapping_repeats.append(repeat)

    total_length = sum(repeat[1] - repeat[0] for repeat in non_overlapping_repeats)

    return total_length / seq_length * 100

def main():
    if len(sys.argv) != 2:
        print("Usage: python tandemrepeatdetector.py <input_fasta_file>")
        sys.exit(1)

    input_fasta_file = sys.argv[1]

    fasta_sequences = read_fasta_sequences_from_file(input_fasta_file)

    # Generate a timestamped output file name
    timestamp = time.strftime("%Y%m%d_%H%M%S")
    output_file = f"tandem_repeat_output_{timestamp}.tsv"

    # Write the results to the output file
    with open(output_file, "w") as f:
        f.write("Accession Number\tTandem Repeat Percentage\n")
        for accession, sequence in fasta_sequences.items():
            percentage = tandem_repeat_percentage(sequence)
            f.write(f"{accession}\t{percentage:.2f}\n")

    print(f"Results have been saved to {output_file}")
if __name__ == "__main__":
    main()
```

Appendix C.4 – Monte Carlo analysis of abiotic CXXC motif probability

Monte Carlo Simulation

```
import numpy as np
from collections import Counter
import time
from concurrent.futures import ThreadPoolExecutor, as_completed

def simulate_peptide_synthesis(num_peptides, peptide_length, aa_probs, reactivity_factors):
    # Adjust probabilities based on reactivity factors
    adjusted_probs = {k: aa_probs[k] * reactivity_factors.get(k, 1) for k in aa_probs}

    # Normalize adjusted probabilities to sum to 1
    total_p = sum(adjusted_probs.values())
    adjusted_probs = {k: v / total_p for k, v in adjusted_probs.items()}

    # Amino acids list and their probabilities
    aa_list = list(adjusted_probs.keys())
    aa_probs = list(adjusted_probs.values())

    # Function to simulate peptide synthesis
    def simulate_peptide():
        return ''.join(np.random.choice(aa_list, size=peptide_length, p=aa_probs))

    # Generate peptides
    peptides = [simulate_peptide() for _ in range(num_peptides)]

    return peptides

def count_cxxc_motifs(peptide_counts):
    cxxc_count = 0
    c_count = 0
    for peptide, count in peptide_counts.items():
        if 'C' in peptide:
            c_count += count
        for i in range(len(peptide) - 3):
            if peptide[i] == 'C' and peptide[i+3] == 'C':
                cxxc_count += count
                break
    return cxxc_count, c_count

# Parameters
num_peptides = 5000000  # Number of peptides per run
peptide_length = 10
num_runs = 50  # Number of runs to ensure 250 million sequences per concentration

# Base amino acid probabilities (lower bound without cysteine)
aa_probs_base = {
    "G": 0.41,
    "A": 0.24,
    "D": 0.12,
    "E": 0.10,
    "V": 0.04,
    "S": 0.02,
    "I": 0.02,
    "L": 0.01,
    "P": 0.03,
    "T": 0.01
}

# Cysteine concentrations to test
cysteine_concentrations = np.linspace(0.001, 0.1, 10)  # From 0.1% to 10%

# Reactivity factors (Cysteine is 7.8 times more reactive)
reactivity_factors = {
    "G": 1,
    "A": 1,
    "D": 1.68,  # Adjusted aspartic acid reactivity
    "E": 1,
    "V": 1,
```

```
    "S": 1,
    "I": 1,
    "L": 1,
    "P": 1,
    "T": 1,
    "C": 7.8
}

# No reactivity factors
no_reactivity_factors = {k: 1 for k in aa_probs_base.keys()}
no_reactivity_factors["C"] = 1

# Store results
results = []

def run_simulation(cysteine_concentration):
    aa_probs = aa_probs_base.copy()
    aa_probs["C"] = cysteine_concentration

    # Normalize the probabilities to sum to 1
    total_prob = sum(aa_probs.values())
    aa_probs = {k: v / total_prob for k, v in aa_probs.items()}

    all_peptide_counts_no_reactivity = Counter()
    all_peptide_counts_with_reactivity = Counter()

    for run in range(num_runs):
        print(f"Running simulation for cysteine concentration {cysteine_concentration:.4f}, run {run + 1}/{num_runs}")

        # Simulate peptide synthesis without reactivity factors
        peptides_no_reactivity = simulate_peptide_synthesis(num_peptides, peptide_length, aa_probs,
no_reactivity_factors)

        # Count occurrences of each unique peptide
        peptide_counts_no_reactivity = Counter(peptides_no_reactivity)

        # Aggregate counts across all runs
        all_peptide_counts_no_reactivity.update(peptide_counts_no_reactivity)

        # Simulate peptide synthesis with reactivity factors
        peptides_with_reactivity = simulate_peptide_synthesis(num_peptides, peptide_length, aa_probs, reactivity_factors)

        # Count occurrences of each unique peptide
        peptide_counts_with_reactivity = Counter(peptides_with_reactivity)

        # Aggregate counts across all runs
        all_peptide_counts_with_reactivity.update(peptide_counts_with_reactivity)

    # Count motifs and cysteine occurrences
    count_no_reactivity, c_count_no_reactivity = count_cxxc_motifs(all_peptide_counts_no_reactivity)
    count_with_reactivity, c_count_with_reactivity = count_cxxc_motifs(all_peptide_counts_with_reactivity)

    # Calculate probabilities
    prob_no_reactivity_cxxc = count_no_reactivity / (num_peptides * num_runs)
    prob_with_reactivity_cxxc = count_with_reactivity / (num_peptides * num_runs)
    prob_no_reactivity_c = c_count_no_reactivity / (num_peptides * num_runs)
    prob_with_reactivity_c = c_count_with_reactivity / (num_peptides * num_runs)

    result = {
        "cysteine_concentration": cysteine_concentration,
        "prob_no_reactivity_cxxc": prob_no_reactivity_cxxc,
        "prob_with_reactivity_cxxc": prob_with_reactivity_cxxc,
        "prob_no_reactivity_c": prob_no_reactivity_c,
        "prob_with_reactivity_c": prob_with_reactivity_c
    }

    output_str = (f"Cysteine concentration {cysteine_concentration:.4f} completed.\n"
                  f"Probability of CXXC motif (no reactivity): {prob_no_reactivity_cxxc:.4%}\n"
                  f"Probability of CXXC motif (with reactivity): {prob_with_reactivity_cxxc:.4%}\n"
                  f"Probability of at least one C (no reactivity): {prob_no_reactivity_c:.4%}\n"
                  f"Probability of at least one C (with reactivity): {prob_with_reactivity_c:.4%}\n")

    return result, output_str
```

```
# Using ThreadPoolExecutor for multithreading
with ThreadPoolExecutor(max_workers=6) as executor:
    future_to_concentration = {executor.submit(run_simulation, conc): conc for conc in cysteine_concentrations}
    with open('cysteine_concentration_results.txt', 'w') as output_file:
        for future in as_completed(future_to_concentration):
            conc = future_to_concentration[future]
            try:
                result, output_str = future.result()
                results.append(result)
                print(output_str)
                output_file.write(output_str + "\n")
            except Exception as exc:
                print(f"An error occurred for concentration {conc:.4f}: {exc}")
# Print the final results
with open('cysteine_concentration_final_results.txt', 'w') as f:
    for result in results:
        final_output = (f"Cysteine concentration: {result['cysteine_concentration']:.4f}\n"
                        f"Probability of CXXC motif (no reactivity): {result['prob_no_reactivity_cxxc']:.4%}\n"
                        f"Probability of CXXC motif (with reactivity): {result['prob_with_reactivity_cxxc']:.4%}\n"
                        f"Probability of at least one C (no reactivity): {result['prob_no_reactivity_c']:.4%}\n"
                        f"Probability of at least one C (with reactivity): {result['prob_with_reactivity_c']:.4%}\n")
        print(final_output)
        f.write(final_output + "\n")
```

Bibliography

1. Abraham A, Mihaliuk E, Kumar B, Legleiter J and Gullion T (2010) Solid-State NMR Study of Cysteine on Gold Nanoparticles. J. Phys. Chem. C. 114, 18109–18114.
2. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung C, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Žídek A, Bapst V, Kohli P, Jaderberg M, Hassabis D and Jumper JM (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 630, 493–500.
3. Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P (2002) The Universal Features of Cells on Earth. Molecular Biology of the Cell. 4th edition. Garland Science
4. Allwood AC, Walter MR, Burch IW and Kamber BS (2007) 3.43-billion-year-old stromatolite reef from the Pilbara Craton of Western Australia: Ecosystem-scale insights to early life on Earth. Precambrian Research. 158, 198–227.
5. Almeida AV, Jacinto JP, Guerra JPL, Vieira BJC, Waerenborgh JC, Jones NC, Hoffmann SV, Pereira AS and Tavares P (2021) Structural features and stability of apo- and holo-forms of a simple iron-sulfur protein. Eur Biophys J. 50, 561–570.
6. Altman S (1990) Enzymatic cleavage of RNA by RNA. Biosci Rep. 10, 317–337.
7. Alva V, Söding J and Lupas AN (2015) A vocabulary of ancient peptides at the origin of folded proteins. eLife. 4, e09410.
8. Amend A and Shock S (1998) Energetics of Amino Acid Synthesis in Hydrothermal Ecosystems. Science. 281, 1659–1662.
9. Anfinsen CB, Haber E, Sela M and White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A. 47, 1309–1314.
10. Bada JL (2013) New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. Chem. Soc. Rev. 42, 2186–2196.
11. Barge LM, Flores E, Baum MM, VanderVelde DG and Russell MJ (2019) Redox and pH gradients drive amino acid synthesis in iron oxyhydroxide mineral systems. Proceedings of the National Academy of Sciences. 116, 4828–4833.
12. Benner SA, Devine KG, Matveeva LN and Powell DH (2000) The missing organic molecules on Mars. Proc Natl Acad Sci U S A. 97, 2425–2430.
13. Berg JM (1990) Zinc fingers and other metal-binding domains. Elements for interactions between macromolecules. J Biol Chem. 265, 6513–6516.
14. Bernal JD (1949) The Physical Basis of Life. Proc. Phys. Soc. B. 62, 597.
15. Bernhardt HS (2012) The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). Biology direct. 7, 23.

16. Bernstein MP, Dworkin JP, Sandford SA, Cooper GW and Allamandola LJ (2002) Racemic amino acids from the ultraviolet photolysis of interstellar ice analogues. Nature. 416, 401–403.
17. Bertini I, Kurtz Jr. DM, Eidsness MK, Liu G, Luchinat C, Rosato A and Scott RA (1998) Solution structure of reduced Clostridium pasteurianum rubredoxin. JBIC Journal of Biological Inorganic Chemistry. 3, 401–410.
18. Blyholder G and Bowen DO (1962) Infrared Spectra of Sulfur Compounds Adsorbed on Silica Supported Nickel . J.Phys.Chem. 66, 1288–1292.
19. Brack A (1993) Liquid water and the origin of life. Orig Life Evol Biosph. 23, 3–10.
20. Bromberg Y, Aptekmann AA, Mahlich Y, Cook L, Senn S, Miller M, Nanda V, Ferreiro DU and Falkowski PG (2022) Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. Science Advances.
21. Brotzel F and Mayr H (2007) Nucleophilicities of amino acids and peptides. Org. Biomol. Chem. 5, 3814–3820.
22. Brownlee D (2014) The Stardust Mission: Analyzing Samples from the Edge of the Solar System. Annual Review of Earth and Planetary Sciences. 42, 179–205.
23. Cable M, Neveu M, Hsu H and Hoehler T (2020) Enceladus. In: Planetary Astrobiology, ed. Meadows V, pp. 217–246, University of Arizona Press, Tucson.
24. Carr MH, Belton MJS, Chapman CR, Davies ME, Geissler P, Greenberg R, McEwen AS, Tufts BR, Greeley R, Sullivan R, Head JW, Pappalardo RT, Klaasen KP, Johnson TV, Kaufman J, Senske D, Moore J, Neukum G, Schubert G, Burns JA, Thomas P and Veverka J (1998) Evidence for a subsurface ocean on Europa. Nature. 391, 363–365.
25. Carr MH and Head JW (2003) Oceans on Mars: An assessment of the observational evidence and possible fate. J.Geophys.Res. 108, .
26. Catling DC and Zahnle KJ (2020) The Archean atmosphere. Science Advances. 6, eaax1420.
27. Cech TR (1990) Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. Bioscience Reports. 10, 239–261.
28. Chen Y and Ma W (2020) The origin of biological homochirality along with the origin of life. PLOS Computational Biology. 16, e1007592.
29. Chino M, Di Costanzo LF, Leone L, La Gatta S, Famulari A, Chiesa M, Lombardi A and Pavone V (2023) Designed Rubredoxin miniature in a fully artificial electron chain triggered by visible light. Nature Communications. 14, 2368.
30. Choblet G, Tobie G, Sotin C, Běhounková M, Čadek O, Postberg F and Souček O (2017) Powering prolonged hydrothermal activity inside Enceladus. Nat Astron. 1, 841–847.
31. Christensen H and Pain RH (1991) Molten globule intermediates and protein folding. European Biophysics Journal. 19, 221–229.
32. Chyba C and Sagan C (1992) Endogenous production, exogenous delivery, and impact-shock synthesis of organic molecules: an inventory for the origins of life. Nature. 355, 125–132.

33. Cleaves HJ, Aubrey AD and Bada JL (2009) An Evaluation of the Critical Parameters for Abiotic Peptide Synthesis in Submarine Hydrothermal Systems. Origins of Life and Evolution of Biospheres. 39, 109–126.
34. Cleaves HJ and Chalmers JH (2004) Extremophiles May Be Irrelevant to the Origin of Life. Astrobiology. 4, 1–9.
35. Cleaves HJ, Chalmers J, Lazcano A, Miller SL and Bada JL (2008) A Reassessment of Prebiotic Organic Synthesis in Neutral Planetary Atmospheres. Orig Life Evol Biosph. 38, 105–115.
36. Clifford SM and Parker TJ (2001) The Evolution of the Martian Hydrosphere: Implications for the Fate of a Primordial Ocean and the Current State of the Northern Plains. Icarus. 154, 40–79.
37. Cockell CS (2014) Trajectories of Martian Habitability. Astrobiology. 14, 182–203.
38. Coker JA (2019) Recent advances in understanding extremophiles. F1000Res. 8, F1000 Faculty Rev–1917.
39. Cottrell TL (1958) The strengths of chemical bonds
40. Criado-Reyes J, Bizzarri BM, García-Ruiz JM, Saladino R and Di Mauro E (2021) The role of borosilicate glass in Miller–Urey experiment. Sci Rep. 11, 21009.
41. Cronin JR, Cooper GW, and Pizzarello S (1995) Characteristics and formation of amino acids and hydroxy acids of the Murchison meteorite. Advances in Space Research. 15, 91–97.
42. Crooks GE, Hon G, Chandonia J and Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.
43. Damer B and Deamer D (2020) The Hot Spring Hypothesis for an Origin of Life. Astrobiology. 20, 429–452.
44. DasSarma S and DasSarma P (2015) Halophiles and their enzymes: Negativity put to good use. Curr Opin Microbiol. 25, 120–126.
45. De Gregorio BT, Stroud RM, Nittler LR, Alexander CMO, Kilcoyne ALD and Zega TJ (2010) Isotopic anomalies in organic nanoglobules from Comet 81P/Wild 2: Comparison to Murchison nanoglobules and isotopic anomalies induced in terrestrial organics by electron irradiation. Geochimica et Cosmochimica Acta. 74, 4454–4470.
46. Deamer D and Damer B (2017) Can Life Begin on Enceladus? A Perspective from Hydrothermal Chemistry. Astrobiology. 17, 834–839.
47. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J and Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 6, 277–293.
48. Deng M, Yu J and Blackmond DG (2024) Symmetry breaking and chiral amplification in prebiotic ligation reactions. Nature. 626, 1019–1024.
49. Des Marais DJ, Nuth JA, Allamandola LJ, Boss AP, Farmer JD, Hoehler TM, Jakosky BM, Meadows VS, Pohorille A, Runnegar B and Spormann AM (2008) The NASA Astrobiology Roadmap. Astrobiology. 8, 715–730.
50. Di Achille G and Hynek BM (2010) Ancient Ocean on Mars supported by global distribution of deltas and valleys. Nature Geosci. 3, 459–463.
51. Di Mauro E, Dunker AK, and Trifonov EN (2012) Disorder to Order, Nonlife to Life: In the Beginning There Was a Mistake. In: Genesis - In the

Beginning: Precursors of Life, Chemical Models and Early Biological Evolution, ed. Seckbach J, pp. 415–435, Springer Netherlands, Dordrecht.
52. Dokmanić I, Sikić M and Tomić S (2008) Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr D Biol Crystallogr. 64, 257–263.
53. Duan C and Wang R (2024) A Unified Description of Salt Effects on the Liquid–Liquid Phase Separation of Proteins. ACS Cent. Sci. 10, 460–468.
54. Edwards H, Abeln S and Deane CM (2013) Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. PLOS Computational Biology. 9, e1003325.
55. Ehlmann BL, Mustard JF and Murchie SL (2010) Geologic setting of serpentine deposits on Mars. Geophys.Res.Lett. 37, .
56. Elsila JE, Aponte JC, Blackmond DG, Burton AS, Dworkin JP and Glavin DP (2016) Meteoritic Amino Acids: Diversity in Compositions Reflects Parent Body Histories. ACS Cent. Sci. 2, 370–379.
57. Elsila JE, Glavin DP and Dworkin JP (2009) Cometary glycine detected in samples returned by Stardust. Meteoritics & Planetary Science. 44, 1323–1330.
58. Erastova V, Degiacomi MT, Fraser D,G. and Greenwell HC (2017) Mineral surface chemistry control for origin of prebiotic peptides. Nat Commun. 8, 1–9.
59. Fairén AG, Fernández-Remolar D, Dohm JM, Baker VR and Amils R (2004) Inhibition of carbonate synthesis in acidic oceans on early Mars. Nature. 431, 423–426.
60. Fan K and Wang W (2003) What is the Minimum Number of Letters Required to Fold a Protein? Journal of Molecular Biology. 328, 921–926.
61. Farías-Rico JA and Mourra-Díaz CM (2022) A Short Tale of the Origin of Proteins and Ribosome Evolution. Microorganisms. 10, 2115.
62. Fitz D, Jakschitz T and Rode BM (2011) Salt-Induced Peptide Formation in Chemical Evolution: Building Blocks Before RNA – Potential of Peptide Splicing Reactions. In: Origins of Life: The Primal Self-Organization, eds. Egel R, Lankenau D and Mulkidjanian AY, pp. 109–127, Springer Berlin Heidelberg, Berlin, Heidelberg.
63. Flach CR, Gericke A and Mendelsohn R (1997) Quantitative Determination of Molecular Chain Tilt Angles in Monolayer Films at the Air/Water Interface: Infrared Reflection/Absorption Spectroscopy of Behenic Acid Methyl Ester. J Phys Chem B. 101, 58–65.
64. Flament N, Coltice N and Rey PF (2008) A case for late-Archaean continental emergence from thermal evolution models and hypsometry. Earth and Planetary Science Letters. 275, 326–336.
65. Flegmann AW and Tattersall R (1979) Energetics of peptide bond formation at elevated temperatures. J Mol Evol. 12, 349–355.
66. Fleminger G, Yaron T, Eisenstein M and Bar-Nun A (2005) The Structure and Synthetic Capabilities of a Catalytic Peptide Formed by Substrate-Directed Mechanism – Implications to Prebiotic Catalysis. Origins of Life and Evolution of Biospheres. 35, 369–382.
67. Flynn GJ (1996) The delivery of organic matter from asteroids and comets to the early surface of Mars. Earth Moon Planet. 72, 469–474.

68. Foden F, Islam I, Fernández-García, Maugeri M, Sheppard S and Powner P (2020) Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. Science. 370, 865–869.
69. Freissinet C, Glavin DP, Mahaffy PR, Miller KE, Eigenbrode JL, Summons RE, Brunner AE, Buch A, Szopa C, Archer Jr. PD, Franz HB, Atreya SK, Brinckerhoff WB, Cabane M, Coll P, Conrad PG, Des Marais DJ, Dworkin JP, Fairén AG, François P, Grotzinger JP, Kashyap S, ten Kate IL, Leshin LA, Malespin CA, Martin MG, Martin-Torres F, McAdam AC, Ming DW, Navarro-González R, Pavlov AA, Prats BD, Squyres SW, Steele A, Stern JC, Sumner DY, Sutter B, Zorzano M- and the MSL ST (2015) Organic molecules in the Sheepbed Mudstone, Gale Crater, Mars. J.Geophys.Res.Planets. 120, 495–514.
70. Frenkel-Pinter M, Samanta M, Ashkenasy G and Leman LJ (2020) Prebiotic Peptides: Molecular Hubs in the Origin of Life. Chem.Rev. 120, 4707–4765.
71. Fried SD, Fujishima K, Makarov M, Cherepashuk I and Hlouchova K (2022a) Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. Journal of the Royal Society, Interface. 19, 20210641.
72. Friedrich D, Perodeau J, Nieuwkoop AJ and Oschkinat H (2020) MAS NMR detection of hydrogen bonds for protein secondary structure characterization. J.Biomol. NMR. 74, 247–256.
73. Fujikawa T, Sasamoto T, Zhao F, Yamagishi A and Akanuma S (2024) Comparative analysis of reconstructed ancestral proteins with their extant counterparts suggests primitive life had an alkaline habitat. Sci Rep. 14, 398.
74. Gamaleldien H, Wu L, Olierook HKH, Kirkland CL, Kirscher U, Li Z, Johnson TE, Makin S, Li Q, Jiang Q, Wilde SA, and Li X (2024) Onset of the Earth's hydrological cycle four billion years ago or earlier. Nat. Geosci. 17, 560–565.
75. Giacobelli VG, Fujishima K, Lepšík M, Tretyachenko V, Kadavá T, Makarov M, Bednárová L, Novák P and Hlouchová K (2022) In Vitro Evolution Reveals Noncationic Protein–RNA Interaction Mediated by Metal Ions. Mol.Biol.Evol. 39, msac032.
76. Gilbert W (1986) Origin of life: The RNA world. Nature. 319, 618–618.
77. Glavin DP, Elsila JE, Burton AS, Callahan MP, Dworkin JP, Hilts RW and Herd CDK (2012) Unusual nonterrestrial l-proteinogenic amino acid excesses in the Tagish Lake meteorite. Meteoritics & Planetary Science. 47, 1347–1364.
78. Glavin DP, Elsila JE, McLain HL, Aponte JC, Parker ET, Dworkin JP, Hill DH, Connolly Jr. HC, and Lauretta DS (2021) Extraterrestrial amino acids and L-enantiomeric excesses in the CM2 carbonaceous chondrites Aguas Zarcas and Murchison. Meteorit Planet Sci. 56, 148–173.
79. Glavin DP, McLain HL, Dworkin JP, Parker ET, Elsila JE, Aponte JC, Simkus DN, Pozarycki CI, Graham HV, Nittler LR and Alexander CMO' (2020) Abundant extraterrestrial amino acids in the primitive CM carbonaceous chondrite Asuka 12236. Meteorit Planet Sci. 55, 1979–2006.
80. Glein CR, Baross JA and Waite JH (2015) The pH of Enceladus' Ocean. Geochimica et Cosmochimica Acta. 162, 202–219.

81. Glein CR and Waite JH (2020) The Carbonate Geochemistry of Enceladus' Ocean. Geophys.Res.Lett. 47, e2019GL085885.
82. Goldford JE, Hartman H, Marsland R and Segrè D (2019) Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. Nat Ecol Evol. 3, 1715–1724.
83. Gondim A, Antunes JMA, Lopes L, Carepo M, Pauleta S and Moura J (2019) Metal substitutions of rubredoxin metal cluster, stability studies and structural implications.
84. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc. 1, 2876–2890.
85. Grotzinger JP and Kasting JF (1993) New constraints on Precambrian Ocean composition. J Geol. 101, 235–243.
86. Hao J, Glein CR, Huang F, Yee N, Catling DC, Postberg F, Hillier JK and Hazen RM (2022) Abundant phosphorus expected for possible life in Enceladus's Ocean. Proceedings of the National Academy of Sciences. 119, e2201388119.
87. Hao J, Sverjensky DA and Hazen RM (2019) Redox states of Archean surficial environments: The importance of H2,g instead of O2,g for weathering reactions. Chemical Geology. 521, 49–58.
88. Harel A, Bromberg Y, Falkowski PG and Bhattacharya D (2014) Evolutionary history of redox metal-binding domains across the tree of life. Proceedings of the National Academy of Sciences. 111, 7042–7047.
89. Harrison SA, Rammu H, Liu F, Halpern A, Palmeira RN, and Lane N (2023) Life as a Guide to Its Own Origins. Annual Review of Ecology, Evolution, and Systematics. 54, 327–350.
90. Hazen RM, Filley TR and Goodfriend GA (2001) Selective adsorption of l- and d-amino acids on calcite: Implications for biochemical homochirality. Proceedings of the National Academy of Sciences. 98, 5487–5490.
91. Hazen RM and Sverjensky DA (2010) Mineral Surfaces, Geochemical Complexities, and the Origins of Life. Cold Spring Harb Perspect Biol. 2, a002162.
92. Hecht MH, Kounaves SP, Quinn RC, West SJ, Young SMM, Ming DW, Catling DC, Clark BC, Boynton WV, Hoffman J, Deflores LP, Gospodinova K, Kapit J and Smith PH (2009) Detection of perchlorate and the soluble chemistry of martian soil at the Phoenix lander site. Science. 325, 64–67.
93. Hendrix AR and House CH (2023) Low effective ultraviolet exposure ages for organics at the surface of Enceladus. Commun Earth Environ. 4, 1–8.
94. Hidese R, Mihara H, Kurihara T and Esaki N (2014) Global identification of genes affecting iron-sulfur cluster biogenesis and iron homeostasis. J.Bacteriol. 196, 1238–1249.
95. Higgs PG and Pudritz RE (2009) A Thermodynamic Basis for Prebiotic Amino Acid Synthesis and the Nature of the First Genetic Code. Astrobiology. 9, 483–490.
96. Holland HD (2006) The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society B: Biological Sciences. 361, 903–915.

97. Honniball CI, Lucey PG, Li S, Shenoy S, Orlando TM, Hibbitts CA, Hurley DM and Farrell WM (2021) Molecular water detected on the sunlit Moon by SOFIA. Nat Astron. 5, 121–127.
98. Hsu H, Postberg F, Sekine Y, Shibuya T, Kempf S, Horányi M, Juhász A, Altobelli N, Suzuki K, Masaki Y, Kuwatani T, Tachibana S, Sirono S, Moragas-Klostermeyer G and Srama R (2015) Ongoing hydrothermal activities within Enceladus. Nature. 519, 207–210.
99. Hurowitz JA and McLennan SM (2007) A ~3.5 Ga record of water-limited, acidic weathering conditions on Mars. Earth and Planetary Science Letters. 260, 432–443.
100. Hwang J and Han J (2015) 57Fe-Mössbauer analysis of biological and synthetic Fe/S clusters. Food Agric Sci Technol. 1, 53–61.
101. Hynek BM, Beach M and Hoke MRT (2010) Updated global map of Martian valley networks and implications for climate and hydrologic processes. J.Geophys.Res. 115, .
102. Ioppolo S, Fedoseev G, Chuang K-, Cuppen HM, Clements AR, Jin M, Garrod RT, Qasim D, Kofman V, van Dishoeck EF and Linnartz H (2021) A non-energetic mechanism for glycine formation in the interstellar medium. Nat Astron. 5, 197–205.
103. Janda M, Morvova M, Machala Z and Morva I (2008) Study of plasma induced chemistry by DC discharges in CO2/N2/H2O mixtures above a water surface. Orig Life Evol Biosph. 38, 23–35.
104. Jiang L, Dziedzic P, Spacil Z, Zhao G, Nilsson L, Ilag LL and Córdova A (2014) Abiotic synthesis of amino acids and self-crystallization under prebiotic conditions. Scientific reports. 4, 6769.
105. Jocelyn PC (1967) The Standard Redox Potential of Cysteine-Cystine from the Thiol-Disulphide Exchange Reaction with Glutathione and Lipoic Acid. European Journal of Biochemistry. 2, 327–331.
106. Jordan SF, Ioannou I, Rammu H, Halpern A, Bogart LK, Ahn M, Vasiliadou R, Christodoulou J, Maréchal A and Lane N (2021) Spontaneous assembly of redox-active iron-sulfur clusters at low concentrations of cysteine. Nat Commun. 12, 1–14.
107. Joyce GF (2002) The antiquity of RNA-based evolution. Nature. 418, 214–221.
108. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P and Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589.
109. Kamber BS (2015) The evolving nature of terrestrial crust from the Hadean, through the Archaean, into the Proterozoic. Precambrian Research. 258, 48–82.
110. Kang W, Mittal T, Bire S, Campin J and Marshall J (2022) How does salinity shape ocean circulation and ice geometry on Enceladus and other icy satellites? Science Advances. 8, eabm4665.
111. Kasting JF (1993) Earth's early atmosphere. Science. 259, 920–926.

112. Kasting JF and Howard MT (2006) Atmospheric composition and climate on the early Earth. Philosophical Transactions of the Royal Society B: Biological Sciences. 361, 1733.
113. Kebukawa Y, Chan QHS, Tachibana S, Kobayashi K and Zolensky ME (2017) One-pot synthesis of amino acid precursors with insoluble organic matter in planetesimals with aqueous activity. Science Advances. 3, e1602093.
114. Keller R (2004) The Computer Aided Resonance Assignment Tutorial. Cantina Verlag.
115. Kerber L, Forget F and Wordsworth R (2015) Sulfur in the early martian atmosphere revisited: Experiments with a 3-D Global Climate Model. Icarus. 261, 133–148.
116. Khawaja N, Postberg F, Hillier J, Klenner F, Kempf S, Nölle L, Reviol R, Zou Z and Srama R (2019) Low-mass nitrogen-, oxygen-bearing, and aromatic compounds in Enceladean ice grains. Monthly Notices of the Royal Astronomical Society. 489, 5231–5243.
117. Khurana KK, Kivelson MG, Stevenson DJ, Schubert G, Russell CT, Walker RJ, and Polanskey C (1998) Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto. Nature. 395, 777–780.
118. Kitadai N (2015) Energetics of Amino Acid Synthesis in Alkaline Hydrothermal Environments. Orig.Life Evol.Biosph. 45, 377–409.
119. Kitadai N and Maruyama S (2018) Origins of building blocks of life: A review. Geoscience Frontiers. 9, 1117–1153.
120. Kitadai N, Oonishi H, Umemoto K, Usui T, Fukushi K and Nakashima S (2017) Glycine Polymerization on Oxide Minerals. Origins of Life and Evolution of Biospheres. 47, 123–143.
121. Knauth LP (2005) Temperature and salinity history of the Precambrian Ocean: implications for the course of microbial evolution. Palaeogeography, Palaeoclimatology, Palaeoecology. 219, 53–69.
122. Knoll AH and Nowak MA (2017) The timetable of evolution. Science Advances. 3, e1603076.
123. Kocher CD and Dill KA (2024) Origins of life: The Protein Folding Problem all over again? Proceedings of the National Academy of Sciences. 121, e2315000121.
124. Koga T and Naraoka H (2017) A new family of extraterrestrial amino acids in the Murchison meteorite. Sci Rep. 7, 636.
125. Kozłowski H, Bal W, Dyba M and Kowalik-Jankowska T (1999) Specific structure–stability relations in metallopeptides. Coord.Chem.Rev. 184, 319–346.
126. Krissansen-Totton Joshua, Arney Giada N and Catling David C (2022) Constraining the climate and ocean pH of the early Earth with a geological carbon cycle model. Proceedings of the National Academy of Sciences. 115, 4105–4110.
127. Kwok S (2016) Complex organics in space from Solar System to distant galaxies. Astron Astrophys Rev. 24, 8.
128. Lahav N and Chang S (1976) The possible role of solid surface area in condensation reactions during chemical evolution: Reevaluation. J Mol Evol. 8, 357–380.
129. Lahav N and White DH (1980) A possible role of fluctuating clay-water systems in the production of ordered prebiotic oligomers. J Mol Evol. 16, 11–21.

130. Lambert J (2008) Adsorption and polymerization of amino acids on mineral surfaces: a review. Orig Life Evol Biosph. 38, 211–242.
131. Le Son H, Suwannachot Y, Bujdak J and Rode BM (1998) Salt-induced peptide formation from amino acids in the presence of clays and related catalysts. Inorganica Chimica Acta. 272, 89–94.
132. Le Vay K and Mutschler H (2019) The difficult case of an RNA-only origin of life. Emerging Topics in Life Sciences. 3, 469–475.
133. Lemke KH, Rosenbauer RJ and Bird DK (2009) Peptide Synthesis in Early Earth Hydrothermal Systems. Astrobiology. 9, 141–146.
134. Levy M and Miller SL (1998) The stability of the RNA bases: Implications for the origin of life. Proceedings of the National Academy of Sciences. 95, 7933–7938.
135. Liu C, Xu W, Zhang Z, Robinson K, Lau M, Huang F, Huang F, Glein CR and Hao J (2024) The Potential for Organic Synthesis in the Ocean of Enceladus. ApJ. 971, 51.
136. Liu J, Michalski JR, Wang Z and Gao W (2024) Atmospheric oxidation drove climate change on Noachian Mars. Nat Commun. 15, 5648.
137. Longo LM and Blaber M (2014) Prebiotic protein design supports a halophile origin of foldable proteins. Front. Microbiol. 0, .
138. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N and Tawfik DS (2020) Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. Proc.Natl.Acad.Sci.USA. 117, 15731.
139. Longo LM, Lee J and Blaber M (2013) Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc.Natl.Acad.Sci.U.S.A. 110, 2135–2139.
140. Lopes I, Piao L, Stievano L and Lambert J (2009) Adsorption of Amino Acids on Oxide Supports: A Solid-State NMR Study of Glycine Adsorption on Silica and Alumina. J.Phys.Chem.C. 113, 18163–18172.
141. Lukács M, Csilla Pálinkás D, Szunyog G and Várnagy K (2021) Metal Binding Ability of Small Peptides Containing Cysteine Residues. ChemistryOpen. 10, 451–463.
142. Maher M, Cross M, Wilce MCJ, Guss JM and Wedd AG (2004) Metal-substituted derivatives of the rubredoxin from Clostridium pasteurianum. Acta Crystallogr D Biol Crystallogr. 60, 298–303.
143. Maiti BK, Almeida RM, Moura I and Moura JJG (2017) Rubredoxins derivatives: Simple sulphur-rich coordination metal sites and its relevance for biology and chemistry. Coord.Chem.Rev. 352, 379–397.
144. Marshall WL (1994) Hydrothermal synthesis of amino acids. Geochim.Cosmochim.Acta. 58, 2099–2106.
145. Martin W, Baross J, Kelley D and Russell MJ (2008a) Hydrothermal vents and the origin of life. Nature Reviews Microbiology. 6, 805–814.
146. Martin W and Russell MJ (2006) On the origin of biochemistry at an alkaline hydrothermal vent. Philosophical Transactions of the Royal Society B: Biological Sciences. 362, 1887.
147. Martin W and Russell MJ (2003) On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic

prokaryotes, and from prokaryotes to nucleated cells. Philosophical Transactions of the Royal Society B: Biological Sciences. 358, 59.
148. Matsue S, Fujii T and Miyawaki O (2001) Effects of water activity and aqueous solvent ordering on thermal stability of lysozyme, α-chymotrypsinogen A, and alcohol dehydrogenase. International Journal of Biological Macromolecules. 28, 343–349.
149. Maury CPJ (2018) Amyloid and the origin of life: self-replicating catalytic amyloids as prebiotic informational and protometabolic entities. Cell. Mol. Life Sci. 75, 1499–1507.
150. Mayer-Bacon C, Meringer M, Havel R, Aponte JC, and Freeland S (2022) A Closer Look at Non-random Patterns Within Chemistry Space for a Smaller, Earlier Amino Acid Alphabet. J.Mol.Evol. 90, 307–323.
151. McDonald GD and Storrie-Lombardi M (2010) Biochemical Constraints in a Protobiotic Earth Devoid of Basic Amino Acids: The “BAA(-) World”. Astrobiology. 10, 989–1000.
152. McKay CP (2014) Requirements and limits for life in the context of exoplanets. Proc.Natl.Acad.Sci.USA. 111, 12628.
153. Merino N, Aronson HS, Bojanova DP, Feyhl-Buska J, Wong ML, Zhang S and Giovannelli D (2019) Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. Front.Microbiol. 0, .
154. Micsonai A, Moussong É, Wien F, Boros E, Vadászi H, Murvai N, Lee Y, Molnár T, Réfrégiers M, Goto Y, Tantos Á and Kardos J (2022) BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Res. 50, W90–W98.
155. Miller S (1953) A Production of Amino Acids Under Possible Primitive Earth Conditions. Science. 117, 528–529.
156. Miller SL and Urey HC (1959) Organic Compound Synthesis on the Primitive Earth. Science. 130, 245–251.
157. Mirceta S, Signore AV, Burns JM, Cossins AR, Campbell KL and Berenbrink M (2013) Evolution of mammalian diving capacity traced by myoglobin net surface charge. Science. 340, 1234192.
158. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S and Steinegger M (2022) ColabFold: making protein folding accessible to all. Nat Methods. 19, 679–682.
159. Miyawaki O, Dozen M and Hirota K (2016) Cooperative hydration effect causes thermal unfolding of proteins and water activity plays a key role in protein stability in solutions. Journal of Bioscience and Bioengineering. 122, 203–207.
160. Mojarro A, Aponte JC, Dworkin JP, Glavin DP, Elsila JE, Connolly HC and Lauretta DS (2024) "Early OSIRIS-REx Mission Results from Standard and Wet Chemistry Pyrolysis of Samples Returned from Asteroid Bennu", , 2024/03/1, pp. 1219.
161. Monteux J, Andrault D, Guitreau M, Samuel H and Demouchy S (2020) A mushy Earth's mantle for more than 500 Myr after the magma ocean solidification. Geophys J Int. 221, 1165–1181.
162. Moody ERR, Álvarez-Carretero S, Mahendrarajah TA, Clark JW, Betts HC, Dombrowski N, Szánthó LL, Boyle RA, Daines S, Chen X, Lane N, Yang Z, Shields GA, Szöllősi GJ, Spang A, Pisani D, Williams TA, Lenton TM and

Donoghue PCJ (2024) The nature of the last universal common ancestor and its impact on the early Earth system. Nat Ecol Evol. 8, 1654–1666.
163. Moore EK, Jelen BI, Giovannelli D, Raanan H and Falkowski PG (2017) Metal availability and the expanding network of microbial metabolisms in the Archaean eon. Nature Geosci. 10, 629–636.
164. Morleo A, Bonomi F, Iametti S, Huang VW, and Kurtz DM (2010) Iron-Nucleated Folding of a Metalloprotein in High Urea: Resolution of Metal Binding and Protein Folding Events. Biochemistry. 49, 6627–6634.
165. Murphy LR, Wallqvist A and Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Engineering, Design and Selection. 13, 149–152.
166. Nainytė M, Müller F, Ganazzoli G, Chan C, Crisp A, Globisch D and Carell T (2020) Amino Acid Modified RNA Bases as Building Blocks of an Early Earth RNA-Peptide World. Chemistry – A European Journal. 26, 14856–14860.
167. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG and Degrado WF (2005) De novo design of a redox-active minimal rubredoxin mimic. J Am Chem Soc. 127, 5804–5805.
168. National Research Council (2007) The Limits of Organic Life in Planetary Systems. The National Academies Press, Washington, DC.
169. Neumann W and Kruse A (2019) Differentiation of Enceladus and Retention of a Porous Core. ApJ. 882, 47.
170. Oba Y, Takano Y, Furukawa Y, Koga T, Glavin DP, Dworkin JP and Naraoka H (2022) Identifying the wide diversity of extraterrestrial purine and pyrimidine nucleobases in carbonaceous meteorites. Nat Commun. 13, 2008.
171. Ohtomo Y, Kakegawa T, Ishida A, Nagase T and Rosing MT (2014) Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. Nature Geosci. 7, 25–28.
172. Oivanen M, Kuusela S and Lönnberg H (1998) Kinetics and Mechanisms for the Cleavage and Isomerization of the Phosphodiester Bonds of RNA by Brønsted Acids and Bases. Chem Rev. 98, 961–990.
173. Ojha L, Karunatillake S, Karimi S and Buffo J (2021) Amagmatic hydrothermal systems on Mars from radiogenic heat. Nat Commun. 12, 1754.
174. Orgel LE (1998) The origin of life—a review of facts and speculations. Trends in Biochemical Sciences. 23, 491–495.
175. Otake T, Taniguchi T, Furukawa Y, Kawamura F, Nakazawa H and Kakegawa T (2011) Stability of Amino Acids and Their Oligomerization Under High-Pressure Conditions: Implications for Prebiotic Chemistry. Astrobiology. 11, 799–813.
176. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, Hendricks MM, Iimura S, Gajiwala K, Scholtz JM and Grimsley GR (2011) Contribution of Hydrophobic Interactions to Protein Stability. J.Mol.Biol. 408, 514–528.
177. Pace NR (2001) The universal nature of biochemistry. Proceedings of the National Academy of Sciences of the United States of America. 98, 805.
178. Parker ET, McLain HL, Glavin DP, Dworkin JP, Elsila JE, Aponte JC, 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 and Nakamura T (2023) Extraterrestrial amino acids

and amines identified in asteroid Ryugu samples returned by the Hayabusa2 mission. Geochimica et Cosmochimica Acta. 347, 42–57.
179. Pavlov AA, McLain HL, Glavin DP, Roussel A, DworkIn JP, Elsila JE and Yocum KM (2022) Rapid Radiolytic Degradation of Amino Acids in the Martian Shallow Subsurface: Implications for the Search for Extinct Life. Astrobiology. 22, 1099–1115.
180. Pearson RG (1963) Hard and Soft Acids and Bases. J.Am.Chem.Soc. 85, 3533–3539.
181. Permyakov EA (2021) Metal Binding Proteins. Encyclopedia. 1, 292.
182. Peter JS, Nordheim TA, and Hand KP (2024) Detection of HCN and diverse redox chemistry in the plume of Enceladus. Nature Astronomy. 8, 164–173.
183. Petkowski JJ, Bains W and Seager S (2020) On the Potential of Silicon as a Building Block for Life. Life (Basel). 10, .
184. Piette BMAG and Heddle JG (2020) A Peptide–Nucleic Acid Replicator Origin for Life. Trends in ecology & evolution (Amsterdam). 35, 397–406.
185. Pillot G, Santiago Ó, Kerzenmacher S and Liebgott P (2023) Spark of Life: Role of Electrotrophy in the Emergence of Life. Life. 13, 356.
186. Piqueux S, Buz J, Edwards CS, Bandfield JL, Kleinböhl A, Kass DM, Hayne PO, The MCS and Teams T (2019) Widespread Shallow Water Ice on Mars at High Latitudesand Midlatitudes. Geophys.Res.Lett. 46, 14290–14298.
187. Pizzarello S, Huang Y and Alexandre MR (2008) Molecular asymmetry in extraterrestrial chemistry: Insights from a pristine meteorite. Proceedings of the National Academy of Sciences of the United States of America. 105, 3700.
188. Pizzarello S and Shock E (2010) The Organic Composition of Carbonaceous Meteorites: The Evolutionary Story Ahead of Biochemistry. Cold Spring Harb Perspect Biol. 2, a002105.
189. Plankensteiner K, Righi A, Rode BM, Gargallo R, Jaumot J and Tauler R (2004) Indications towards a stereoselectivity of the salt-induced peptide formation reaction. Inorganica Chimica Acta. 357, 649–656.
190. Pohorille A and Pratt LR (2012) Is Water the Universal Solvent for Life? Orig Life Evol Biosph. 42, 405–409.
191. Postberg F, Clark RN, Hansen CJ, Coates AJ, Dalle Ore CM, Scipioni F, Hedman MM and Waite JH (2018) Plume and Surface Composition of Enceladus. In: , pp. 129, .
192. Postberg F, Khawaja N, Abel B, Choblet G, Glein CR, Gudipati MS, Henderson BL, Hsu H, Kempf S, Klenner F, Moragas-Klostermeyer G, Magee B, Nölle L, Perry M, Reviol R, Schmidt J, Srama R, Stolz F, Tobie G, Trieloff M and Waite JH (2018) Macromolecular organic compounds from the depths of Enceladus. Nature. 558, 564–568.
193. Postberg F, Sekine Y, Klenner F, Glein CR, Zou Z, Abel B, Furuya K, Hillier JK, Khawaja N, Kempf S, Noelle L, Saito T, Schmidt J, Shibuya T, Srama R and Tan S (2023) Detection of phosphates originating from Enceladus's Ocean. Nature. 618, 489–493.
194. 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 and Nakamura E (2023) Insights into the formation and

evolution of extraterrestrial amino acids from the asteroid Ryugu. Nat Commun. 14, 1482.
195. Pozdnyakova I and Wittung-Stafshede P (2001) Biological Relevance of Metal Binding before Protein Folding. J. Am. Chem. Soc. 123, 10135–10136.
196. Pross A (2004) Causation and the origin of life. Metabolism or replication first? Orig Life Evol Biosph. 34, 307–321.
197. Reed CJ, Lewis H, Trejo E, Winston V and Evilia C (2013) Protein Adaptations in Archaeal Extremophiles. Archaea. 2013, e373275.
198. Rigden DJ and Galperin MY (2004) The DxDxDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. Journal of Molecular Biology. 343, 971–984.
199. Rimola A, Costa D, Sodupe M, Lambert J and Ugliengo P (2013) Silica surface features and their role in the adsorption of biomolecules: computational modeling and experiments. Chem Rev. 113, 4216–4313.
200. Rimola A, Sodupe M and Ugliengo P (2016) Amide and Peptide Bond Formation: Interplay between Strained Ring Defects and Silanol Groups at Amorphous Silica Surfaces. J.Phys.Chem.C. 120, 24817–24826.
201. Robinson KJ, Bockisch C, Gould IR, Liao Y, Yang Z, Glein CR, Shaver GD, Hartnett HE, Williams LB, and Shock EL (2021) Quantifying the extent of amide and peptide bond synthesis across conditions relevant to geologic and planetary environments. Geochim.Cosmochim.Acta. 300, 318–332.
202. Rode BM, Son HL and Suwannachot Y (1999) The combination of salt induced peptide formation reaction and clay catalysis: a way to higher peptides under primitive earth conditions. Orig Life Evol Biosph. 29, 273–286.
203. Rodriguez-Garcia M, Surman AJ, Cooper GJT, Suárez-Marina I, Hosni Z, Lee MP, and Cronin L (2015) Formation of oligopeptides in high yield under simple programmable conditions. Nature communications. 6, 8385.
204. Romero P, Obradovic Z and Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett. 462, 363–367.
205. Ross DS and Deamer D (2016) Dry/Wet Cycling and the Thermodynamics and Kinetics of Prebiotic Polymer Synthesis. Life. 6, 28.
206. Rout SK, Rhyner D, Riek R and Greenwald J (2022) Prebiotically Plausible Autocatalytic Peptide Amyloids. Chem.Eur.J. 28, e202103841.
207. Russell MJ (2021) The “Water Problem”(sic), the Illusory Pond and Life’s Submarine Emergence—A Review. Life. 11, .
208. Sagan C and Khare BN (1971) Long-Wavelength Ultraviolet Photoproduction of Amino Acids on the Primitive Earth. Science. 173, 417–420.
209. Sakata K, Kitadai N and Yokoyama T (2010) Effects of pH and temperature on dimerization rate of glycine: Evaluation of favorable environmental conditions for chemical evolution of life. Geochimica et Cosmochimica Acta. 74, 6841–6851.
210. Santos TCB and Futerman AH (2023) The fats of the matter: Lipids in prebiotic chemistry and in origin of life studies. Prog.Lipid Res. 92, 101253.
211. Saur J, Duling S, Roth L, Jia X, Strobel DF, Feldman PD, Christensen UR, Retherford KD, McGrath MA, Musacchio F, Wennmacher A, Neubauer FM, Simon S and Hartkorn O (2015) The search for a subsurface ocean in

Ganymede with Hubble Space Telescope observations of its auroral ovals. J.Geophys.Res.Space Physics. 120, 1715–1737.
212. Sauterey B, Charnay B, Affholder A, Mazevet S and Ferrière R (2022) Early Mars habitability and global cooling by H2-based methanogens. Nat Astron. 6, 1263–1271.
213. Schwieterman EW, Kiang NY, Parenteau MN, Harman CE, DasSarma S, Fisher TM, Arney GN, Hartnett HE, Reinhard CT, Olson SL, Meadows VS, Cockell CS, Walker SI, Grenfell JL, Hegde S, Rugheimer S, Hu R and Lyons TW (2018) Exoplanet Biosignatures: A Review of Remotely Detectable Signs of Life. Astrobiology. 18, 663–708.
214. Segré D, Ben-Eli D, Deamer DW and Lancet D (2001) The lipid world. Orig Life Evol Biosph. 31, 119–145.
215. Segura TL, Toon OB, and Colaprete A (2008) Modeling the environmental effects of moderate-sized impacts on Mars. J.Geophys.Res. 113, .
216. Sekine Y, Shibuya T, Postberg F, Hsu H, Suzuki K, Masaki Y, Kuwatani T, Mori M, Hong PK, Yoshizaki M, Tachibana S and Sirono S (2015) High-temperature water-rock interactions and hydrothermal environments in the chondrite-like core of Enceladus. Nature communications. 6, 8604.
217. Sen C, Logashree V, Makde RD and Ghosh B (2024) Amino acid propensities for secondary structures and its variation across protein structures using exhaustive PDB data. Computational Biology and Chemistry. 110, 108083.
218. Shalayel I, Youssef-Saliba S, Vazart F, Ceccarelli C, Bridoux M and Vallée Y (2020) Cysteine Chemistry in Connection with Abiogenesis. Eur.J.Org.Chem. 2020, 3019–3023.
219. Shanker U, Bhushan B, Bhattacharjee G and Kamaluddin (2012) Oligomerization of Glycine and Alanine Catalyzed by Iron Oxides: Implications for Prebiotic Chemistry. Origins of Life and Evolution of Biospheres. 42, 31–45.
220. Sharma A (2022) Enantiomeric excess of amino acids in interstellar ice analogues-asymmetric photolysis of precursors by circularly polarized UV light. Mon Not R Astron Soc. 517, 6112–6120.
221. Shock EL (1992) Stability of peptides in high-temperature aqueous solutions. Geochimica et Cosmochimica Acta. 56, 3481–3491.
222. Sibilska IK, Chen B, Li L and Yin J (2017) Effects of Trimetaphosphate on Abiotic Formation and Hydrolysis of Peptides. Life (Basel). 7, .
223. Siddiq MA, Hochberg GKA and Thornton JW (2017) Evolution of protein specificity: Insights from ancestral protein reconstruction. Curr Opin Struct Biol. 47, 113–122.
224. Speight JG (2018) Chapter 6 - Hydrolysis. In: Reaction Mechanisms in Environmental Engineering, ed. Speight JG, pp. 203–229, Butterworth-Heinemann, .
225. Squyres S, Arvidson A, Bell B, Brückner, Cabrol C, Calvin C, Carr C, Christensen C, Clark C, Crumpler C, Marais M, d'Uston, Economou E, Farmer F, Farrand F, Folkner F, Golombek G, Gorevan G, Grant G, Greeley G, Grotzinger G, Haskin H, Herkenhoff H, Hviid H, Johnson J, Klingelhöfer, Knoll K, Landis L, Lemmon L, Li L, Madsen M, Malin M, McLennan M, McSween M, Ming M, Moersch M, Morris M, Parker P, Rice R, Richter R, Rieder R, Sims S,

Smith S, Smith S, Soderblom S, Sullivan S, Wänke, Wdowiak W, Wolff W and Yen Y (2004) The Opportunity Rover's Athena Science Investigation at Meridiani Planum, Mars. Science. 306, 1698–1703.
226. Steel EL, Davila A and McKay CP (2017) Abiotic and Biotic Formation of Amino Acids in the Enceladus Ocean. Astrobiology. 17, 862–875.
227. Sterner R and Liebl W (2001) Thermophilic adaptation of proteins. Crit Rev Biochem Mol Biol. 36, 39–106.
228. Stewart JE (2019) The Origins of Life: The Managed-Metabolism Hypothesis. Found Sci. 24, 171–195.
229. Strecker A (1850) Ueber die künstliche Bildung der Milchsäure und einen neuen, dem Glycocoll homologen Körper; Justus Liebigs Ann.Chem. 75, 27–45.
230. Stüeken EE, Anderson RE, Bowman JS, Brazelton WJ, Colangelo-Lillis J, Goldman AD, Som SM and Baross JA (2013) Did life originate from a global chemical reactor? Geobiology. 11, 101–126.
231. Sun PD, Foster CE and Boyington JC (2004) Overview of Protein Structural and Functional Folds. Curr Protoc Protein Sci. 35, 1711–171189.
232. Sushko T, Kavaleuski A, Grabovec I, Kavaleuskaya A, Vakhrameev D, Bukhdruker S, Marin E, Kuzikov A, Masamrekh R, Shumyantseva V, Tsumoto K, Borshchevskiy V, Gilep A and Strushkevich N (2021) A new twist of rubredoxin function in M. tuberculosis. Bioorg.Chem. 109, 104721.
233. Tagami S and Li P (2023) The origin of life: RNA and protein co-evolution on the ancient Earth. Develop Growth Differ. 65, 167–174.
234. Takahagi W, Seo K, Shibuya T, Takano Y, Fujishima K, Saitoh M, Shimamura S, Matsui Y, Tomita M and Takai K (2019) Peptide Synthesis under the Alkaline Hydrothermal Conditions on Enceladus. ACS Earth Space Chem. 3, 2559–2568.
235. Talley K and Alexov E (2010) On the pH-optimum of activity and stability of proteins. Proteins. 78, 2699–2706.
236. Taubner R, Pappenreiter P, Zwicker J, Smrzka D, Pruckner C, Kolar P, Bernacchi S, Seifert AH, Krajete A, Bach W, Peckmann Jö, Paulik C, Firneis MG, Schleper C and Rittmann SK-R (2018) Biological methane production under putative Enceladus-like conditions. Nat Commun. 9, 1–11.
237. Tenelanda-Osorio L, Parra JL, Cuartas-Restrepo P and Zuluaga JI (2021) Enceladus as a Potential Niche for Methanogens and Estimation of Its Biomass. Life (Basel). 11, .
238. Tessera M (2018) Is pre-Darwinian evolution plausible? Biology Direct. 13, 18.
239. Tian Z, Magna T, Day JMD, Mezger K, Scherer EE, Lodders K, Hin RC, Koefoed P, Bloom H and Wang K (2021) Potassium isotope composition of Mars reveals a mechanism of planetary volatile retention. Proceedings of the National Academy of Sciences. 118, e2101155118.
240. Tjoa JNKY, Mueller M and Tak FFS (2020) The subsurface habitability of small, icy exomoons. A&A. 636, A50.
241. Tosca NJ, Fischer WW, Milliken RE and Hurowitz JA (2010) Origin of acidic surface waters and the evolution of atmospheric chemistry on early Mars. Nature geoscience. 3, 323–326.
242. Tosca NJ, Knoll AH and McLennan SM (2008) Water Activity and the Challenge for Life on Early Mars. Science. 320, 1204.

243. Tosca NJ and Tutolo BM (2023) Hydrothermal vent fluid-seawater mixing and the origins of Archean iron formation. Geochimica et Cosmochimica Acta. 352, 51–68.
244. Tostevin R and Ahmed IAM (2023) Micronutrient availability in Precambrian oceans controlled by greenalite formation. Nat. Geosci. 16, 1188–1193.
245. Tretyachenko V, Vymětal J, Neuwirthová T, Vondrášek J, Fujishima K and Hlouchová K (2022) Modern and prebiotic amino acids support distinct structural profiles in proteins. Open Biol. 12, 220040.
246. Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene. 261, 139–151.
247. Trigo-Rodríguez JM, Rimola A, Tanbakouei S, Soto VC and Lee M (2019) Accretion of Water in Carbonaceous Chondrites: Current Evidence and Implications for the Delivery of Water to Early Earth. Space Sci Rev. 215, 18.
248. Tuff J, Wade J and Wood BJ (2013) Volcanism on Mars controlled by early oxidation of the upper mantle. Nature. 498, 342–345.
249. Vallee Y, Shalayel I, Ly K, Rao KVR, De Paëpe G, Märker K and Milet A (2017) At the very beginning of life on Earth: the thiol-rich peptide (TRP) world hypothesis. The International Journal of Developmental Biology. 61, 471–478.
250. van der Gulik P,T.S. and Speijer D (2015) How amino acids and peptides shaped the RNA world. Life (Basel, Switzerland). 5, 230–246.
251. Van Kranendonk MJ, Philippot P, Lepot K, Bodorkos S and Pirajno F (2008) Geological setting of Earth's oldest fossils in the ca. 3.5 Ga Dresser Formation, Pilbara Craton, Western Australia. Precambrian Research. 167, 93–124.
252. Vázquez-Salazar A, Tan G, Stockton A, Fani R, Becerra A and Lazcano A (2017) Can an Imidazole Be Formed from an Alanyl-Seryl-Glycine Tripeptide under Possible Prebiotic Conditions? Origins of Life and Evolution of Biospheres. 47, 345–354.
253. Vieira AdN, Kleinermanns K, Martin WF and Preiner M (2020) The ambivalent role of water at the origins of life. FEBS Lett. 594, 2717–2733.
254. Wachnowsky C, Fidai I and Cowan JA (2018) Iron-sulfur cluster biosynthesis and trafficking - Impact on human disease conditions. Metallomics. 10, 9–29.
255. Wächtershäuser G (1992) Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog Biophys Mol Biol. 58, 85–201.
256. Wade J, Byrne DJ, Ballentine CJ and Drakesmith H (2021) Temporal variation of planetary iron as a driver of evolution. Proceedings of the National Academy of Sciences. 118, e2109865118.
257. Waite JH, Glein CR, Perryman RS, Teolis BD, Magee BA, Miller G, Grimes J, Perry ME, Miller KE, Bouquet A, Lunine JI, Brockwell T and Bolton SJ (2017) Cassini finds molecular hydrogen in the Enceladus plume: Evidence for hydrothermal processes. Science. 356, 155–159.
258. Waite J, Magee B, Brockwell T, Zolotov M, Teolis B and Lewis W (2011) Enceladus Plume Composition. LPI Contrib. 1774, 61.
259. Watenpaugh KD, Sieker LC, Herriott JR and Jensen LH (1972) The structure of a non-heme iron protein: rubredoxin at 1.5 Angstrom resolution. Cold Spring Harb Symp Quant Biol. 36, 359–367.

260. Watenpaugh K, Sieker L and Jensen H (1980) Crystallographic refinement of rubredoxin at 1·2 Å resolution. J.Mol.Biol. 138, 615–633.
261. Wilde SA, Valley JW, Peck WH and Graham CM (2001) Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. Nature. 409, 175–178.
262. Williamson MP (2013) Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spectrosc. 73, 1–16.
263. Wochner W, Attwater A, Coulson C and Holliger H (2011) Ribozyme-Catalyzed Transcription of an Active Ribozyme. Science. 332, 209–212.
264. Wogan NF, Catling DC, Zahnle KJ, and Lupu R (2023) Origin-of-life Molecules in the Atmosphere after Big Impacts on the Early Earth. Planet. Sci. J. 4, 169.
265. Wordsworth R (2016) The Climate of Early Mars. Annual Review of Earth and Planetary Sciences, 44, 381–408.
266. Xiao Z, Lavery MJ, Ayhan M, Scrofani SDB, Wilce MCJ, Guss JM, Tregloan PA, George GN and Wedd AG (1998) The Rubredoxin from Clostridium pasteurianum: Mutation of the Iron Cysteinyl Ligands to Serine. Crystal and Molecular Structures of Oxidized and Dithionite-Treated Forms of the Cys42Ser Mutant. J.Am.Chem.Soc. 120, 4135–4150.
267. Yadav M, Kumar R and Krishnamurthy R (2020) Chemistry of Abiotic Nucleotide Synthesis. Chem.Rev. 120, 4766–4805.
268. Zaia DAM, Zaia CTBV and De Santana H (2008) Which Amino Acids Should Be Used in Prebiotic Chemistry Studies? Orig.Life Evol.Biosph. 38, 469–488.
269. Zartler ER, Jenney FE, Terrell M, Eidsness MK, Adams MWW and Prestegard JH (2001) Structural Basis for Thermostability in Aporubredoxins from Pyrococcus furiosus and Clostridium pasteurianum. Biochemistry. 40, 7279–7290.