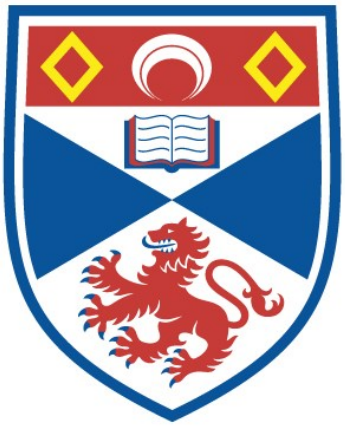
Exploring biosynthetic routes to novel histidine containing cyclodipeptides

Emmajay Sutherland

A thesis submitted for the degree of PhD
at the
University of St Andrews



2023

Candidate's declaration

I, Emmajay Sutherland, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 48,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2019.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date: 19/02/2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date: 19/02/2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Emmajay Sutherland, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

Embargo on all of print copy for a period of 3 years on the following ground(s):

- Publication would preclude future publication.

Supporting statement for printed embargo request

The data featured in this thesis will be included manuscripts currently being drafted for publication.

Electronic copy

Embargo on all of electronic copy for a period of 3 years on the following ground(s):

- Publication would preclude future publication.

Supporting statement for electronic embargo request

The data featured in this thesis will be included manuscripts currently being drafted for publication.

Title and Abstract

- I require an embargo on the abstract only.

Date: 19/02/2023 Signature of candidate

Date: 19/02/2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Emmajay Sutherland, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date: 19/02/2023 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date: 19/02/2023 Signature of candidate

Date: 19/02/2023 Signature of supervisor

Abstract

Histidine containing cyclodipeptides have gathered attention for their important biological potential as anti-cancer drugs and neuroprotective agents. Thus, the creation of facile synthetic routes to these compounds are essential to exploit their potential as therapeutic targets. Biosynthetic pathways of natural products recently revealed a family of enzymes called cyclodipeptide synthases (CDPSs) which use aminoacylated tRNAs as substrates to yield an array of cyclodipeptides. From the current set of characterised CDPSs however, only two CDPSs accept histidine as a substrate: *Para*CDPS and *Parcu*CDPS. Extensive research has led to thorough characterisation of the catalytic mechanism of CDPSs, however their exact substrate recognition and specificity determinants remain poorly understood.

Initial investigation of CDPSs demonstrated the inherent promiscuity of this class by generating a library of cyclodipeptides from canonical and non-canonical amino acids. By solving the structure of *Parcu*CDPS, key residues for substrate recognition were unveiled in the first binding pocket of the CDPS. Rationally engineered mutants subsequently displayed altered substrate scope, rejecting histidine, and accepting phenylalanine and leucine. This research is the first instance of successfully engineering a CDPS to yield multiple products and paves the way for directed modifications to enhance the promiscuity of these enzymes to produce molecules of our choosing. Furthermore, tailoring enzymes are counterparts to CDPSs in nature and create an additional layer of diversity to cyclodipeptides. Preliminary investigations sought to characterise a cyclodipeptide oxidase to identify the substrate specificity of these rather mysterious enzymes.

As a final observation, the work presented here complements the current understanding of biosynthetic pathways to novel cyclodipeptides. Cyclodipeptide synthases pose as a facile tailored route with the potential to generate a large range of cyclic products from just one enzyme. Additionally, the introduction of tailoring enzymes enhances the diversity possible from such a small family of enzymes.

Acknowledgements

First and foremost, I would like to thank Dr Clarissa Melo Czekster for agreeing to re-train a chemist to be a biochemist. Thank you for teaching me how to properly pipette on the first day of my PhD and for everything you have taught me since. I appreciate that this type of mentorship is rare and I cannot thank you enough for allowing me to be a scientist in the best possible way. I will always appreciate your unwavering optimism, enthusiasm for science, endless knowledge, and last (but certainly not least) all of the cake over these past 3.5 years.

A special thank you to everyone across the BSRC who has helped me in the lab throughout my time at the University of St Andrews. To Dr Rafael Guimaraes da Silva, thank you for allowing me to learn from you, for all the life advice and for always giving me an honest answer. To Dr Verena Oehler, thank you for always being kind, and I apologise that my cloning took over 4 months to work but you never once doubted it. To Dr Magnus Alphey, thank you for your helpful expertise on x-ray crystallography and looking after the temperamental robots. To Dr Bela Bode, thank you for the EPR training and helping me get promising results on an elusive radical. Thank you to Dr Sally Shirran and Dr Silvia Synowsky from the BSRC Mass spectrometry and proteomics facility for their help on this work, specifically the trapped acyl-enzyme intermediate experiments.

To the Czekster lab members, the office wouldn't be half as fun without you. A special mention to Dr Chris Harding for helping me achieve my first ever protein crystals and for always having an answer to my endless questions. My crystal fishing skills will never be as good as yours, but my music taste will always be better. Thank you to Martha for answering my silly questions, telling me if my figures look bad, showing me how to bake amazing cookies and always saying yes to weekend plans even when that meant we sat front row at the world's worst performance of the Fringe 2022. Thank you to Greice, Lynette and Thu for lovely lunchtime chats and mid-day breaks to get cake – good luck on your journeys to becoming a doctor. A mention to honorary members of the Czekster lab – Gemma and Teresa – for all their kind words and wise advice, especially during thesis write-up. And finally, thanks to the new and old PhD students of level 2 – Flora, Catriona, Shannon, Anna, Peter and Ben to name a few.

I'd like to extend my thanks to my friends from home for sticking around (ML1 represent!). Thank you to Craig for the immense number of voice notes and videos you have sent me to keep me going through good and bad days. Thank you to Caroline for always being just a message away and an eager gig partner. I appreciate both of your endless support and I can't wait until you visit me in sunny Seattle. Thank you to Dr Taylor Swift for releasing multiple albums over the past 3 years which have been on repeat during my PhD.

I'm especially grateful to my parents whose support has been unwavering since I started this university journey in 2014. Thank you for allowing me to pursue my goals even though this was an entirely new experience for all of us. You never doubted my capabilities and were always so supportive in more ways than I can describe, you both deserve the honorary title of doctor! I appreciate all the opportunities you have given me throughout my life, sometimes at the expense of your own - I can only hope to repay you in the future.

To Liam, firstly thank you for agreeing to move with me to Dundee in 2019. Who was to know that just a few months later, I'd be cutting your hair in the bathroom and feeding the nearby swans was our only hobby. Thank you for always being on my side and reminding me I can do this – there is no one I would rather have spent this time with (or moved house 4 times with!) All's well that ends well to end up with you and you certainly bring out the best in me. This is only the beginning and I can't wait to explore Seattle with you – go Seahawks!

The research underpinning this thesis received funding from the Cunningham Trust (Grant: PhD-CT-18-41).

Table of Contents

Table of Figures

Table of Tables

Abbreviations

aaRS	Aminoacyl-tRNA synthetase
aa-AMP	Aminoacyl-adenylate intermediate
aa-DBE	Amino acid-dinitrobenzyl ester
aa-tRNA	Amino acid-transfer ribonucleic acid
ACN	Acetonitrile
AMP	Adenosine 5'-monophosphate
APS	Ammonium persulfate
ATP	Adenosine 5'-triphosphate
AU	Absorbance unit
CAPS	*N*-cyclohexyl-3-aminopropanesulfonic acid
CASTp	Computed Atlas of Surface Topography of proteins
CDO	Cyclodipeptide oxidase
CDP	Cyclodipeptide
CDPS	Cyclodipeptide synthase
Cryo-EM	Cryo-electron microscopy
CTP	Cytidine 5'-triphosphate
CV	Column volume
Da	Dalton
DEPC	Diethyl pyrocarbonate
DLS	Dynamic light scattering
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
DBE	3,5-Dinitrobenzyl ester
DKP	2,5-Diketopiperazine
DSF	Differential scanning fluorimetry
DTT	Dithiothreitol
ε280	Extinction coefficient measured at 280 nm
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EIC	Extracted ion chromatogram
EPR	Electron paramagnetic resonance
ESI	Electrospray ionisation
FA	Formic acid
FPLC	Fast protein liquid chromatography
GTP	Guanosine 5'-triphosphate
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	High performance liquid chromatography
IDT	Integrated DNA Technologies
IPTG	Isopropyl β-D-1-thiogalactopyranoside
k_{cat}	Steady-state catalytic constant
k_{cat}/K_M	Catalytic efficient constant
K_M	Michaelis constant

KCl	Potassium chloride
KLD	Kinase ligase digest
KOAc	Potassium acetate
KOH	Potassium hydroxide
LB	Lysogeny broth
LC-MS	Liquid chromatography-mass spectrometry
LeuEnk	Leucine Enkephalin
m^1A	1-methyladenosine
MES	2-(*N*-morpholino)ethanesulfonic acid
$MgCl_2$	Magnesium chloride
$MgOAc_2$	Magnesium acetate
MR	Molecular replacement
MS	Mass spectrometry
MW	Molecular weight
MWCO	Molecular weight cut-off
NaCl	Sodium chloride
NaOAc	Sodium acetate
Ni-NTA	Nickel-nitrilotriacetic acid
NCBI	National Centre for Biotechnology Information
Ndas	*Nocardiopsis dassonvillei*
NEB	New England BioLabs
NH_4OAc	Ammonium acetate
$(NH_4)_2SO_4$	Ammonium sulfate
NMR	Nuclear magnetic resonance
NRPS	Non-ribosomal peptide synthetase
OD	Optical density
Para-	*Parabacteroides sp. 20_3*
Parcu-	*Parcubacteria bacterium RAAC4_OD1_1*
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein data bank
pK_a	Acid dissociation constant
PPi	Pyrophosphate
QTof	Quadrupole time-of-flight
RNA	Ribonucleic acid
S. aureus	*Staphylococcus aureus*
SAD	Single wavelength anomalous dispersion
SAM	*S*-adenosyl-L-methionine
SAH	*S*-adenosyl-L-homocysteine
SDM	Site directed mutagenesis
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC-SAXS	Size exclusion chromatography-single angle x-ray scattering
T_M	Melting temperature
TAE	Tris-acetate-EDTA

TB	Terrific broth
TBE	Tris-borate-EDTA
TEMED	Tetramethylethylenediamine
TFA	Trifluoroacetic acid
TLC	Thin layer chromatography
TrmK	*Bacillus subtilis* methyltransferase
tRNA	Transfer ribonucleic acid
UPLC	Ultra high performance liquid chromatography
UTP	Uridine 5'-triphosphate
UV	Ultraviolet
v	Initial rate
WT	Wild type
6xHis tag	Hexahistidine tag

Throughout this thesis, amino acids are either referred to using their triple or single letter codes:

Amino acid	**3 letter code**	**1 letter code**
L-Alanine	Ala	A
L-Arginine	Arg	R
L-Asparagine	Asn	N
L-Aspartic acid	Asp	D
L-Cysteine	Cys	C
L-Glutamic acid	Glu	E
L-Glutamine	Gln	Q
L-Glycine	Gly	G
L-Histidine	His	H
L-Isoleucine	Ile	I
L-Leucine	Leu	L
L-Lysine	Lys	K
L-Methionine	Met	M
L-Phenylalanine	Phe	F
L-Proline	Pro	P
L-Serine	Ser	S
L-Threonine	Thr	T
L-Tryptophan	Trp	W
L-Tyrosine	Tyr	Y
L-Valine	Val	V

Chapter 1 - Introduction

Organisms across all domains of life are capable of producing natural products of unsurmountable value, especially in a therapeutic context[1]. Even the most ancient of civilisations could understand that these natural products held a wealth of important properties, with the earliest records being found in Mesopotamia (2600 B.C.) which detailed oils from *Cupressus sempervirens* and *Commiphora* species to treat coughs and colds[2]. Natural products are classed as secondary metabolites i.e. they are not required for organism growth or regulation and are typically unique to the organism of which they are found[3]. Secondary metabolites are thought to aid the organism in survival, either by adapting to its surroundings or defending against predators[4]. However, humans have also taken advantage of these molecules with their highly useful properties, creating a field devoted entirely to this known as pharmacognosy. Therefore, the importance of natural products is abundantly clear with applications in nearly all forms of disease including bacterial, viral, and even cancerous.

Notably, the widespread use of antibiotics has been termed as the greatest medical breakthrough of the 20th century leading to the "Golden Age" (1940-1960), where microorganisms were investigated for further clinically relevant molecules[5]. From the 1970s, there was a notable decline in natural product isolation due to many factors including the emergence of antimicrobial resistance and discovery of high throughput combinatorial methods[3,5]. Additionally, difficulties were encountered as many organisms encode more natural product pathways than they produce under laboratory conditions. These pathways are known as 'cryptic' with around three quarters being inactivated *in vitro*[6]. Therefore, pharmaceutical corporations began to transition away from extract-based screening, claiming time constraints and lack of novel discoveries meant this method was no longer viable from a cost or synthetic point of view.

The advent of combinatorial chemistry in the 1990s seemed like the future of drug design by allowing a tailored systematic approach to yielding unique compounds from multiple libraries of known molecules. Consequently, during the mid-1990s to the mid-2000s, millions of compounds were tested using these newly developed high-throughput methods[7]. However, this endeavour did not produce any approved drugs to date and scientists soon realised that their combinatorial libraries lacked the means to create the truly complex structures featured in natural products[3,7]. This was evidenced further by the fact that from 1981-2006, 52% of new chemical entities were either a natural product, a mimic or a modification to an original natural product structure compared to the 30% that were synthetically made[2].

The development of genomics-based approaches, starting with two *Streptomyces* genomes in the early 2000s, propelled efforts into understanding the secondary metabolite gene clusters of

organisms[7–9]. Throughout evolution, similar patterns have been found in the biosynthetic pathways for many secondary metabolites, thereby allowing scientists to track and exploit these pathways[10]. Of course, this did not always mean the molecules in question would display desirable properties - hence the need for characterisation remains as research continues to struggle with the prediction of natural product properties.

Regardless of the discovery route to them, secondary metabolites from organisms have helped extend the human life span and are clearly fundamental to our survival as a species. Time and time again, researchers have referred back to the original scaffold of natural products for inspiration and diversification that is unrivalled elsewhere. Interestingly, scientists have barely scratched the surface of what is readily available given that over 95% of organisms found in nature have not been cultured under laboratory conditions as of yet[11]. Therefore, it is key that research continues to further uncover new molecules and highlight potential opportunities to utilise these natural products for their novelty and complexity.

1.1 Cyclodipeptides

1.1.1 Overview of cyclodipeptides

First discovered in 1880 by E. Fischer, cyclodipeptides (CDPs) have since earned well-deserved recognition for their impressive properties[12]. They are the smallest class from the cyclic peptide family - sometimes referred to as 2,5-diketopiperazines – and are formed from the cyclisation of two amino acids with 'head-to-tail' peptide bonds (Figure 1.1)[13]. CDPs feature an integral 6-membered ring which can be derivatised at up to 6 positions with controlled stereoselectivity at 4 positions[14]. As such, this rigid ring structure contains two hydrogen bond donors and two hydrogen bond acceptors which are important for reducing entropic penalties and binding to targets, for example proteins including receptors[15]. Cyclodipeptides are known for their remarkable biological properties, such as being resistant to proteolysis, permeability to the blood-brain barrier and ability to mimic functional pharmacophores and optimized peptide conformations[16,17]. Moreover, the cyclic backbone imposes a rigid peptide scaffold which increases the stability of this group of molecules to far greater than their linear counterparts. Therefore, given all the advantages, CDPs can be classed as 'privileged' scaffolds, a term first used by Evans *et al.* in 1988 to highlight the beneficial drug-like properties of molecules[18,19]. Substitutions to the ring can influence the preferred conformation by introducing steric compromises to minimize intramolecular interactions between side groups. Furthermore, the overall conformation of these heterocycles has been investigated via crystallography where they were shown to exist in either a flat conformation or a slightly puckered boat form depending on the external environment. For example, cyclo(Gly-Gly) was first solved in a planar formation in 1938 which was confirmed by

density functional theory (DFT) calculations, however data from ab initio calculations and microwave spectra demonstrated a boat conformer when in the gaseous state[20,21].

1.1.2 Origins of cyclodipeptides

Originally, CDPs were disregarded as by-products of peptide degradation, found in many processed food or beverages such as beer, cheese and coffee[22]. During the boom of natural product research in the 20th century, scientists realized that cyclodipeptides were featured in a significant number of molecules, often with complex derivatisations (Figure 1.1). Cyclodipeptide scaffolds have been found mainly in growth media or extracts from microorganisms such as bacteria like *Pseudomonas aeruginosa*[23], or *Bacillus subtilis*[24], and fungi including *Pencillium*[25], but there are also reports from marine sponges[26], and animal venoms[27]. More often than not, the CDP backbone is present with a series of modifications including hydroxylations, N- or C-methylations or α,β-dehydrogenations[28]. Overall, CDPs are now known to be naturally produced by a large variety of organisms across the prokaryote and eukaryote kingdoms[27].

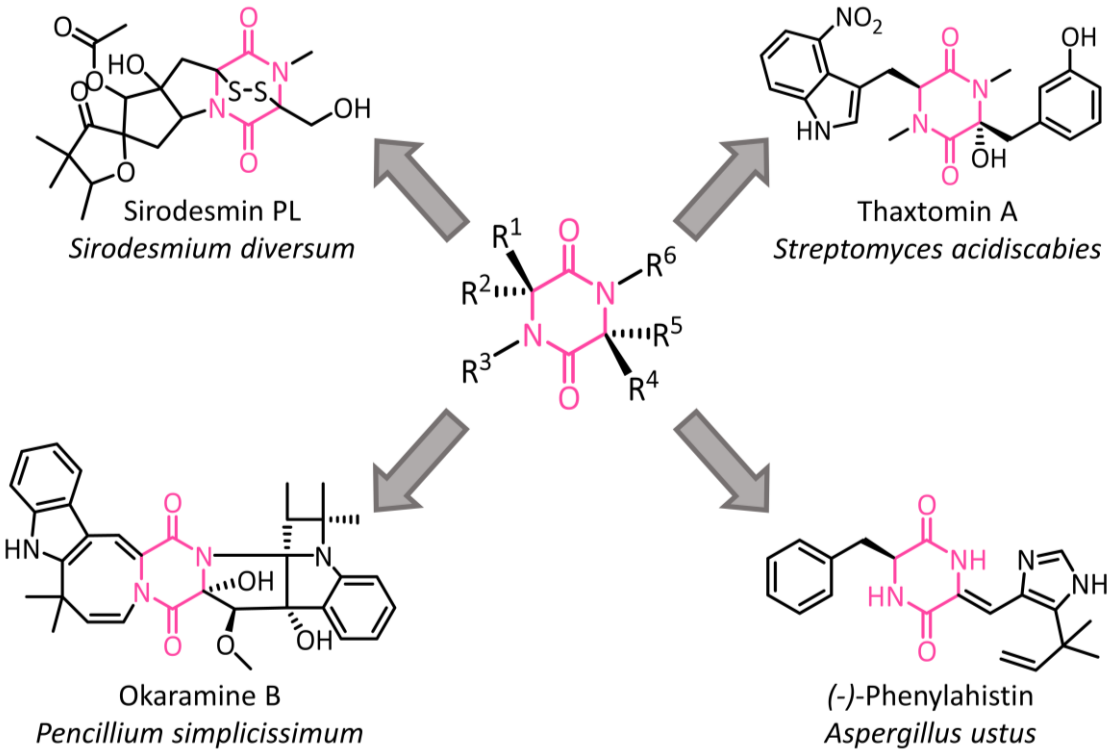


Figure 1.1 Cyclic dipeptides found in nature.
The core 2,5-diketopiperazine ring is depicted by the pink bonds. The name of each molecule is given with the species from which it was extracted underneath in italics.

Despite the exhaustive research into CDPs, their original function in nature remains elusive. It is thought that their role in quorum sensing (QS) allowed them to contribute to conserved signaling pathways in bacteria mediating symbiotic and virulent behaviours[29]. Additionally, interkingdom communication has been observed with these cyclic molecules, for example between bacteria and

plants [30], or even more interestingly between the staphylococcal quorum sensing system and a human vaginal isolate, *Lactobacillus reuteri* RC-14[31]. [31]On a more basic level, it is speculated that CDPs are simply messenger molecules used for cell communications of the intra- and intercellular nature[32]. Their presence in the human body was an interesting discovery and their function has been questioned to mirror the role of bacterial CDPs. Only two CDPs have been reported to be endogenous to the human body – cyclo(His-Pro) and cyclo(Gly-Pro)[29]. The generation of cyclo(His-Pro) is suggested by some to be a spontaneous by-product from the degradation of the thyrotropin-releasing hormone – pGlu-His-Pro – by pyroglutamate aminopeptidase. The resultant dipeptide, linear His-Pro, is fixed in such a conformation by the rigid proline that cyclisation almost always occurs spontaneously to produce cyclo(His-Pro)[33]. Coincidentally, cyclo(Gly-Pro) is also formed from another hormone, insulin-like growth factor-1 (IGF-1) which is a tripeptide composed of linear Gly-Pro-Glu[34].

1.1.3 Applications of cyclodipeptides

Initially, CDPs isolated from microorganisms were tested for activity against different cancer cell lines. This led to a wide range of CDPs discovered to be putative anti-cancer therapeutics including glycine containing peptides such as cyclo(Gly-L-Leu) and cyclo(Gly-L-Val)[35]. However, these were not quite as effective as diketopiperazines composed of aromatic amino acids including cyclo(L-Tyr-L-Cys) and cyclo(L-Phe-L-Pro) which were able to inhibit cancer cell growth in Ht-29, MCF-7 and Hela cells[35]. Unfortunately, natural cyclodipeptides remained second place to known cancer drugs due to the high CDP concentration required to influence the growth of tumour cells. Therefore, design modifications to the scaffold were introduced synthetically. One of the most successful cases of this was plinabulin, a derivative of natural product phenylahistin. First isolated from a culture of Aspergillus ustus NSC-F038, phenylahistin was shown to inhibit the cell cycle and postulated to be an anti-cancer therapeutic [36]. However, a chiral carbon led to an enantiomeric mixture of products where only one configuration (S or commonly referred to as (-)) was active[37]. To remove this chirality issue and dictate biological activity, synthetic derivations of phenylahistin were produced thus leading to the development of plinabulin[38]. Rounds of optimization using a Structure-Activity Relationship study (SAR) uncovered the defining features of plinabulin which allowed it to be such a potent inhibitor: two unsaturated bonds to the side groups of the diketopiperazine ring; the L-form of the benzyl moiety, and the preserved hydrogen bonding interaction between the ring backbone amide hydrogen and the π-nitrogen of the imidazole ring (Figure 1.2)[39]. This drug is currently undergoing extensive stage III clinical trials in combination with docetaxel, a tubulin stabilizing taxane, with the most recent study focusing on non-small cell lung cancer (ClinicalTrials.gov Identifier: NCT02504489).

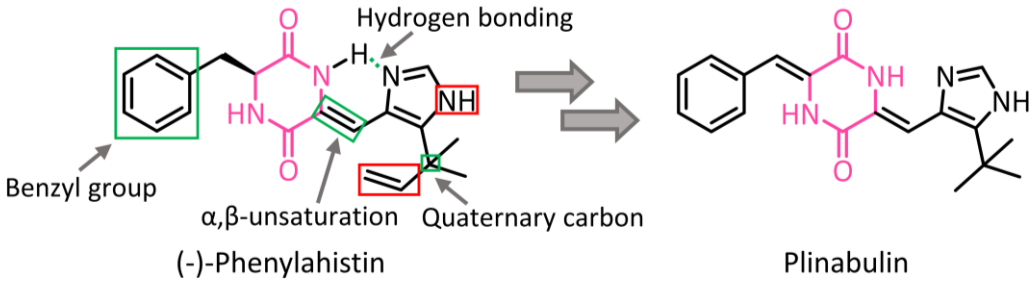


Figure 1.2 Structures of (-)-Phenylahistin and its derivative, Plinabulin.
Structural components essential for potency are highlighted in green boxes whilst features deemed unnecessary are in red[39].

Additionally, cyclodipeptides are known for their antimicrobial properties, an obvious characteristic of molecules produced for an organism's defence and survival. Interestingly, some amino acid components such as proline, arginine and tryptophan are more likely to be featured in antimicrobial compounds. A range of proline containing dipeptides were shown to be effective against both medical and agricultural bacteria as well as being antifungal against different organisms[35]. Again, the stereochemistry of these CDPs has proven to be a deciding factor in the efficiency of these compounds which appears to be a trend across applications[40].

The previously mentioned cyclo(His-Pro) has displayed positive anti-inflammatory and anti-oxidative effects on the central nervous system in the body[41]. The ability to cross the blood-brain barrier has led to the proposal that this molecule could function as a potential neuroprotective agent against degenerative conditions such as Alzheimer's disease and amyotrophic lateral sclerosis (ALS)[42–44]. Moreover, literature suggests that there is a correlation between type 2 diabetes and an increased risk of developing Alzheimer's disease. Both are reliant on efficient transportation and absorption of zinc which is usually deficient in patients with these conditions. Therefore, trials have used cyclo(His-Pro) to sequester zinc and act as a transporter through the gastrointestinal tract, adding to the countless possibilities of applications for these molecules[45].

As highlighted above, cyclodipeptides possess a variety of untapped potential owing to their impressive properties. In the literature, CDPs are reported as anti-cancer, antiviral, antimicrobial and anti-inflammatory agents proving themselves to be valuable therapeutics for various conditions. Naturally occurring CDPs are versatile and clear starting points for further modifications which add to the complexity of their structure and function. Therefore, it is imperative to conquer the synthetic route to cyclodipeptides with simple routes to ring derivatisations which would surely propel modern medicine efforts.

1.1.4 Synthesis of Cyclodipeptides

Conventional solution based chemistry has always been the first approach when aiming to synthesise a specific product and this has been successful for CDPs. Linear dipeptides with unprotected termini were shown to undergo cyclisation in acidic or basic conditions and indeed this method was successful in yielding plinabulin[14]. Harsh conditions, however, have been shown to produce a mix of enantiomers, therefore these methods were not always suitable. Microwave-assisted heating of dipeptide esters for 10 minutes was also demonstrated to give good product yields whilst removing the use of environmentally harmful components and possible risks of epimerisation[46]. Peptide synthesis using solid phase supports has been employed since the 1960's which allowed for the synthesis of many different CDPs[47]. Unfortunately, resin based chemistry includes multiple step processes increasing the overall cost whilst decreasing the final product yield. Complex derivatisations such as those seen in natural products can be mimicked with chemical synthesis, often with a multitude steps and uncommon reagents. Consequently, research has been directed towards simple processes taking advantage of naturally occurring biocatalytic methods.

There are two biosynthetic pathways to cyclic dipeptides which utilise different enzymes and different substrates. Nonribosomal peptide synthetases (NRPSs) have been found mostly in bacteria but there are instances reported in eukarya and archaea[48]. NRPSs are characterised by their large complexes where each module typically contains an adenylation domain, a peptidyl carrier protein or thiolation domain and lastly a condensation domain which allow these proteins to successfully catalyse a multitude of peptide synthesis processes. The substrates of these enzymes are not limited to the 20 proteinogenic amino acids and around 500 different monomers, including nonproteinogenic amino acids and fatty acids, have been accepted for peptide synthesis[49]. NRPSs have been extensively studied with reviews available in the literature[48], however they are not the focus of this research and discussion will be limited here.

1.2 Cyclodipeptide Synthases (CDPSs)

1.2.1 Overview of CDPSs

In 2002, Lautru et al. were investigating the biosynthetic pathway of the natural product albonoursin and how microorganisms were synthesizing cyclodipeptides for future derivatisation to complex molecules. Their research led to the characterization of *Snou*-CDPS (formerly called AlbC) as the sole enzyme necessary for the formation of the CDP, cyclo(L-Leu-L-Phe)[12]. This discovery enabled the identification of a new enzyme class termed cyclodipeptide synthases (CDPSs) - a family of small proteins capable of hijacking aminoacylated tRNAs from the ribosome to produce CDPs. This type of enzyme was unprecedented at the time and since then, publications have focused on uncovering the

unique properties of CDPSs. However, it took seven years to confirm the substrates were aminoacylated tRNAs which aided the understanding as to why these enzymes could function whilst being so small in comparison to other enzymes which make similar CDP products i.e NRPSs (Figure 1.3)[50].

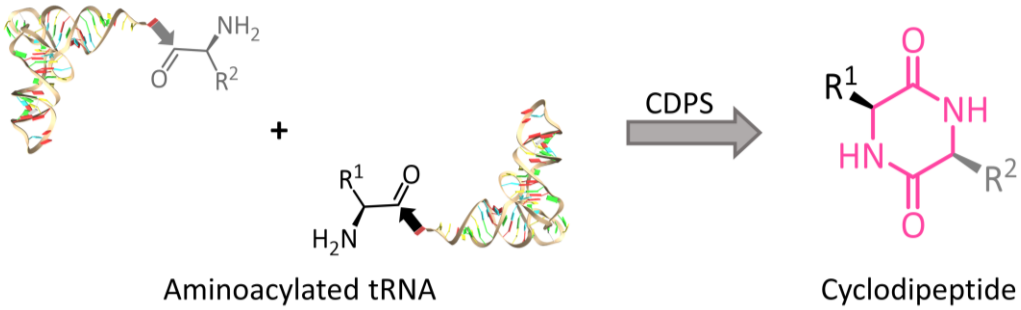


Figure 1.3 Simplified route to cyclodipeptides using cyclodipeptide synthases.
Aminoacylated tRNA is represented by an amino acid (R^1 in black or R^2 in grey) coupled to a molecule of tRNA (PDB:1EHZ[51]) via the terminal adenosine represented by a small arrow. The resultant CDP scaffold has the common 2,5-dikeotpiperazine architecture highlighted in pink.

All CDPSs are approximately 20-30 kDa in size whilst NRPSs are known to be greater than 100 kDa due to their large modular structure. For nonribosomal peptide synthetases to produce cyclodipeptides, they first need to activate their amino acid substrates to aminoacyl thioesters whereas cyclodipeptide synthases use pre-activated substrates available in the cell allowing them to be conservative in size[52]. Alignment of CDPSs to known domains highlighted a strong similarity to class-I aminoacyl-tRNA synthetases (aaRSs), specifically tyrosyl-RS and tryptophanyl-RS[53]. The first published structure of a CDPS in 2010 confirmed the presence of a Rossman fold, also common to class-I aaRSs, which supported the possibility that CDPSs evolved from these precursors[54]. There are, however, distinct differences between the two owing to the different ways in which they interact with their substrates. For example, CDPSs do not require an ATP binding motif (amino acid HIGH sequence at the N-terminus) which is conserved across the class-I aaRSs as well as the tRNA-binding domain located at the C-terminus of aaRSs[55]. The lack of this HIGH sequence allows the CDPS structure to position two catalytically important residues at the active site for substrate binding, therefore highlighting the different substrate interaction networks employed by these enzymes[54].

1.2.2 Mechanism of Cyclodipeptide Synthases

Efforts to elucidate the mechanism of CDPSs were extensive, spanning over 4 years and resulted in a good understanding of how *Snou*-CDPS, and more generally CDPSs, were able to generate cyclic dipeptide products. Originally, Vetting et al. investigated the mechanism of Rv2275, a cWW-producing CDPS, by performing mutagenesis to confirm the role of a serine residue. A lack of protein activity was reported upon substitution of this particular residue with alanine, thus highlighting its essentiality

within the mechanistic process[54]. However, this research group postulated a preliminary catalytic mechanism for the CDPS family which was ultimately proven incorrect. Their hypothesis involved covalent attachment of the first tyrosine substrate to the serine followed by a rearrangement to place the serine-tyrosyl ester into the 'second surface depression' as they termed it. This would then allow the second tyrosine substrate to occupy the now empty pocket and interact with the first substrate, forming a di-tyrosine peptide coupled to the tRNA moiety. Vetting argued the enzyme must support the deprotonation of the 3' hydroxyl group of the tRNA and then cyclisation of the dipeptide would occur spontaneously[54]. This mechanism we know now to be incorrect thanks to additional work to uncover the second reaction intermediate performed by Moutiez *et al*[56,57]. The active site serine of *Snou*-CDPS was mutated to a cysteine and this new thiol group participated in a coupling reaction with an intermediate mimic molecule which was unable to be turned over by the enzyme[56]. This trapped state was investigated using crystallographic methods to probe the key residues responsible for turning over the substrate to product. Computational analyses then confirmed the role of the four catalytic residues - S37, Y178, E182 and Y202 (numbering respective to *Snou*-CDPS) - thus strengthening the evidence for this alternative mechanistic pathway[58]. S37 is responsible for the first binding event and removing this residue resulted in an inactive enzyme[54]. Upon the formation of the first aminoacyl-enzyme intermediate, an interaction network with Y178 and E182 is formed to anchor the intermediate in place for productive coupling with the second substrate. Additionally, the second reaction intermediate, the dipeptidyl-enzyme, must be in the cis conformation for cyclisation to occur and these residues ensure this with hydrogen bonding interactions to hold the complex in place[56]. Lastly, the cyclisation step of the reaction requires Y202 to act as a proton relay where the proton from the N-terminal amino group of the dipeptidyl-enzyme complex transfers to the hydroxyl of the tyrosine and subsequently the tyrosyl proton moves to the newly unbound serine oxygen to regenerate the catalytic residue for subsequent turnovers[58].

To give an example of this reaction, it is known that *Snou*-CDPS uses phenylalanyl-tRNA and leucyl-tRNA as the first and second substrate respectively. The mechanism therefore begins by covalently attaching the first substrate, phenylalanine Phe-tRNA^{Phe}, to the catalytic serine (S37) of the CDPS. Y178 and E182 hold the bound substrate in a favorable conformation for second substrate binding, in this case Leu-tRNA^{Leu}. A new peptide bond is formed between the two substrates causing the free tRNA^{Leu} to dissociate from the enzyme. This leaves a dipeptidyl-enzyme complex which undergoes subsequent cyclisation with help from Y202 to generate the final cyclic product, cyclo(L-Phe-L-Leu) (Figure 1.4)[59].

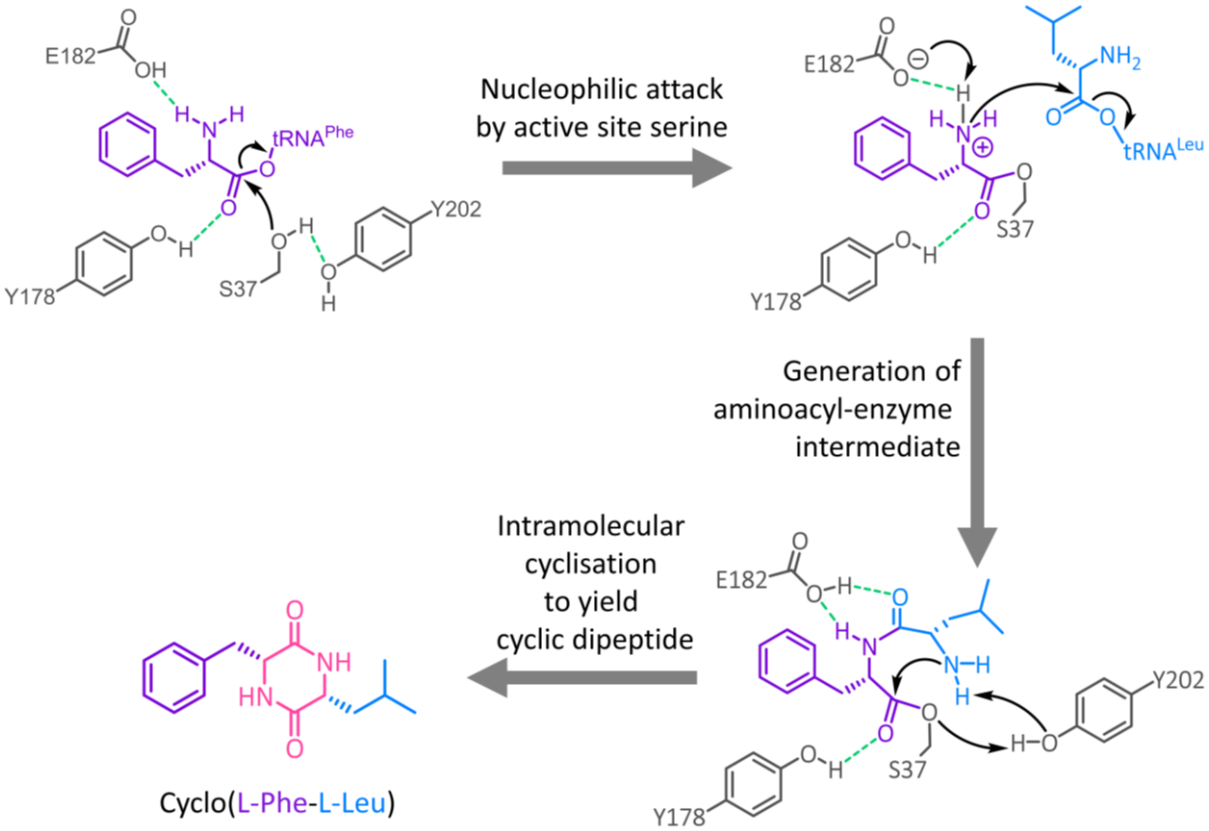


Figure 1.4 Catalytic mechanism of cyclodipeptide synthases.
Using *Snou*-CDPS as an example, the catalytic mechanism is drawn with phenylalanine (purple) and leucine (blue)[60]. The active site residues of *Snou*-CDPS are in grey and H-bonding interactions are depicted in green.

1.2.3 Classification of Cyclodipeptide Synthases

The establishment of the CDPS family includes the discovery of 120 CDPSs with empirically determined function and more than 1000 predicted members with unknown function[61]. Cyclodipeptide synthase genes have been found mostly in bacterial phyla and all but one active CDPS originate from bacteria – this exception was from a sea anemone[62,63]. From this family, a diverse set of cyclodipeptide products could be generated by incorporating a range of proteinogenic amino acids. Early investigations proposed CDPSs to be divided in two distinct groups based on the hydrogen bonding network around the essential Y202 residue. These two subsets have been defined as NYH or XYP where NYH corresponds to residues N40, Y202, and H203 and XYP corresponds to X40, Y202 and P203 (numbering here refers to *Snou*-CDPS)[64]. It is important to note that mutations to convert an XYP-CDPS into an NYH-CDPS resulted in a 50% decrease in overall product yield highlighting that each motif is not interchangeable between the two groups[65]. Regardless of this group classification, all known CDPSs display the 4 key catalytic residues - S37, Y178, E182 and Y202 - and are therefore assumed to participate in the same reaction mechanism depicted in Figure 1.4.

It was hypothesised that divergence was linked to the ways in which the CDPS interacts with the tRNA body and indeed, Bourgeois et al. confirmed this by crystallising an XYP-CDPS with aminoacylated tRNA. Using an inactive mutant of *Cglo*-CDPS, charged phenylalanyl-tRNA (i.e. Phe-$tRNA^{Phe}$) was added to trap the enzyme in an unproductive state, allowing them to investigate potential binding modes[66]. The 5.0 Å resolution structure highlighted that *Cglo*-CDPS-S32A used two beta strands (β2 and β7 in blue) from the first half of the Rossman fold to interact with the major groove of the acceptor stem of phenylalanyl-tRNA (Figure 1.5). To position the phenylalanine group into the catalytic site of the enzyme, the CCA arm of the $tRNA^{Phe}$ adopted a bent configuration (highlighted by the rotated view of the complex in Figure 1.5) to optimally fit in a channel formed by the presence of the two catalytic loops (depicted in blue in Figure 1.5). However, given the low resolution of the *Cglo*-CDPS-S32A tRNA complex, the overall protein structure was refined only as a polyalanine chain to avoid overfitting of the data. Thus, no further information can be inferred from the catalytic residues except that the four are present when the sequence is aligned to other known CDPSs. Overall, this finding did showcase the difference between XYP and NYH-CDPSs, given that NYH-CDPSs have shorter beta strands which are unable to provide favourable interactions with the incoming tRNA moiety. However, it is difficult to definitively state that these two classes interact with tRNA differently without crystallographic proof using an NYH-CDPS complexed to a tRNA substrate. Furthermore, this work emphasised that whilst CDPSs display structural similarity to class-I aaRSs, their modes of binding to substrates are completely different and no further comparisons can be drawn to them regarding mechanism or substrate recognition. Interestingly, when comparing the CDPS to EF-Tu - a molecular competitor for aminoacylated tRNA in the cell - it was shown that EF-Tu recognises its substrates using the opposite end of the tRNA. Bourgeois et al. have postulated that this could, in fact, help the CDPS directly hijack an aminoacylated tRNA already bound to EF-Tu[66].

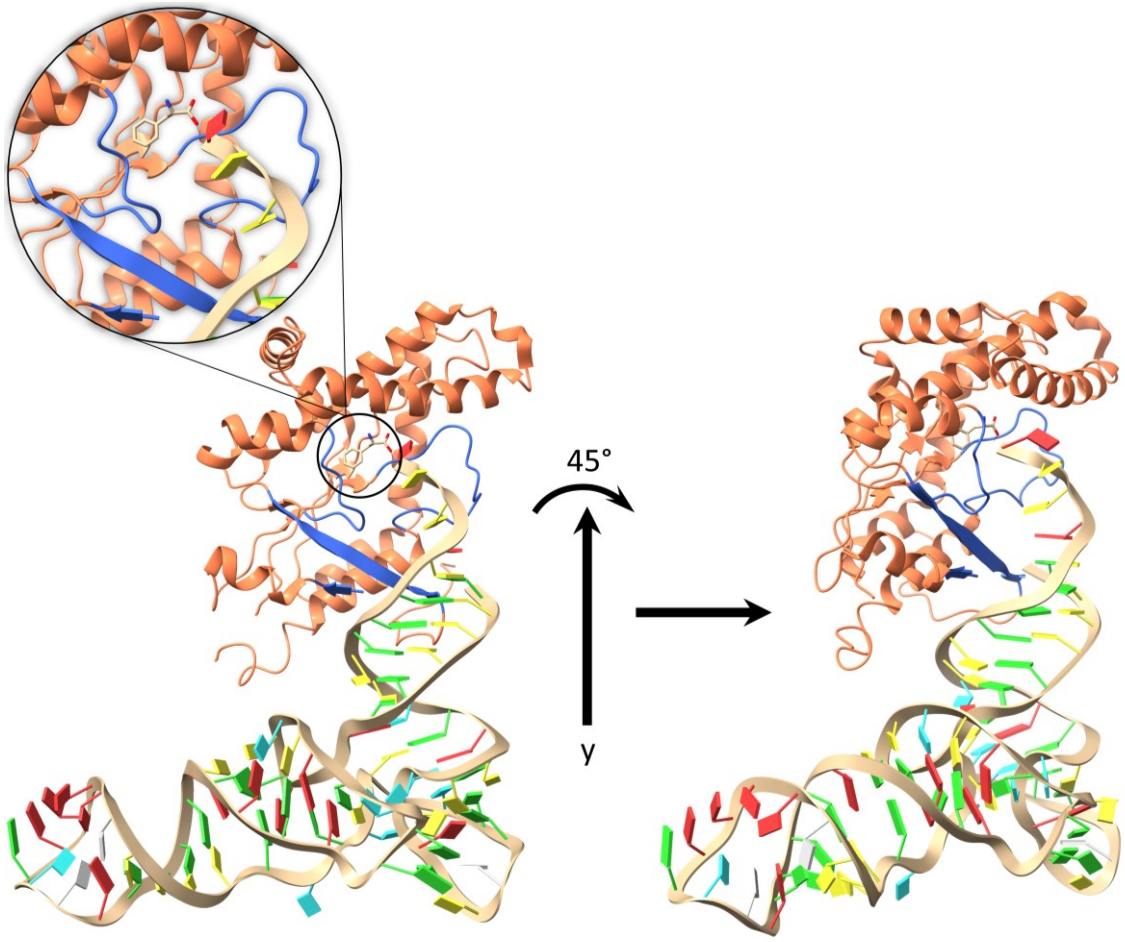


Figure 1.5 Crystallographic substrate of *Cglo*-CDPS-S32A mutant complexed to Phe-tRNA^{Phe}. Phe-tRNA^{Phe} is drawn with a beige backbone and ribonucleotide bases coloured as follows: adenine in red; uracil in blue; cytosine in yellow and guanine in green. The structure of *Cglo*-CDPS-S32A is depicted in orange and the key structural motifs postulated to interact with Phe-tRNA^{Phe} are highlighted in blue (PDB:6Y4B)[66]. The 45 ° rotation highlights the unique bend of the terminal CCA ribonucleotides which is thought to be unique to XYP-CDPSs.

1.2.4 Structural Investigations of Cyclodipeptide Synthases

To date, there are 8 deposited crystal structures from the CDPS family - five NYH-CDPSs: Rv2275[54]; *Snou*-CDPS[56]; YvmC[57]; *Shae*-CDPS[65]; and BtCDPS[67], and 3 XYP-CDPSs: *Rgry*-CDPS; *Nbra*-CDPS, and *Fdum*-CDPS[65]. Direct structural comparison between NYH-CDPSs and XYP-CDPSs revealed the structures diverge the most in the first half of the Rossmann fold which is where the beta strands that are hypothesised to interact with the tRNA (β2 and β7) reside[65]. Additionally, these 3D structures highlighted the two solvent accessible pockets of the CDPS structure which facilitate substrate binding. For the first substrate, pocket 1 (P1) is situated deep within the protein and is rather narrow – the residues lining this pocket contribute to the type of amino acids the enzyme can accept. Pocket 2 (P2) appears to be closer to the surface of the protein and is a lot larger and shallower than P1[59]. It is thought that this large pocket gives the CDPS advanced substrate promiscuity, allowing the enzyme to generate more than one product. Research has focused on predicting the activity of CDPSs using

specific motifs found in P1 and P2, however it was only reliable and accurate for predicting the substrate of P1 and could not accurately give all the minor products of a CDPS[62]. Indeed, it would be advantageous to quickly predict the product of a CDPS just by sequence motifs, however it is clear that there has only been limited success in this endeavour at present[62].

1.2.5 Substrate specificity of Cyclodipeptide Synthases

From the information known about cyclodipeptide synthases, researchers assumed there was one substantial caveat in using these enzymes to produce CDPs – their substrate selectivity[68]. CDPSs use aminoacylated tRNA which is generated by the cell using endogenous amino acids and aminoacyl-tRNA synthetases (aaRSs). Aminoacyl-tRNA synthetases allow amino acids to participate in protein synthesis by catalysing the esterification between the carboxylic acid group of the amino acid and one of the two free hydroxyl groups on the ribose ring of the tRNA terminal adenosine[69]. Overall, this includes a two-step process where the amino acid is first activated with adenosine triphosphate (ATP) to generate an aminoacyl-adenylate intermediate and pyrophosphate. The cognate tRNA contains an anticodon loop with a complementary sequence to the mRNA codon for the amino acid as well as an acceptor stem which non-covalently interacts with the amino acid to form the aminoacylated tRNA molecule which is then ready to the enter the ribosome for protein synthesis[70]. Recently, it was discovered that the mechanism in which these enzymes identify ATP differs based on which class of aaRS they belong to. Class I aaRSs use backbone hydrogen bonding to interact with ATP causing a structural rearrangement upon binding whereas class II aaRSs use a pair of arginine residues to create salt bridges with the ATP molecule[71]. This finding aids the understanding as to why Class I aaRS load the amino acid onto the 2'-hydroxyl group and Class II aaRS load onto the 3'-hydroxyl[72]. Moreover, the most prevalent type of non-covalent interaction each aaRS employs with the amino acid substrate differs with Class I aaRS using hydrophobic interactions the most whilst Class II aaRS use hydrogen bonding[73]. Aminoacyl-tRNA synthetases are incredibly important proteins, directly linked to protein translation, therefore their reactions are tightly regulated to prevent mis-aminoacylation which could be fatal to cells and organisms. To do this, each aaRS employs a specific range of strategies to both select the correct amino acid and rectify any mis-aminoacylation that might occur.

Soon it was realised that the natural promiscuity of aminoacyl-tRNA synthetases could be exploited to accept non-canonical amino acids which were similar to their natural substrate – for example selenomethionine instead of methionine - and use them in peptide synthesis[74]. This is due to the lack of cellular pathways developed for editing amino acids not traditionally found inside the cell. Research has since been devoted to the development of methods to yield non-canonical aminoacylated tRNA molecules for various purposes. Chemical aminoacylation was able to incorporate unnatural amino acids, however it was deemed laborious and difficult to employ protected amino acids for a multiple

step synthesis[75–77]. Alternatively, the synthetases themselves were directly mutated to force the introduction of unnatural amino acids as substrates[78–80]. Unfortunately, this hypothesis was not as successful as hoped and was rather limited by both the number of aaRSs which would tolerate the mutations and the nature of the non-proteinogenic amino acids that were accepted as substrates. A pivotal finding then came from a highly flexible artificial tRNA aminoacylation system using a ribozyme, also known as a flexizyme[81,82]. Suga et al. demonstrated that these flexizymes only needed to recognise two components for successful aminoacylation: the 3'-end of tRNA and a benzylic leaving group[83]. This completely cell-free protocol overcame the previous disadvantages of the techniques at the time by expanding the scope of both the tRNAs and the amino acids available for aminoacylation. Alternatively, there were accounts of non-canonical amino acids being introduced into proteins during expression in vivo. This technique was most useful for metabolic labelling by introducing a reactive group which could then be detected via HPLC or MS analysis[84]. This method, developed by David Tirrell and colleagues, was termed biorthogonal noncanonical amino acid tagging (BONCAT)[85]. It should be noted that the aaRSs were not altered in the cell - instead they were subjected to high concentrations of unnatural amino acids and in the absence of the natural substrate, the synthetase was encouraged to load the tRNA with the analogue instead.

Moreover, an extensive review of all the known non-canonical amino acids accepted by the 20 aminoacyl-tRNA synthetases from E. coli demonstrated that over 250 analogues could be used by aaRS enzymes[86]. It should be highlighted that AsnRS and GlyRS have no confirmed non-canonical amino acid substrate to date and were therefore not included in the review. Having confirmed the successful generation of non-canonical aminoacylated tRNA, research then shifted to investigating if CDPSs were able to use these altered substrates to generate diverse cyclic products. In 2018, Canu et al. detailed the incorporation of non-canonical amino acids into the ring using an in vivo system in E. coli[87]. This resulted in around 200 new cyclic dipeptides containing diverse modifications including halogenations, hydroxylations, azido-functionalisations and methylations. Cyclodipeptide synthases themselves likely pose as an additional sieve to the non-canonical amino acid incorporation as only 26 of the 60 known unnatural amino acids were accepted. Alternatively, it is possible that the tested non-canonical amino acids struggled to enter the cells of this in vivo system thereby directly affecting the product yield of certain analogues. Overall, more information is required about the substrate recognition of CDPSs in order to fully utilise them for the generation of diverse products.

1.3 Tailoring Enzymes

The discovery of one gene from a multi-step biosynthetic process typically leads to the subsequent discovery of the others due to clustering of related genes within the genome. CDPSs of bacterial origin

have been shown to be located close to their modifying partners, hereafter known as tailoring enzymes. Currently there are over 700 putative CDPS-encoding genes in the public domain (InterPro: IPR038622)[88] which have predicted associated tailoring enzymes, however there are only 9 completely characterised biosynthetic pathways to date, six of which are featured in Figure 1.6[89]. These are amongst the smallest natural product gene clusters due to the compact nature of CDPSs and low number of tailoring enzymes required for diversification, owing to their promiscuity[90]. A range of tailoring enzymes have been revealed from these pathways, which are by no means exhaustive, therefore this group are important enzymes with mostly unknown applications.

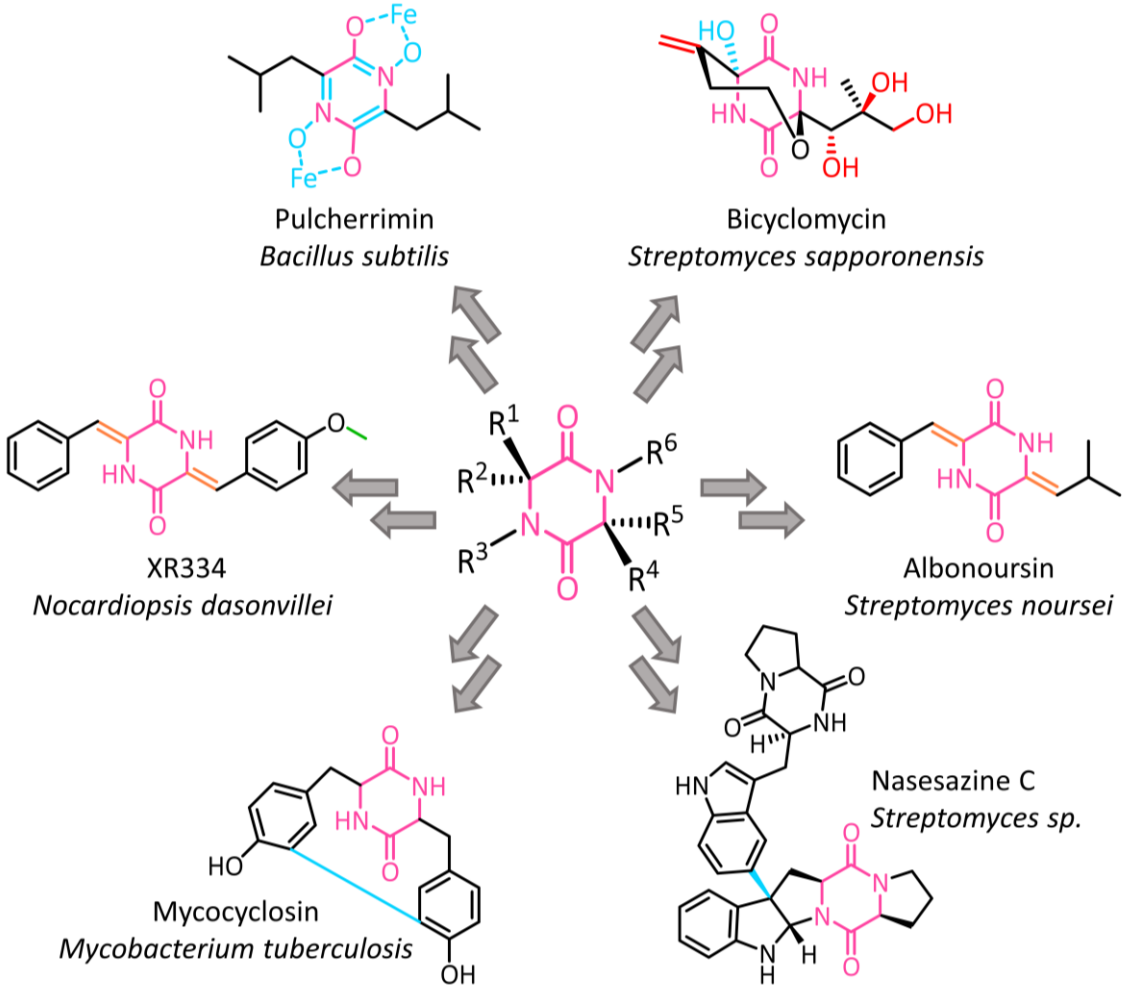


Figure 1.6 Summary of various cyclodipeptides available from tailoring enzymes.
The modifications by tailoring enzymes are colour coded: blue for cytochrome P450 enzymes, orange for cyclodipeptide oxidases, green for methyltransferases and red for α-KG/Fe^{2+}-dependent dioxygenases.

To give a brief overview, the first confirmed biosynthetic pathway was the biosynthesis of albonoursin from which *Snou*-CDPS was found[12,91]. This highlighted a class of tailoring enzymes capable of oxidising C-C single bonds to double bonds which have since been found in other pathways including the nocazines from *Nocardiopsis dassonvillei* and purincyclamide from *Streptomyces*

chrestomyceticus[92,93]. These cyclodipeptide oxidases (CDOs) are highly useful enzymes, composed of two subunits referred to as CDOA and CDOB with the molecular weights of 22 and 11 kDa respectively[94]. They are found together in the genome, typically downstream from the cyclodipeptide synthase encoding gene. Both genes were discovered to have overlapping regions of between 20-30 nucleotides and this is vital for the production of active enzyme as previous research to express each gene separately resulted in an inactive form[92]. Furthermore, this complex requires a flavin cofactor which appeared to be covalently bound and, together with both subunits, constitutes an active unit. With limited experimental evidence, the oligomeric state of this protein complex remains unknown with only one group in the early 2000's claiming that the complex could be greater than 2,000,000 Da after size exclusion chromatography showed the complex eluted in the void of a Superose 6 HR column[91]. Substate dehydrogenation was unsuccessful under anaerobic conditions revealing that oxygen is necessary to act as the electron acceptor for the flavin redox reaction and the by-product from the oxidation is therefore hydrogen peroxide (H_2O_2). Preliminary tests highlighted that the central 2,5-diketopiperazine ring appears essential to the CDO as linear peptides were not oxidised, reinforcing the essential role of cyclodipeptide synthases in natural biosynthetic pathways[91]. Le Chevalier *et al.* proved that CDPSs and CDOs from different organisms can be used in a tailored cascade to produce molecules of interest. Eighteen CDPSs were tested in combination with eight CDOs in *E. coli* which discovered that conversion was more likely in the presence of hydrophobic or aromatic side chains and the rate of reaction was dependent on the nature of the cyclic dipeptide[94]. This research emphasised the relaxed substrate scope of this family, with several CDOs capable of oxidising the same cyclodipeptide.

Beyond cyclodipeptide oxidases, there are other tailoring enzymes responsible for the modification of cyclic dipeptides and generation of diverse natural products. For example, bicyclomycin involves the most complex characterised biosynthetic pathway to date with the involvement of 7 different enzymes[95,96]. The original CDP ring is formed from a CDPS which then undergoes three subsequent hydroxylations, followed by α,β-dehydrogenation and epoxidation of one leucine side group. The exact mechanism to the bicyclic structure is unknown but it has been postulated that it could happen from spontaneous nucleophilic attack by the opposing leucine side chain, or it could be enzyme-catalysed by the same protein responsible for the dehydrogenation and epoxidation steps. The last two manipulations involve a P450 enzyme to incorporate an additional hydroxyl group and then a final dehydrogenation to produce an exo-methylene group[90]. The lesser known *S*-adenosylmethionine (SAM) dependent methyltransferases are also present in other biosynthetic pathways for diverse molecules from the nocazine family[97]. Overall, it is clear that tailoring enzymes enable interesting unique modifications that would be otherwise difficult to achieve via traditional synthetic methods.

Research remains incredibly limited on this enzyme class, especially given the potential for a range of tailoring enzymes present within the 700 known individual biosynthetic pathways involving cyclodipeptide synthases. Yet, there may be overlap between the substrate specificity of these enzymes, as shown by the cyclodipeptide oxidases, therefore leading to less novel products, but this has yet to be demonstrated.

1.4 Project Aims

The focus of this work targets two histidine accepting cyclodipeptide synthases and their biosynthetic pathways. Chapter 3 introduces *Para*CDPS from *Parabacteroides sp. 20_3* which synthesises cyclo(L-His-L-Phe) (cHF) and *Parcu*CDPS from *Parcubacteria bacterium RAAC4_OD1_1* which produces both cyclo(L-His-L-Pro) (cHP) and cyclo(L-His-L-Glu) (cHE) (Figure 1.7). Current literature has glossed over these enzymes, having just characterised them as active with possible classification of being XYP-CDPSs based on sequence alignment alone. This work therefore aims to address the lack of information around the mechanism which these enzymes employ to selectively use histidine as a substrate. The development of non-canonical aminoacylated tRNA in Chapters 4-5 allows us to probe the specificity of both enzymes, seeking to understand residues essential for substrate recognition by these unique CDPSs.

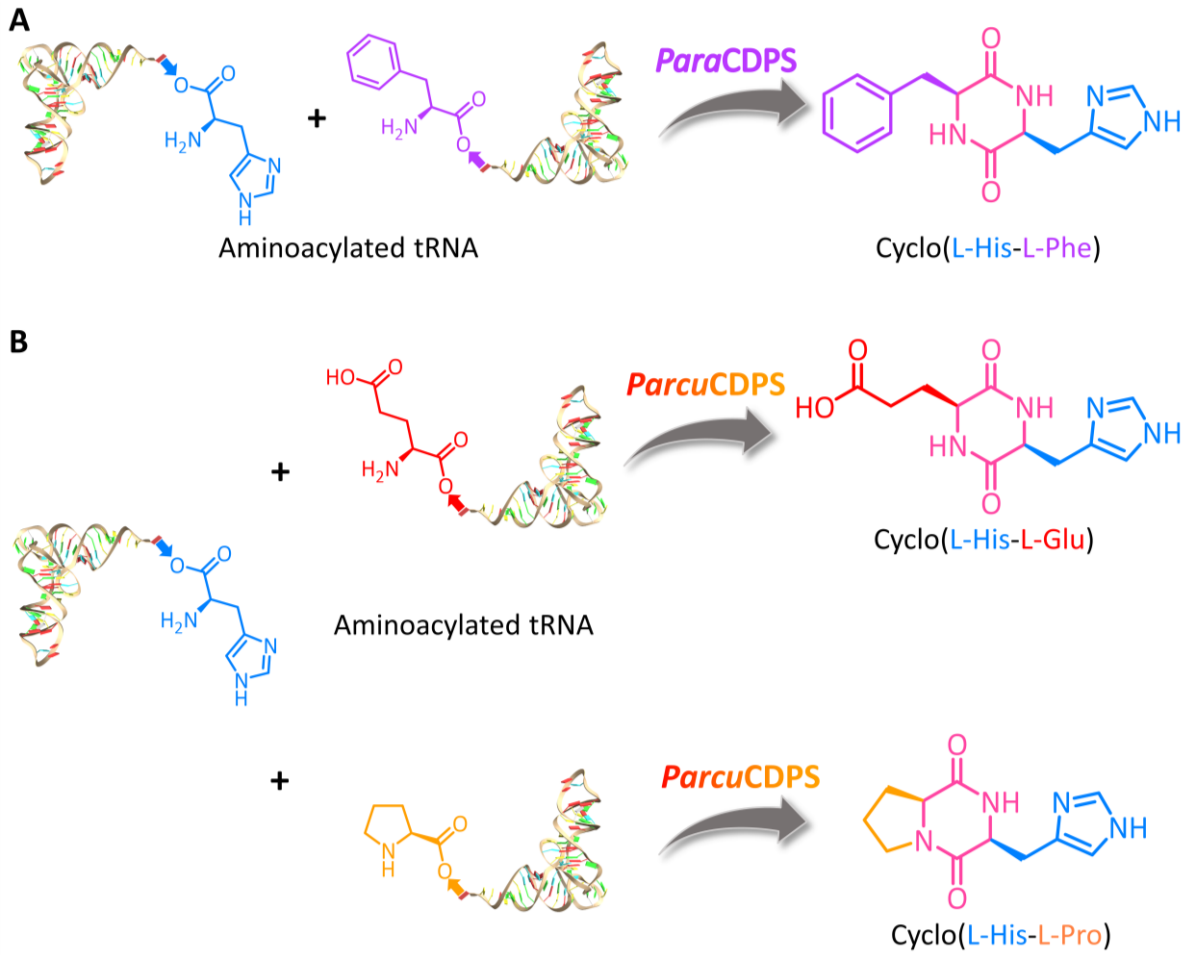


Figure 1.7 Cyclodipeptides expected from the two histidine accepting CDPSs.
A – *Para*CDPS utilises His-tRNA^{His} and Phe-tRNA^{Phe} to synthesise cyclo(L-His-L-Phe) where histidine is drawn in blue and phenylalanine in purple. **B** – *Parcu*CDPS can synthesise two cyclic dipeptides, both containing histidine (in blue): cyclo(L-His-L-Glu) with the glutamate amino acid drawn in red and cyclo(L-His-L-Pro) where the proline amino acid is in orange. The common 2,5-diketopiperazine ring is shown in pink.

Chapter 6 addresses the secondary aims of this project to determine the potential mechanism and substrate preference by cyclodipeptide oxidases, known tailoring enzymes found in some gene clusters with CDPSs. By establishing substrate specificity and key biophysical factors such as steady-state kinetic parameters and potential rate-limiting steps, further information about these enzymes could be revealed. The chemical and catalytic mechanism of members of this enzyme family are yet to be determined, thus their functional characterisation is key to understanding these enzymes. Overall, it was envisioned that this work would expand the biosynthetic toolkit available for diverse cyclic peptide production.

Chapter 2 - Materials and Methods

2.1 Reagents

All commercially available reagents were used without further purification. gBlocks™ and primers were produced by Integrated DNA technologies (IDT). DH5α (high efficiency) *E. coli,* BL21(DE3) chemically competent *E. coli.*, Gibson Assembly Cloning Kit, Q5 DNA polymerase, deoxynucleotide (dNTPs) solution mix, Dpn1, P1 nuclease, Antarctic phosphatase, T4 DNA Ligase, Polynucleotide Kinase, 6X gel purple loading dye, Quick Load Purple 1 kb Plus DNA ladder, and OneTaq Quick-Load 2X Master Mix with Standard Buffer were purchased from New England BioLabs (NEB). PCR Clean Up kit, QIAquick Gel Extraction kit and QIAprep Spin Miniprep Kit were purchased from QIAGEN. Kanamycin sulfate, Dithiothreitol (DTT), and isopropyl β-D-1-thiogalactopyranoside (IPTG) were purchased from Formedium. From Merck - formerly known as Sigma-Aldrich - lysozyme, glycerol, 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES), tris base, imidazole, sodium chloride (NaCl), magnesium chloride ($MgCl_2$), potassium chloride (KCl), 3,5-dinitrobenzyl chloride, boc-phenylalanine, agarose, ethylenediaminetetraacetic acid disodium salt dihydrate (EDTA) and adenosine 5'-triphosphate disodium salt hydrate (ATP) were purchased. DNase I was purchased from Roche. SYBR™ Safe DNA Gel Stain (10000X in DMSO stock), SYPRO™ Orange Protein Gel Stain (5,000X Concentrate in DMSO), RNaseZap™ RNase Decontamination Solution and 4X NuPAGE™ LDS Sample Buffer was purchased from Invitrogen. Acid-Phenol:Chloroform, pH 4.5 (with indole-3-acetic acid (IAA), 125:24:1), N,N'-di-boc-L-histidine dicyclohexylammonium salt, magnesium acetate tetrahydrate and FastDigest™ NdeI, XhoI, EcoRI and NcoI were purchased from Thermofisher Scientific. All commercial crystallography screens and Durham screens (Salt, pH and Osmolyte) were purchased from Molecular Dimensions. The amino acid analogues and cyclic dipeptides used in the following experiments are detailed in Table 2.1.

Table 2.1 List of all amino acids and cyclic dipeptides used, grouped together by vendor.
*(2S,4S)-4-Bromopyrrolidine-2-carboxylic acid is no longer stocked by Fluorochem.

Reagent	CAS number	Vendor
H-β-(2-Thiazolyl)-Ala-OH	134235-86-2	Bachem
L-2-Oxothiazolidine-4-carboxylic acid	19771-63-2	
4-Azido-L-phenylalanine	1241681-80-0	
4-Nitro-L-phenylalanine	949-99-5	
H-3,4-Dehydro-proline	4043-88-3	
Trans-hydroxy-l-proline (H-Hyp-OH)	51-35-4	
Cyclo(L-His-L-Pro) (cHP)	53109-32-3	
Cyclo(L-His-L-Phe) (cHF)	56586-95-9	
Cyclo(L-Leu-L-Pro) (cLP)	2873-36-1	
4-ethynyl-L-phenylalanine hydrochloride	278605-15-5	Cambridge Bioscience
Cyclo(L-Trp-L-Phe) (cWF)	6521-48-8	
(S)-4,4-Difluoropyrrolidine-2-carboxylic acid	52683-81-5	Fluorochem
(2S,4S)-4-Bromopyrrolidine-2-carboxylic acid*	16257-69-5	
(2S,4S)-4-Hydroxypyrrolidine-2-carboxylic acid	618-27-9	
(2R,4S)-4-Hydroxypyrrolidine-2-carboxylic acid	3398-22-9	
Cis-4-hydroxy-D-proline	2584-71-6	
L-Leucine	61-90-5	Merck (Sigma Aldrich)
β-(1,2,4-triazol-3-yl)-DL-alanine	10109-05-4	
1-methyl-L-histidine	332-80-9	
3-methyl-L-histidine	368-16-1	
p-Bromo-DL-phenylalanine	14091-15-7	
p-Chloro-DL-phenylalanine	7424-00-2	
Cyclo(L-Phe-L-Pro) (cFP)	3705-26-8	
(2S,4S)-4-Fluoro-pyrrolidine-2-carboxylic acid	2438-57-5	Santa Cruz Biotechnology
(2S,4R)-4-Fluoro-pyrrolidine-2-carboxylic acid	2507-61-1	
Cyclo(L-Phe-L-Gly) (cFG)	10125-07-2	
Cyclo(L-Trp-L-Gly) (cWG)	7451-73-2	
Cyclo(L-Trp-L-Trp) (cWW)	20829-55-4	
Cyclo(L-Trp-L-Tyr) (cWY)	20829-53-2	
L-Proline	147-85-3	Thermo Fisher Scientific
L-Phenylalanine	63-91-2	
L-Histidine	71-00-1	
L(+)-Glutamic acid	56-86-0	
3-(2-Pyridyl)-L-alanine dihydrochloride	1082692-96-3	
3-(3-Pyridyl)-L-alanine	64090-98-8	
3-(4-Pyridyl)-L-alanine dihydrochloride	178933-04-5	
3-(2-Thienyl)-L-alanine	22951-96-8	
4-Amino-L-phenylalanine	943-80-6	
4-Cyano-L-phenylalanine	167479-78-9	
4-Amino-cis-L-proline dihydrochloride	16257-84-4	Toronto Research Chemicals
(R)-α-Propynyl-proline hydrochloride	1049733-10-9	

2.2 Materials

Ultrafiltration spin units with a 10 kDa molecular weight cut-off (MWCO) and MF-Millipore™ membranes filters (0.8 and 0.22 µm, 45 mm diameter) were manufactured by Millipore. SnakeSkin™ dialysis tubing composed of cellulose membrane was purchased from Thermofisher Scientific. 5 mL HisTrap FF nickel columns were purchased from GE Lifesciences and the gel filtration column (HiLoad® 16/600 Superdex® 200 pg resin) was from Cytiva. Mini-PROTEAN® TGX™ 4-15% Precast Gels and Precision Plus Protein Unstained Standards were both purchased from BioRad. DNA sequencing was performed by either Mix2Seq (Eurofins Genomics) or Sanger Sequencing services (MRC PPU DNA Sequencing and Services, University of Dundee). The instructions provided by the company were followed as given for sample submission. All intact protein mass spectrometry and trypsin digest analysis were performed by the BSRC Mass Spectrometry Facility.

2.3 Equipment

Cell growth was performed in an Incubator Shaker I-26 manufactured by New Brunswick Scientific with temperature control and constant spin of 180 RPM. Cell density was measured using 1 mL of culture in a CO8000 Cell Density Meter from WPA. Large scale cell growth (1 L) was harvested using a JLA 8.1000 rotor in an Avanti J-26S XP centrifuge from Beckman Coulter and soluble proteins (post lysis) were pelleted via centrifugation in an Avanti J-20 XP centrifuge with a JA-25.50 rotor, also from Beckman Coulter. Micro centrifuge tubes were spun using an Eppendorf Centrifuge 5430R equipped with a FA-45-30-11 rotor. To centrifuge both 15 and 50 mL falcon tubes, a benchtop centrifuge 5804R with a S-4-72 rotor from Eppendorf was used. Cell lysis was induced using a high-pressure cell disruptor manufactured by Constant Systems. All protein purification was carried out using an NGC Scout Plus Liquid Chromatography System equipped with a BioFrac Fraction Collector from BioRad. Spectrometric analysis was performed using a DS-11 FX, manufactured by DeNovix. Polymerase chain reactions and small scale incubation steps were executed in a T100 thermal cycler from BioRad. Gel electrophoresis was performed in a Biorad Mini-Sub Cell GT Cell using a PowerPac™ Basic Power Supply and all gels, such as agarose or SDS-PAGE, were visualised with a Biorad ChemiDoc™ MP system.

2.4 Primers, DNA and Protein Sequence for CDPSs

2.4.1 CDPS DNA sequences

Genes encoding *Para*CDPS and *Parcu*CDPS were found using the National Centre for Biotechnology Information (Table 2.2) and with the IDT online tool, the genes were codon-optimised for expression in *E. coli*.

Table 2.2 Details of the two cyclodipeptide synthase genes cloned.

Protein	Genome assembly	Gene name	GenBank accession	UniProtKB entry
*Para*CDPS	GCA_003436875.1	DXC95_17590	EFK64745.1	A0A3E4NXU6
*Parcu*CDPS	GCA_000505065.1	O210_OD1C00001G0241	ETB63777.1	V7PYQ5

***Para*CDPS**

ATG TCT CTG CTT TAT GAA CGT CGT CTG GAA AGC TGT TAT ATC GAA CGC ATT TAT CCT TAC TCT GAG TCT AAT GCG TTT CTT GGA GTA TCC ATC TGC AGC CGT CTG TTC TCA GAG AAA ACA CTG ATC GCG TTG TTC GAC TGG TGT AAG GTA AAC GTC AAG TCA GTC TAC GTG TTA ATC GCT GAT GAA ATT CAG ATG TAC ACG TTT ATG GCT AGT AAG GGG TTG GAG CGT AAA GAA GCG TGT GCA AAG GCA CTT CAA ATT GGT GAC ATT AAA TAC CGT TTT ATT GAA CGT GTC ATT AAA AAG GGC GAT TAT GAC AAC GTC CGT TTA CTT AGT TGG AAA GCC GTG GCC TTA GAG CCG CGT TTC AAA ACA TTG TTA CAG CGT TTG CGC CTG CTG TAT GGT ACT GAA ATT CTG TTC CGT CGT GAG GTT ATT AAG CAA ATC TTA GAG CGT AAC CGT CGT TTG CCT GAG GGT TTT CGT TTT GCA CGT ATT GAC AGC TCT GAT TAT GAT CTG GCT TCG CTT TAT ATT TTG AAC GAA CTG GCT GTC ATC CTG TAC TTT CAT CTG TAC TTC GAC CCC GTT TGC CAA TAT CAG ATC TCC CCC CTG CCT ATG ACA CCT CTG CTT GAA ATC CTG TAC GAT GGG ACA TTC TTA AAG GAC TTA TTA GTA CCA AAG AAA GAT ATT GGT TAC ATT GAG ATT ATG GAG GAG AAA GAC CTT ACA GGA CGT ATG ATT AAA GTC AAC ACA AAA GGT GCA CCG TTA

***Parcu*CDPS**

ATG GAG TTA CAC CAG ATT CGT GGC TGT CAC AAA AAC GAC ATT GAG TTG AAG AAG TAC AAT ATT GGG GTC GCC ATC TCA CTT GGA AAC AAA TGG TTT TCA ATC GAC AAC ATT GAG AAA TTG GTA AAA TGG TCT TTG CTT CAC ACA AAG GAA TAT GTA ATT ATC TAC ATC GCC GAT TCC ATC CAT GGT ATC AAC CTG AGC GTC CGC AAC AAA CTG TCG GAT AGC CAC GCG GAA GAG GTC GCA ATT CGC TAC GGG CGT AAT CTG TTC ATT AAA ATT AAA GAG CGC GTC AGT TTG TCA TTC TCT CAG GAC GAG CAA GCG AAA ATC ATT TAC GCC ACG TGG AGT GAT ATC GCT GAC TCT AAA TAT AAG GAG AAG GTT AAA TAC CTT TAC AAC CTG TAC GAC AAA AAT ATC AAT TTC AAA AAT TAC ATT GAG AAC TTC GTC AAG GAG TGG GTA TCT AAG GAG AAA CGC ACT TTT AAT AAC AAC GAA ATT AAC AAG TTC GGG CGC TAT ATC TTG GAG GAA TTG CCG GAA CTG ATG GTC CAG GTC AAG GCC CGT GGA GTC CTG TTC GAG GCG TAT GTG TAC CCT TAC AAA ACT CGT ATC ACG GAG TTT GTC GGG TTG CTT CAG AAA GGG GAG ATT TTC CCC GAG ATT AAG ACA AAT ATT CTG GAT AAC CAT CCG AAG ATC TTT CTT GAG GTC CGT

2.4.2 Primer sequences

Table 2.3 Primer sequences used to amplify DNA of CDPS encoding gBlock™ genes and the pJ411 vector containing a C-terminus hexahistidine tag. The region complementary to the gBLOCK™ is highlighted in red.

Primer	Sequence	T_M (°C)
*Para*CDPS Forward	5'-ACTTTTAAGGAGGTAAACATATGTCTCTGCTTTATGAACG-3'	52-53
*Para*CDPS Reverse	5'- GTGGTGGTGGTGCTCTAACGGTGCACCTTTTGTG -3'	56-59
pJ411 Plasmid overhang *Para*CDPS Forward	5'-CACCACCACCACCACTGAAAG-3'	63
pJ411 Plasmid overhang *Para*CDPS Reverse	5'-GTTTAACTTTTAAGGAGGTAAACATATGTCTCTGCTTTATGAACG-3'	62-63
*Parcu*CDPS Forward	5'-ACTTTTAAGGAGGTAAACATATGGAGTTACACCAGATTC-3'	53-55
*Parcu*CDPS Reverse	5'-TGGTGGTGGTGGTGGTGCTCACGGACCTCAAGAAAGATC-3'	54-55
pJ411 Plasmid overhang *Parcu*CDPS Forward	5'-GAGCACCACCACCACCAC-3'	61-62
pJ411 Plasmid overhang *Parcu*CDPS Reverse	5'-ATGTTTACCTCCTTAAAAGTTAAACAAAATTATTTCTAGAG-3'	62-63

2.4.3 CDPS Protein Sequences

Protein mass and extinction coefficients at 280nm (ε280) were calculated using ProtParam[98].

Table 2.4 Protein sequences and parameters of *Para*CDPS and *Parcu*CDPS.
The penta- or hexa-histidine tag is highlighted in red.

Protein	Sequence	Molecular weight (Da)	Extinction coefficient, ε280 (M^{-1} cm^{-1})
*Para*CDPS	SLLYERRLESCYIERIYPYSESNAFLGVSICSRLFSEKT LIALFDWCKVNVKSVYVLIADEIQMYTFMASKGLERK EACAKALQIGDIKYRFIERVIKKGDYDNVRLLSWKAVA LEPRFKTLLQRLRLLYGTEILFRREVIKQILERNRRLPEGF RFARIDSSDYDLASLYILNELAVILYFHLYFDPVCQYQISP LPMTPLLEILYDGTFLKDLLVPKKDIGYIEIMEEKDLTGR MIKVNTKGAPLEHHHHH	29816	34840
*Parcu*CDPS	MELHQIRGCHKNDIELKKYNIGVAISLGNKWFSIDNIE KLVKWSLLHTKEYVIIYIADSIHGINLSVRNKLSDSHAEE VAIRYGRNLFIKIKERVSLSFSQDEQAKIIYATWSDIADS KYKEKVKYLYNLYDKNINFKNYIENFVKEWVSKEKRTFN NNEINKFGRYILEELPELMVQVKARGVLFEAYVYPYKTR ITEFVGLLQKGEIFPEIKTNILDN HPKIFLEVREHHHHHH	28115	42860

2.5 Methods for Chapter 3

Materials and methods used to gather the results presented in Chapter 3 have previously been published in: Sutherland, E., Harding, C. J., and Czekster, C. M., Active site remodelling of a cyclodipeptide synthase redefines substrate scope, *Commun Chem* **5**, 101 (2022).

2.5.1 Cloning of CDPSs via Gibson Assembly

The DNA encoding *Para*CDPS and *Parcu*CDPS (Section 2.4.1) were purchased as gBlocks™ and cloned into a modified pJ411 vector, purchased from DNA2.0, which contained an IPTG inducible T7 promoter, a kanamycin resistance gene and a carboxyl-terminal hexahistidine (6xHis) tag. Restriction free cloning was employed when first inserting the new genes into the plasmid[99]. A set of primers were designed to include complementary gene flanking regions as well as complementary plasmid flanking regions in both directions (Table 2.3). The gene inserts and vector backbones were amplified by performing a polymerase chain reaction (PCR). Primer DNA purchased from IDT were resuspended in autoclaved water to a final concentration of 100 µM and a mixture containing 10:10:80 (forward primer:reverse primer:H_2O) was prepared, resulting in a final primer concentration of 4 µM in the PCR

mixture. The optimal annealing temperature for each primer set was calculated using the online NEB T_M Calculator[100]. The reaction conditions were as described in Tables 2.5 and 2.6.

Table 2.5 Component mixture prepared for Polymerase Chain Reaction (PCR).

Reagent	Stock concentration	Final concentration	Volume (µL)
Q5 reaction buffer	5X	1X	5
dNTPs	10 mM	200 µM	0.5
gBlock™ DNA **or** Vector DNA	10 ng/µL	0.4 ng/µL	1
Primer mix	10 µM	0.5 µM	1.25
Water			To 25 µL (16.75)
Q5 High-Fidelity DNA polymerase	1 U/µL	0.02 U/µL	0.5

Table 2.6 General PCR method used for DNA amplification.

Step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	98	30	1
Denaturation	98	10	Can be repeated 25-35 times
Annealing	55-72	30	
Extension	72	20-30 seconds per kilobase of amplification	
Final extension	72	120	1
Hold	4	∞	

Subsequent PCR products were mixed with 6X Purple loading dye and loaded onto a 1% (w/v) agarose-TAE gel (1x Tris-acetate-EDTA buffer featuring 40 mM Tris pH 7.6, 20 mM acetic acid and 1 mM EDTA) stained with 1X SYBR™ Safe DNA Gel Stain. Quick Load Purple 1 kb Plus DNA Ladder (5 µL) was also added as a molecular weight reference ladder. Gel electrophoresis was performed using a Mini-Sub® Cell GT Cell (Bio-Rad) in 1X TAE buffer at a constant voltage of 100 V for 30 minutes. Visualisation of the resultant gel used the SYBR™ Safe mode on a ChemiDoc™ MP system.

The amplified vector backbone was digested by incubating 1 µL of Dpn1 in 20 µL of amplified PCR product at 37 °C for two hours to remove any methylated template plasmid DNA. Both gene insert and vector backbones were purified following the manufacturer's protocol for the PCR Clean Up kit (QIAGEN). Afterwards, Gibson Assembly was performed using a 500:100 ng ratio of Gene:Vector in 5 µL total followed by the addition of NEBuilder® HiFi DNA Assembly Master Mix (5 µL)[101]. The mixture was incubated at 50 °C for 1 hour before transforming (See Section 2.5.2 for details) the entire reaction volume into chemically competent DH5α *Escherichia coli* (*E. coli*) cells (100 µL).

2.5.2 Transformation

To transform the newly ligated DNA plasmids resulting from the Gibson Assembly process, the total 10 µL reaction volume was gently mixed with 100 µL of pre-thawed chemically competent DH5α *E. coli* cells. This mixture was incubated on ice for 30 minutes, heat-shocked at 42 °C for 1 minute using a heat block and placed back on ice for a further 5 minutes. 1 mL of lysogeny broth (LB) was added to the reaction and this was incubated at 37 °C for 1 hour with shaking at 180 RPM. The mixture was centrifuged for 1 minute at 17950 x g, 1 mL was removed and the remaining 100 µL was spread onto an LB 1.5% (w/v) agar plate supplemented with 50 µg mL^{-1} kanamycin antibiotic. The plate was incubated at 37 °C overnight in a static incubator and stored at 4 °C for later use.

2.5.3 Colony PCR

Following transformation of the desired plasmid into DH5α *E. coli*, an array of colonies were chosen for analysis using colony PCR to determine if the plasmid containing the desired insert was present. Briefly, a colony was picked, spotted onto a new LB 1.5% (w/v) agar plate containing the appropriate antibiotic and then mixed into a 25 µL total reagent mixture containing both primers for insert amplification and OneTaq Quick-Load 2X Master Mix with Standard Buffer. The exact reaction conditions are detailed in Table 2.7 and the thermocycler method in Table 2.8. The newly spotted LB 1.5% (w/v) agar plate was incubated at 37 °C overnight later use if the colonies contained the plasmid with insert.

Table 2.7 Component mixture prepared for colony PCR.

Reagent	Stock concentration	Final concentration	Volume (µL)
Forward primer	100 µM	5 µM	1.25
Reverse primer	100 µM	5 µM	1.25
OneTaq Quick-Load 2X Master Mix with Standard Buffer	2X	1X	12.5
Water			To 25

Table 2.8 Colony PCR method used in the thermocycler.

Step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	98	30	1
Denaturation	98	30	30
Annealing	55-72	30	
Extension	72	60	
Final extension	72	300	1
Hold	4	∞	

Post-thermocycling, each sample was loaded onto a 1% (w/v) agarose gel supplemented with 1X 1X SYBR™ Safe DNA Gel Stain – 2X OneTaq mix contained a green dye therefore the samples were not mixed with loading dye beforehand. Gel electrophoresis was performed in. Quick Load Purple 1 kb Plus DNA Ladder (5 µL) was also added as a molecular weight reference ladder. Gel electrophoresis was performed using a Mini-Sub® Cell GT Cell (Bio-Rad) in 1X TAE buffer at a constant voltage of 100 V for 30 minutes. Visualisation of the resultant gel used the SYBR™ Safe mode on a ChemiDoc™ MP system.

Only colonies which led to a band in the gel corresponding to the correct mass expected for the insert were used to make starter cultures from the newly spotted LB 1.5% (w/v) agar plate. LB (10 mL) with kanamycin (50 µg mL^{-1}) was inoculated with a single colony and incubated overnight at 37 °C with shaking at 180 RPM. Following this, 4 mL of the culture was harvested by centrifugation at 17950 x g for 10 minutes at 4 °C. Plasmid DNA was extracted by following the instructions from the QIAprep Spin Miniprep Kit and stored at -20 °C.

2.5.4 DNA Sequencing

Two different DNA sequencing services were used: Mix2Seq Sanger Sequencing (Eurofins Genomics) or Sanger Sequencing (MRC PPU DNA Sequencing and Services, University of Dundee). For the Mix2Seq service, 15 uL total was sent containing 50 ng μL^{-1} final concentration of plasmid DNA and 5 pmol uL^{-1} of the relevant sequencing primer (Table 2.9 for primer sequences). For the Sanger Sequencing service provided by the University of Dundee, only 30 ng μL^{-1} of plasmid DNA was required and the primer was added by the facility.

Table 2.9 Primers used to sequence DNA.

Primer	Sequence
T7 promoter	5'-TAATACGACTCACTATAGGG-3'
T7 terminator	5'-AAACCCCTCAAGACC-3'

2.5.5 Small scale protein purification trials

All expression trials were performed using BL21(DE3) *E. coli* chemically competent cells. A starter culture containing LB media (10 mL) with kanamycin (50 µg mL^{-1}) was inoculated with a single colony and grown overnight at 37 °C. The following day, 50 mL of LB supplemented with kanamycin was inoculated with 500 µL of starter culture and grown at 37 °C until the optical density at 600 nm (OD_{600}) had reached 0.6-0.8. Protein expression was induced upon the addition of 0.5 mM IPTG and the 50 mL culture was split into 4 x 10 mL samples, each in a 50 mL falcon tube. These cultures were grown at 3 different temperatures and 2 lengths of time (Table 2.10) before harvesting at 3234 x g for 10 min

at 4 °C. Each pellet was resuspended in 500 µL of buffer composed of 50 mM HEPES pH 7, 200 mM NaCl and 20 mM imidazole, and lysed via sonication for 7 s at 25% amplitude (Sonic Dismembrator, Fisherbrand). This was performed three times with the sample being placed on ice in between each round to avoid over-heating. Cell debris was isolated via centrifugation at 17950 x g for 10 minutes at 4 °C. The supernatant was analysed with MagneHis™ Ni^{2+} particles (Promega) in a BioSprint15® instrument (QIAGEN) on the Ni-NTA setting using two buffers: the wash buffer contained 50 mM HEPES pH 7, 200 mM NaCl and 20 mM imidazole, and the elution buffer contained 50 mM HEPES pH 7, 200 mM NaCl and 200 mM imidazole. The elution samples were analysed via Sodium Dodecyl-Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE, Section 2.5.6 for details) to confirm which condition appeared best for protein expression.

Table 2.10 Growth conditions tested during small scale expression trial.

Expression temperature (°C)	Expression duration
37	4 hours
25	4 hours
25	Overnight
16	Overnight

The influence of purification buffers on the isolation of the protein of interest was investigated using a diverse buffer screen. The array of different lysis and elution buffers used here were provided by Dr Tracey Gloster's lab (University of St Andrews) who kindly shared them for this trial (Table 2.11). Cells pellets harvested from the small scale growth step mentioned above were resuspended in 500 µL of each individual lysis buffer (14 in total). The cells were lysed via sonication for 7 s at 25% amplitude three times (Sonic Dismembrator, Fisherbrand) with the cells being incubated on ice in between each replicate. The lysate was centrifuged at 17950 x g for 10 minutes at 4 °C and the supernatant was analysed with MagneHis™ Ni^{2+} particles (Promega) in a BioSprint15® instrument (QIAGEN) on the Ni-NTA setting. The beads were washed with the respective lysis buffers – containing 20 mM imidazole - and the related elution buffer – containing 250 mM imidazole - for each was used to elute the protein. The elution samples were analysed via SDS-PAGE (Section 2.5.6 for details) to confirm the optimal buffer system for protein purification.

Table 2.11 Compositions of the buffers in the Gloster buffer screen.
All lysis buffers contained 20 mM imidazole and all elution buffers contained 250 mM imidazole.

Buffer	Buffer composition
1	20 mM HEPES, 150 mM NaCl, pH 7
2	PBS (137 mM NaCl, 10 mM Phosphate, 2.7 mM KCl), pH 7.4
3	20 mM Tris, 150 mM NaCl, pH 8.0
4	20 mM Tris, 500 mM NaCl, pH 8.0
5	20 mM Tris, 350 mM NaCl, pH 8.5
6	20 mM HEPES, 150 mM NaCl, pH 7, 10% (v/v) glycerol
7	PBS, pH 7.4, 10% (v/v) glycerol
8	20 mM Tris, 150 mM NaCl, pH 8.0, 10% (v/v) glycerol
9	20 mM HEPES, 150 mM NaCl, pH 7, 50 mM L-Arg, 50 mM L-Glu
10	PBS, pH 7.4, 50 mM L-Arg, 50 mM L-Glu
11	20 mM Tris, 150 mM NaCl, pH 8.0, 50 mM L-Arg, 50 mM L-Glu
12	20 mM HEPES, 150 mM NaCl, pH 7, 0.1 M L-Gly
13	PBS, pH 7.4, 0.1 M L-Gly
14	20 mM Tris, 150 mM NaCl, pH 8, 0.1 M L-Gly

2.5.6 Sodium Dodecyl-Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Protein samples were prepared for gel electrophoresis by mixing with 4X NuPAGE™ LDS Sample Buffer supplemented with 5% (v/v) 2-mercaptoethanol. The samples were boiled at 100 °C for 5 minutes before loading onto the gel. Mini-PROTEAN® TGX™ 4-15% Precast Gels were placed into a Mini-PROTEAN Tetra Cell gel tank containing 1X SDS-PAGE running buffer (25 mM Tris, 192 mM glycine and 0.1 (w/v) SDS). As a molecular weight marker, Precision Plus Protein Unstained Standards were loaded onto the gel beside the other relevant samples. 300 V was passed through the gel until the dye front had almost reached the bottom – approximately 20 minutes.

2.5.7 Large scale expression of CDPSs

Large scale expression of CDPSs typically involved between 4 – 6 litre of LB each time. A starter culture (10mL) containing LB and kanamycin (50 µg mL^{-1}) was inoculated with a previously prepared glycerol stock and grown overnight at 37 °C. 10 mL of this starter culture was transferred to 1 L of LB media supplemented with kanamycin (50 µg mL^{-1}) and grown at 37 °C until the optical density at 600nm (OD_{600}) reached a measurement between 0.6 – 0.8. Protein expression was induced with IPTG to a final concentration of 0.5 mM and grown at 16 °C overnight before cells were harvested via centrifugation at 6900 x g for 15 minutes at 4 °C. Cell pellets were collected and stored at -20 °C until further purification.

2.5.8 Protein purification using nickel affinity chromatography

All purification steps were performed at 4 °C throughout this protocol and buffers were filtered through 0.22 µm membranes before use. Previously collected cell pellets were resuspended in their respective lysis buffers (Table 2.12) using the ratio of 1 L cell culture to 30 mL lysis buffer. Lysozyme (1 mg mL^{-1}) and DNase I (0.02 mg mL^{-1}) were added to the resuspension which was stirred to homogeneity at 4 °C for 30 minutes. The cell mixture was lysed by passing it through a high pressure cell disrupter - pre-equilibrated with lysis buffer - at 30 kpsi three times. Soluble proteins were separated from cell debris via centrifugation at 32800 x g for 30 minutes at 4 °C. The supernatant post-centrifugation was vacuum filtered through a 0.8 µm membrane before being loaded onto a pre-equilibrated 5 mL HisTrap FF nickel column. The column was washed with 20 column volumes (CV) of lysis buffer and any adsorbed proteins were eluted using a stepped method of 10%, 20% and finally 100% elution buffer (See Table 2.12 for specific buffer components). Fractions containing protein were analysed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE, Section 2.5.6) to assess the molecular weight and purity of the target protein.

2.5.9 Protein Dialysis

Fractions of interest were pooled together in an appropriate length of pre-wet 10 kDa MWCO dialysis tubing and placed in 2 L of dialysis buffer and incubated at 4 °C for 3 hours (See Table 2.12 for specific buffer components). The protein dialysis bag was transferred to a new 2 L of dialysis buffer and left overnight at 4 °C. It was assumed that by following this overnight dialysis step, the resultant imidazole concentration would be less than 1 mM final.

2.5.10 Size Exclusion Chromatography

Post dialysis, the protein sample was prepared for size exclusion chromatography by concentrating to 5 mL using a 10 kDa MWCO ultrafiltration unit. The concentrated sample was loaded into a 5 mL sample loop before being injected onto a HiLoad® 16/600 Superdex® 200 pg column pre-equilibrated with the appropriate dialysis buffer. The protein was eluted in 7 mL fractions, collected after 0.25 CVs had initially passed through the column. Chosen fractions were analysed by SDS-PAGE following the protocol detailed in Section 2.5.6 and pooled, concentrated, aliquoted and flash frozen in liquid nitrogen for storage at -80 °C. The concentration was determined spectrophotometrically at 280 nm using the corresponding theoretical extinction coefficient (ε280) as calculated by the online tool, ProtParam (Expasy)[98].

Table 2.12 Purification buffers used for each CDPS.

Protein	Buffer	Buffer components
ParaCDPS	Lysis	20 mM HEPES pH 7.0, 150 mM NaCl, 20 mM imidazole, 5% (v/v) glycerol
	Elution	20 mM HEPES pH 7.0, 150 mM NaCl, 300 mM imidazole, 5% (v/v) glycerol
	Dialysis	20 mM HEPES pH 7.0, 150 mM NaCl, 5% (v/v) glycerol, 5 mM 2-mercaptoethanol
ParcuCDPS	Lysis	50 mM HEPES pH 7.0, 250 mM NaCl, 20 mM imidazole, 5% (v/v) glycerol
	Elution	50 mM HEPES pH 7.0, 250 mM NaCl, 300 mM imidazole, 5% (v/v) glycerol
	Dialysis	20 mM HEPES pH 7, 250 mM NaCl, 5 mM 2-mercapoethanol

2.5.11 Expression of CDPS in minimal media

Before cell growth, a minimal media base was made from dissolving SelenoMet™ Medium Base® powder (Molecular Dimensions) in water (21.6 g into 1 L) and autoclaving. The corresponding SelenoMet™ Medium Nutrient Mix (Molecular Dimensions) was also prepared alongside a 20 mg mL^{-1} stock of L-methionine and both were sterile filtered before use. A negative control using an empty plasmid was set up alongside the plasmids of interest containing a CDPS gene. From a starter culture in LB inoculated with a single colony from either plate containing the plasmid with the gene of interest or control plasmid, 100 µL was spiked into a mixture of minimal media (10 mL) with nutrient mix (500 µL), L-methionine (20 µL) and kanamycin (50 µg mL^{-1}). The cells were grown at 37 °C until the OD_{600} had reached between 0.6 and 0.8 whereupon protein expression was induced by the addition of 0.5 mM IPTG. Each sample was grown overnight at 16 °C and then pelleted via centrifugation at 3234 x g for 10 minutes at 4 °C. The supernatant from this was diluted 1:10 in water and trifluoroacetic acid was added to a final concentration of 2%. Cyclic dipeptides were detected via LC-MS as described in Section 2.5.18 using both HSS T3 and C18 (Waters) columns.

2.5.12 Skyline analysis of CDPS activity

LC-MS data generated from the minimal media growth experiments were subjected to analysis in Skyline. Firstly, a list of all the possible CDPs expected from the 20 proteinogenic amino acids was generated in the Czekster lab by drawing the compounds in ChemDraw 21.0.0.28 (PerkinElmer Informatics) and calculating the exact mass to 4 decimal places. Unique peaks from MS analysis were determined by comparison of the chromatograms with the negative control containing an empty plasmid. Peaks were only considered real if they showed a peak area greater than 1000 and complied with a strict ± 5 ppm deviation from the expected m/z value. Furthermore, each hit from this was confirmed by checking the isotope distribution pattern matched the natural abundance expected for the ions present.

Skyline was initially developed to tackle quantitative mass spectrometry of proteomes and improve MS methods for optimal data collection[102]. The software was designed to allow the end user to import raw data from a number of different popular mass spectrometers and perform targeted mass analysis. It was not intended to be used for the traditional nontargeted metabolomic workflow, especially since a list of expected ions from both the precursors and products must be provided before any peak analysis can be executed. Recently, Adams and colleagues published details using Skyline for small molecule detection thus strengthening the evidence that Skyline is hugely advantageous and versatile for applications in proteomics and metabolomics alike[103]. It should be noted that due to the original design focusing on proteomics, Skyline assumed ionisation only by protonation i.e. the most common ionisation form for proteins. However, for small molecules, this was not the case whereby other adducts could be observed, for example metal ions or volatile ammonium and formate ions. To circumvent this, the software used direct specification of ionisation states with the respective m/z values expected of such an adduct. Therefore, Skyline was the easiest and quickest way to analyse the data from the CDPS minimal media assay and can be concluded as a useful "vendor-neutral" tool. Using such a software to develop methods and analyse data has improved transparency and reproducibility across research groups from different institutions. Hence, it can be argued that Skyline is a valuable tool that should be utilized for biological mass spectrometry.

2.5.13 tRNA Source Preparation

2.5.13.1 In vitro transcription

The following protocol was originally published by Beckert *et al*[104]*.* tRNA sequences were found on NCBI: *E. coli*[105] and *Parcubacteria bacterium RAAC4_OD1_*1[106] (Table 2.13 for RNA sequences and Table 2.14 for primers used). There were no details of tRNAs specific to *Parabacteriodes sp. 20_3*. Template DNA encoding these RNA sequences was amplified using a polymerase chain reaction (See Table 2.15 for reaction components and Table 2.16 for thermocycler conditions) where the minimum T_M as calculated by Primerize[107,108] To perform *in vitro* transcription, the following components were added to give a reaction with a total volume of 100 µL: 47 µL PCR product (DNA yield was not quantified as it was not purified); 20 mM $MgCl_2$; 50 mM HEPES pH 7.5; 2 mM spermidine; 20 mM dithiothreitol (DTT); 5 mM ATP; 5 mM UTP; 5 mM CTP; 6 mM GTP; 5 mM RNA polymerase $\Delta 172\text{-}173_{49}$ and incubated overnight at 37 °C. DNase I was added per manufacturer's instructions and incubated for 1 hour at 37 °C.

Table 2.13 RNA Sequences of each tRNA prepared using *in vitro* transcription.
The anti-codon sequence for *Parcubacteria* tRNA is given in brackets.

tRNA	Organism	RNA Sequence
His	*E. coli*	GUGGCUAUAGCUCAGUUGGUAGAGCCCUGGAUUGUGAUU CCAGUUGUCGUGGGUUCGAAUCCCAUUAGCCACCCCA
His	*Parcubacteria bacterium RAAC4_OD1_1* (GTG)	GUGCCUAUAGUGUAAUGGUUAGCACUAGAGUUUGUGGAA CUCUCAGUCUGGGUUCAAAUCCCAGUAGGCACCA
Pro	*E. coli*	CGGUGAUUGGCGCAGCCUGGUAGCGCACUUCGUUCGGGAC GAAGGGGUCGGAGGUUCGAAUCCUCUAUCACCGACCA
Pro	*Parcubacteria bacterium RAAC4_OD1_1* (CGG)	UCGGAGUAGAAGGAUUUGAACCUCCGCCCCUUUGGCUCCU AACCAAAUACUCUAGCCAGCUGAGCUACACUCCGCCA
Phe	*E. coli*	GCCCGGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUC CCCGUGUCCUUGGUUCGAUUCCGAGUCCGGGCACCA
Glu	*E. coli*	GUCCCCUUCGUCUAGAGGCCCAGGACACCGCCCUUUCACGG CGGUAACAGGGGUUCGAAU CCCCUAGGGGACGCCA
Glu	*Parcubacteria bacterium RAAC4_OD1_1* (CTC)	UGACCUCACGGAGAAUCGAACUCCGAUUAGCGGAUUGAGA ACCCGCUGUCCUAACCGUUAGACGAUGAGGCCA
Glu	*Parcubacteria bacterium RAAC4_OD1_1* (TTC)	UGACCCUACCGGGAAUCGAACCCGGAUUACCGGCUUGAAAA GCCGAUGUCCUAACCGUUAGACGAUAAGGCCA

Table 2.14 List of primers for *in vitro* transcription.
These were designed by Primerize and used to generate the DNA sequence encoding the desired tRNA[107,108]. The letters 'F' and 'R' represent forward and reverse primers respectively.

Primer	Organism	Sequence
His 1-F	*E. coli*	TAATACGACTCACTATAGTGGCTATAGCTCAGTTGGTAGA
His 2-R		GGAATCACAATCCAGGGCTCTACCAACTGAGCTATAGCCACTATAGTGAGT
His 3-F		GCCCTGGATTGTGATTCCAGTTGTCGTGGGTTCGAATCCCAT
His 4-R		TGGGGTGGCTAATGGGATTCGAACCCACGACA
His 1-F	*Parcubacteria bacterium RAAC4_OD1_1* (GTG)	TAATACGACTCACTATAGTGCCTATAGTGTAATG
His 2-R		AGTTCCACAAACTCTAGTGCTAACCATTACACTATAGGCACTATAGTGAGTCGT
His 3-F		AGCACTAGAGTTTGTGGAACTCTCAGTCTGGGTTCAAATCCCAGTA
His 4-R		TGGTGCCTACTGGGATTTGAACCCAGA
Pro 1-F	*E. coli*	TAATACGACTCACTATAGCGGTGATTG
Pro 2-R		CTACCAGGCTGCGCCAATCACCGCTATAGTGAGTCGTATTA
Pro 3-F		GCGCAGCCTGGTAGCGCACTTCGTTCGGGACGAAGGGGTCGGAGGTTCGAATCCTCT
Pro 4-R		TGGTCGGTGATAGAGGATTCGAACCTCCGACCCCTT
Pro 1-F	*Parcubacteria bacterium RAAC4_OD1_1* (CGG)	TAATACGACTCACTATAGTCGGAGTAGAAGGAT
Pro 2-R		GCCAAAGGGGCGGAGGTTCAAATCCTTCTACTCCGACTATAGTGAGTCGT
Pro 3-F		CCGCCCCTTTGGCTCCTAACCAAATACTCTAGCCAGCTGAGCTACA
Pro 4-R		TGGCGGAGTGTAGCTCAGCTGGCTAGAGTATTTGGTTA
Phe 1-F	*E. coli*	TTCTAATACGACTCACTATATAATACGACTCACTATAGGCCCGGATAGCTCAGT
Phe 2-R		TCCCCTGCTCTACCGACTGAGCTATCCGGGCCTA
Phe 3-F		CGGTAGAGCAGGGGATTGAAAATCCCCGTGTCCTTGGTT
Phe 4-R		TGGTGCCCGGACTCGGAATCGAACCAAGGACACGGGGATTT
Glu 1-F	*E. coli*	TAATACGACTCACTATAGGTCCCCTTCGTCT
Glu 2-R		GGTGTCCTGGGCCTCTAGACGAAGGGGACCTATAGT
Glu 3-F		AGGCCCAGGACACCGCCCTTTCACGGCGGTAACAGGGGTTCGAATC
Glu 4-R		TGGCGTCCCCTAGGGGATTCGAACCCCTGTTACCGCCGTGAAA
Glu (CTC) 1-F	*Parcubacteria bacterium RAAC4_OD1_1* (CTC)	TAATACGACTCACTATAGTGACCTCACG
Glu (CTC) 2-R		CGGGTTCTCAATCCGCTAATCGGAGTTCGATTCTCCGTGAGGTCACTATAGTGAGTCGTA
Glu (CTC) 3-F		GCGGATTGAGAACCCGCTGTCCTAACCGTTAGACGATG
Glu (CTC) 4-R		TGGCCTCATCGTCTAACGGTTAGGACA
Glu (TTC) 1-F	*Parcubacteria bacterium RAAC4_OD1_1* (TTC)	TAATACGACTCACTATAGTGACCCTACCGGGAAT
Glu (TTC) 2-R		GCCGGTAATCCGGGTTCGATTCCCGGTAGGGTCACTATAGTGAG
Glu (TTC) 3-F		ACCCGGATTACCGGCTTGAAAAGCCGATGTCCTAACCGTTAGA
Glu (TTC) 4-R		TGGCCTTATCGTCTAACGGTTAGGACATCGGCTTTTCAA

Table 2.15 Reaction conditions for DNA amplification of tRNA sequences.

Reagent	Stock concentration	Final concentration	Volume (µL)
Q5 reaction buffer	5X	1X	20
dNTPs	10 mM	200 µM	2
Primers 1-F and 4-R	100 µM	2 µM	2.0
Primers 2-R and 3-F	10 µM	0.2 µM	2.0
Water			To 100 µL (69 µL)
Q5 High-Fidelity DNA polymerase	1 U/µL	0.02 U/µL	1

Table 2.16 Thermocycler method used for DNA amplification of tRNA sequences.

Step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	98	30	1
Denaturation	98	10	Repeat x30
Annealing	60	30	
Extension	72	30	
Final extension	72	600	1
Hold	4	∞	

To assess the RNA products, a urea-TBE-PAGE gel (Tris-borate-EDTA Polyacrylamide Gel Electrophoresis) was prepared using the standard gel casting protocol. Briefly, 5X TBE (45mM tris-borate and 1mM EDTA, 5 mL), 40% acrylamide (7.5 mL), urea (10.5 g) and water (5 mL) were mixed together before the addition of 10% ammonium persulfate (APS, 125 µL) and tetramethylethylenediamine (TEMED, 25 µL). The mixture was quickly stirred before being poured into empty gel cassettes (BioRad). The gels were left to harden at room temperature and subsequently placed inside a gel tank with 1X TBE buffer. 100 V was applied through the gel for 30 minutes before loading the samples to remove any residual urea. Samples for gel analysis were mixed with 2X RNA dye and boiled for 5 minutes at 100 °C. DynaMarker® Prestain Marker for Small RNA Plus ladder (Biodynamics) was loaded alongside the gel which were ran at 100 V until the bottom dye band had reached the bottom of the gel. To visualise the bands, the gel was stained with 1X SYBR™ Safe DNA Gel Stain (10000X stock in DMSO) in 1X TBE for 30 minutes before visualising.

Samples displaying a band of correct molecular weight in the gel were taken for further purification before use. Firstly, an equal volume of acid-phenol:chloroform mix pH 4.5 (125:24:1) was added to the RNA, vortexed for 2 minutes (setting 2, Genuis 3 vortex, IKA) and then centrifuged for 1 minute at 17950 x g at 4 °C. The bottom (organic) layer was carefully removed using a 1 mL syringe and needle (BD Microlance™ 3, 0.6 mm x 30 mm) and discarded. Following from this, an equal volume of 100%

chloroform was added, vortexed for 2 minutes, centrifuged for 1 minute at 17950 x g at 4 °C and the bottom layer removed. Lastly, 0.1 volumes of 3M sodium acetate and 3 volumes of cold ethanol were added to the sample and incubated overnight at -80 °C. To recover the tRNA, the samples were centrifuged for 10 minutes at 17950 x g at 4 °C. The supernatant was removed and cold 70% ethanol (1 mL) was added on top whereupon the samples were centrifuged for a further 10 minutes. The supernatant was again removed, and any residual ethanol was gently dried from the pellet under nitrogen. The resultant pellet was resuspended in an appropriate volume of DEPC-treated water and the concentration was calculated by measuring absorbance with the RNA mode on the spectrophotometer and converted to molar concentration using the extinction coefficient of the tRNA sequence as calculated by OligoEvaluator™ (Table 2.17, Sigma-Aldrich, http://www.oligoevaluator.com/LoginServlet).

Table 2.17 Parameters of each tRNA calculated by OligoEvaluator™.

tRNA	Length	MW / Da	Extinction coefficient / $mM^{-1}cm^{-1}$
His (*E. coli)*	76	24330.5	737.0
His (*Parcu)*	73	23430.1	726.5
Pro (*E. coli)*	77	24829.9	740.2
Pro *(Parcu-CGG)*	77	24510.7	726.5
Phe (*E. coli)*	76	24468.7	737.3
Glu *(E. coli)*	76	24361.7	717.8
Glu *(Parcu-CTC)*	73	23471.2	716.4
Glu *(Parcu-TTC)*	73	23454.2	716.4

2.5.13.2 tRNA Pool Extraction

The protocol was adapted from an original method by Mechulam *et al.*[109]. To begin, a starter culture was prepared by inoculating 15 mL of LB media with 15 µL of BL21(DE3) competent cells and grown overnight at 37 °C with shaking. Following this, 1 L of terrific broth (TB) was inoculated with starter culture (10 mL) and split into two separate 2 L baffled erlenmeyer flasks (Corning). These were incubated at 37 °C until the OD_{600} was greater than 4. To measure such a high optical density, a 1:10 dilution was performed prior to measurement. Cells were subsequently pelleted via centrifugation at 6900 x g for 15 minutes at 4 °C. All buffers used in this procedure were prepared using DEPC-treated water and all equipment was sprayed with RNaseZap™ RNase Decontamination Solution before use. Pellets were gently resuspended in 1mM HEPES-KOH pH 7.5 and 10mM magnesium acetate (8.6 mL) whereupon an equal volume of acid-phenol:chloroform mix, pH 4.5 with IAA (125:24:1) was added and this mixture was vortexed to form a viscous white solution. This was centrifuged at 15000 x g for 30 minutes at room temperature. The top layer containing RNA was transferred to a new falcon tube where 0.1 volumes of 5 M NaCl and 2.2 volumes of cold ethanol were added. This solution was

centrifuged at 15000 x g for 30 minutes at 4 °C and the pellet was isolated for resuspension in 5mL of 1 M NaCl. The solution was centrifuged again at 15000 x g for 30 minutes at 4 °C and the resultant supernatant was moved to a new falcon tube. To precipitate the tRNA, 2 volumes of cold ethanol was added, and the mixture was centrifuged at 15000 x g for 30 minutes at 4 °C. The pellet was resuspended in 2 mL of 100 mM HEPES-KOH at pH 8 and the tRNA was deacylated by incubating for one hour at 37 °C. Precipitation was performed by adding by 0.1 volumes of 5 M NaCl and 2.2 volumes of cold ethanol and the pellet was isolated via centrifugation at 17950 x g for 10 minutes at 4 °C. The final precipitation step involved resuspending the pellet in 2 mL of cold 70% ethanol and incubating at -80 °C for 2 days. Lastly, the solution was centrifuged at 17950 x g for 10 minutes at 4 °C and resuspended in an appropriate volume of DEPC-treated water to yield a highly concentrated stock of purified tRNA. The concentration of the tRNA stock was calculated by measuring the absorbance under the RNA mode of the spectrophotometer and using the Beer-Lambert law where the extinction coefficient used was a sum of the 4 individual extinction coefficients of the RNA nucleobases present within the stock.

2.5.13.3 S30 extract

This method was modified from Krinsky *et al.* who first published a procedure to make a bacterial lysate for the *in vitro* synthesis of proteins[110]. All buffers used in this procedure were sterile filtered and prepared using DEPC-treated water. Additionally, all equipment was sprayed with RNaseZap™ RNase Decontamination Solution before use. To begin, a starter culture was prepared by inoculating 25 mL of LB with 25 µL of BL21(DE3) cells and grown overnight at 37 °C with shaking. 1 L of TB was inoculated the following day using 20 mL of the starter culture and split into two separate 2 L baffled flasks before incubating at 37 °C until OD_{600} had reached 4. Cells were pelleted via centrifugation at 6900 x g for 15 minutes at 4 °C and the resultant pellet was resuspended in 1 L of S30 buffer containing 10 mM Tris acetate pH 7.4, 14 mM magnesium acetate, 60 mM potassium acetate, 1 mM DTT and 0.5 mL L^{-1} 2-mercapoethanol. The cells were centrifuged again at 6900 x g for 15 minutes at 4 °C and the pellet was resuspended in fresh S30 buffer (15 mL). To lyse the cells, the resuspension was passed through a high pressure cell disruptor three times at 15 kpsi. 0.1 M DTT was added to the lysate – 100 µL of DTT for every 10 mL of lysate – and this was centrifuged for 30 minutes at 32794 x g at 4 °C. The protein content of the supernatant was measured spectrometrically and this was aliquoted into appropriate volumes for storage at -80 °C.

2.5.14 Aminoacyl-tRNA synthetase purification

Overall, five different aminoacyl-tRNA synthetases (aaRSs) were required for the *in vitro* activity assays with cyclodipeptide synthase enzymes. The following plasmids containing the specific genes

encoding each specific aaRS were transformed into *E. coli* BL21(DE3) competent cells following the protocol described in Section 2.5.2:

1. HisRS - pET21a-HisRS-His was a gift from Sebastian Maerki and Takuya Ueda (Addgene plasmid # 124111 ; http://n2t.net/addgene:124111 ; RRID:Addgene_124111)[111].
2. ProRS – pJ414-ProRS was a gift from Dr Huanting Liu.
3. PheRS-A294G - Pkpy514 was a gift from David Tirrell (Addgene plasmid # 62598 ; http://n2t.net/addgene:62598 ; RRID:Addgene_62598)[112].
4. GluRS – pET21a-GluRS-His was a gift from Sebastian Maerki and Takuya Ueda (Addgene plasmid # 124109 ; http://n2t.net/addgene:124109 ; RRID:Addgene_124109)[111].
5. LeuRS – pET21a-LeuRS-His was a gift from Dr Andres Palencia at the National Institute of Health and Medical Research (Inserm)[113].

Large scale expression of aaRss typically involved the growth of 4 – 6 litres of LB each time. A starter culture (10mL) containing LB and kanamycin (50 μg mL^{-1}) was inoculated with a single colony from the transformed LB 1.5% (w/v) agar plates and grown overnight at 37 °C with shaking at 180 RPM. 10 mL of this starter culture was transferred to 1 L of LB media supplemented with kanamycin (50 μg mL^{-1}) and grown at 37 °C until the optical density at 600nm (OD_{600}) reached a measurement between 0.6 – 0.8. Protein expression was induced with IPTG to a final concentration of 0.5 mM and grown at 16 °C overnight before cells were harvested via centrifugation at 6900 x g for 15 minutes at 4 °C. A generic protein purification protocol employing a C-terminus 6xHis tag for isolation was followed as detailed below for all aaRS excluding HisRS which required specific buffers.

2.5.14.1 aaRS generic expression and purification

Cell growth was performed following the previously described method above using a starter culture to inoculate litre quantities of LB with the appropriate antibiotic. Subsequent cell pellets were resuspended in lysis buffer containing 50 mM HEPES pH 8, 500 mM NaCl and 20 mM imidazole whereupon lysozyme and DNase I were added. Cell lysis was induced using a high pressure cell disruptor at 30 kpsi three successive times. Soluble proteins were separated by centrifugation at 32800 x g for 30 minutes, 4 °C and filtered through a 0.8 μm membrane under vacuum. The lysate was applied to a 5 mL HisTrap FF nickel column after equilibrating the column with lysis buffer. Following sample application, the column was washed with 20 CVs of lysis buffer and elution was performed with a gradient of elution buffer from 10 – 100% final. Elution buffer was composed of 50 mM HEPES pH 8, 500 mM NaCl and 300 mM imidazole. Fractions with high absorbance at 280 nm from the column were analysed by SDS-PAGE (Section 2.5.6) to ensure only those containing the protein of interest were chosen for dialysis. The protein samples were dialysed into 2 L of buffer containing 20 mM Tris pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 7 mM 2-mercaptoethanol and 10% (v/v) glycerol and incubated

overnight at 4 °C. The resultant protein was concentrated to a suitable working concentration (100 or 200 µM), aliquoted, snap frozen in liquid nitrogen and stored at -80 °C for future use.

2.5.14.2 HisRS specific purification

The expression of HisRS required the growth of cells at 37 °C post-induction for 4 hours before harvesting via centrifugation. An alternative buffer system was used here, adapted from Shimizu *et al*[111]. Lysis buffer contained 50 mM HEPES-KOH pH 7.6, 10 mM $MgCl_2$, 2 mM 2-mercaptoethanol and 10 mM imidazole whilst the elution buffer was 50 mM HEPES-KOH pH 7.6, 10 mM $MgCl_2$, 2 mM 2-mercaptoethanol and 400 mM imidazole. Before resuspending the pellet, protease tablets (cOmplete Tablets, EDTA free, Roche) were added to the lysis buffer. The same cell lysis and nickel chromatography procedure was used here as described above and the isolated protein was dialysed into 50mM HEPES-KOH pH 7.6, 10 mM $MgCl_2$, 100 mM KCl, 7 mM 2-mercaptoethanol and 30% (v/v) glycerol overnight at 4 °C. Again, the final protein sample was concentrated, aliquoted, snap-frozen and stored at – 80 °C.

2.5.15 Activity assay with different sources of tRNA

To measure the activity of the CDPS enzymes *in vitro*, the following reaction systems were performed as detailed below (Table 2.18). Before the addition of any enzymes, the pH of the reaction was confirmed by spotting onto a pH strip (pH 0-14 range, PANPEHA, Whatman®) and any deviation from pH 7 was corrected using 10 M NaOH. DEPC-treated water was added to the desired final volume, and the reaction was incubated at room temperature overnight.

Table 2.18 Specific experimental setup of the *in vitro* CDPS activity assays.
AA1 and AA2 refer to amino acid 1 and amino acid 2 respectively.

***In vitro* transcribed tRNA**	
Component	**Final concentration**
HEPES pH 7.0	100 mM
KCl	100 mM
$MgCl_2$	10 mM
ATP	5 mM
DTT	10 mM
Amino acid 1	500 µM
Amino acid 2	500 µM
AA1-tRNA	5 µM
AA2-tRNA	5 µM
aaRS 1	5 µM
aaRS 2	5 µM
CDPS	5 µM

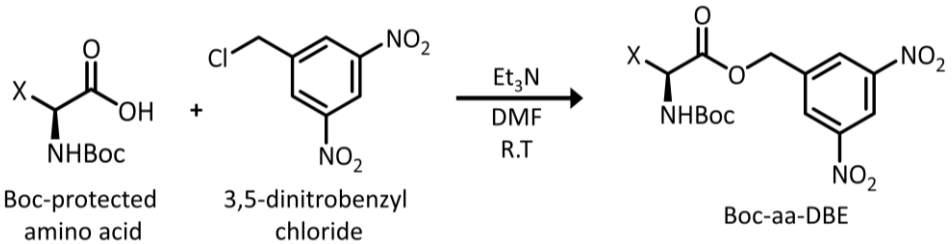
tRNA pool	
Component	**Final concentration**
HEPES pH 7.0	100 mM
KCl	100 mM
$MgCl_2$	10 mM
ATP	5 mM
DTT	10 mM
Amino acid 1	500 µM
Amino acid 2	500 µM
tRNA pool	50 µM
aaRS 1	5 µM
aaRS 2	5 µM
CDPS	5 µM

S30 extract	
Component	**Final concentration**
HEPES-KOH pH 8.0	55 mM
$MgOAc_2$	14 mM
KOAc	50 mM
NH_4OAc	155 mM
ATP	5 mM
Amino acid 1	2.5 mM
Amino acid 2	2.5 mM
S30 extract	22 mg/mL
CDPS	5 µM

2.5.16 Synthesis of small molecule substrates

The synthesis of amino acids coupled to a dinitrobenzyl ester (aa-DBE) was taken from Harding *et al.* who followed the original method from Peacock *et al* [67,114]. Pro-DBE was already available in the Czekster lab[67], and His-DBE and Phe-DBE were then synthesised following the same protocol:

Step 1. Synthesis of Boc-aa-DBE

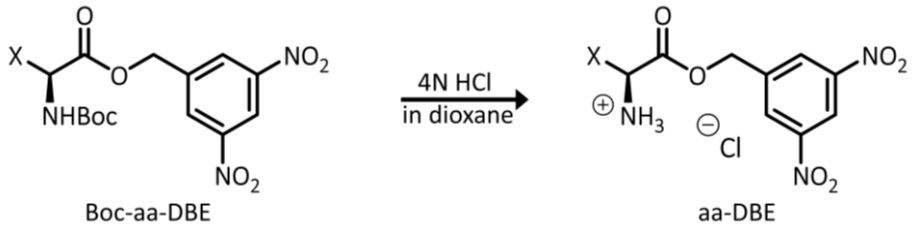


Boc-protected amino acid + 3,5-dinitrobenzyl chloride → Boc-aa-DBE

N-Boc-L-amino acid (1.97 mmol) was added to a round bottom flask with 3,5-dinitrobenzyl chloride (1.65 mmol). The reaction was placed under argon before dimethylformamide (DMF) (0.4 mL) was added to form a yellow solution. Triethylamine (Et3N, 3.98 mmol) was added last, and the reaction was stirred overnight at room temperature. The mixture was diluted in diethyl ether (25mL) and extracted with 1M hydrochloric acid (HCl) (2 x 25mL) followed by saturated sodium bicarbonate

(NaHCO3, 2 x 25mL). A brine wash was performed, and the resultant organic layer was dried over $MgSO_4$. The filtered solution was concentrated *in vacuo* to yield an orange oil.

Step 2. Boc deprotection



Boc-aa-DBE aa-DBE

Crude N-Boc-amino acid-3,5-dinitrobenzyl ester (aa-DBE) was added to a round bottom flask whereupon 4 N HCl in 1,4-dioxane (2.5 mL) was added. The mixture was stirred at room temperature for 1 hour and then the solvent removed *in vacuo*. The resulting oil was triturated with diethyl ether to yield the final compound.

His-DBE: orange solid, 39% yield, ^{1}H NMR (400 MHz, Methanol-d_4) δ 8.93 (t, J = 2.0 Hz, 1H), 8.85 (s, 1H), 8.59 (d, J = 2.0 Hz, 2H), 7.49 (s, 1H), 5.48 (d, J = 6.4 Hz, 2H), 4.57 (t, J = 6.8 Hz, 1H), 3.54 – 3.35 (m, 2H) – spectra in Appendix I.

Phe-DBE: white solid, 66% yield, ^{1}H NMR (400 MHz, Methanol-d_4) δ 8.96 (t, J = 2.1 Hz, 1H), 8.51 (d, 2H), 7.30 – 7.19 (m, 5H), 5.41 (s, 2H), 4.45 (t, J = 7.2 Hz, 1H), 3.23 (dd, J = 7.2, 2.5 Hz, 2H) – spectra in Appendix I.

2.5.17 Activity assay with aa-DBE substrates

Reactions involving aa-DBE substrates had two possible configurations: one aa-DBE substrate and one aa-tRNA substrate or two aa-DBE substrates. When both provided substrates contained the DBE moiety, the tRNA source and aminoacyl-tRNA synthetase was omitted from the reaction (Table 2.19). Here, the pH was not adjusted due to the hydrolysis of aa-DBE which was accelerated by the addition of an external base. The reactions were prepared in eppendorfs and incubated overnight at room temperature.

Table 2.19 Reaction components required for CDPS activity assay using a combination of substrates.

2 x aa-DBE	
Component	**Final Concentration**
HEPES pH 7.0	20 mM
NaCl	500 mM
DTT	10 mM
Amino acid 1-DBE	500 µM
Amino acid 2-DBE	500 µM
CDPS	5 µM

aa-DBE + aa-tRNA	
Component	**Final Concentration**
HEPES pH 7.0	100 mM
KCl	100 mM
$MgCl_2$	10 mM
ATP	5 mM
DTT	10 mM
Amino acid 1	500 µM
Amino acid 2-DBE	500 µM
tRNA pool	50 µM
aa-tRNA synthetase 1	5 µM
CDPS	5 µM

2.5.18 Cyclodipeptide identification via liquid chromatography mass spectrometry analysis

Samples containing cyclodipeptides produced by a CDPS following the protocols in either Section 2.5.15 or 2.5.17 were first quenched by the addition of cold methanol to the total volume of 80% methanol. The samples were incubated at -80 °C for 15 minutes and centrifuged at 17950 x g for 10 minutes to pellet any precipitate. The supernatant from this was transferred to a new Eppendorf which was subsequently dried under nitrogen. The resultant residue was reconstituted in 50 µL of LC-MS grade water (or equal volume to the original reaction volume).

Calibration curves for product quantification were prepared by spiking a commercial cyclic dipeptide stock into a matrix containing all reaction components excluding the cyclodipeptide synthase enzyme (Section 2.5.15 for details). This mixture was subjected to the same methanol extraction protocol as above. By doing this, any loss of CDP or interference from the reaction components was considered when preparing the linear calibration for conversion from peak area to product concentration.

All samples were analysed using a Waters ACQUITY UPLC liquid chromatography system coupled to a Xevo G2-XS QTof mass spectrometer equipped with an electrospray ionization (ESI) source. 1 µL of sample was injected onto the appropriate column and ran at 40 °C. All histidine containing compounds were injected onto a HSS-T3 column (2.1 x 100 mm, 1.8 µm, Waters Acquity) whilst non-polar CDPs such as cFP or cLP were run on a BEH C18 column (2.1 x 100 mm, 1.7 µm, Waters Acquity). CDPs were separated from the mixture using a gradient mobile phase from 1% B to 50% B where the two mobile phases consisted of A - 0.1% formic acid in water and B – 0.1% formic acid in acetonitrile at a flow rate of 400 µL min^{-1}. The capillary voltage was set at 2.5 kV in positive ion mode. The source

and desolvation gas temperatures of the mass spectrometer were set at 120 °C and 500 °C respectively. The cone gas flow was set to 50 L/hr whilst the desolvation gas flow was set at 1000 L/hr. An MS^E scan was performed between 50 – 700 m/z where function 1 employed MS analysis whilst function 2 applied a collision energy ramp from 15 to 30 V to perform MS/MS fragmentation. In addition, a lockspray signal was measured and a mass correction was applied by collecting every 10 s, averaging 3 scans of 1 s each using 50 pg mL^{-1} of Leucine Enkephalin dissolved in water:acetonitrile (50:50) and 0.1% formic acid as a standard (556.2771 m/z, Waters).

2.5.19 Time course experiment using aa-DBE substrates

Large scale reactions were prepared using aa-DBE substrates following the same method as Section 2.6.14, fixed volumes of 50 µL were removed at each required time point and quenched by adding 80% methanol. The same extraction protocol detailed in Section 2.5.18 was followed here for CDP identification. A calibration curve was employed using commercial peptide standards to quantify the product yield during the experiment. Kinetic parameters were determined by fitting the LC-MS data to an exponential one phase association relationship (Equation 2.1, GraphPad Prism 9.4.1) where *K* is the rate constant in min^{-1}.

$$y = y_0 + (Plateau - y_0)\, x\, (1 - e^{-Kx}) \qquad \text{Equation 2.1}$$

2.5.20 Analysis of trapped acyl-enzyme intermediate

Reactions of 50 µL total volume were prepared by incubating 20 µM of CDPS enzyme with one aa-DBE substrate at 1mM overnight at room temperature. For analysis, the sample was diluted to 1 µM final of enzyme in water and 10 µL was injected onto a Waters MassPrep column cartridge on a Waters G2TOF LC-MS system. The system used a two solvent system composed of solvent A - water + 1% formic acid in acetonitrile (95:5) and solvent B - water + 1% formic acid in acetonitrile (5:95). A gradient was employed from 2% B to 98% B over 4 minutes, held for 0.5 minutes and returned finally to 2% B. MS data was collected in ESI+ mode and scanned from 500–2500 m/z. The raw data were combined and processed to mass using Waters MaxEnt algorithm at 0.1 resolution using a peak width of half height of 0.4 Da. The instrument was calibrated externally against a solution of horse heart myoglobin, and an internal lock mass of Leucine Enkephalin (556.2771 m/z, Waters) was additionally used.

2.6 Methods for Chapter 4

2.6.1 Detection of aminoacylated tRNA via LC-MS analysis

Charging tRNA with the desired amino acid was performed using the following concentrations of reagents in 100 µL reaction mix: 100 mM HEPES; 100 mM KCl; 10 mM $MgCl_2$; 5 mM ATP; 10 mM DTT; 500 µM amino acid (AA); 5 µM $aaRS^{AA}$ and 50 µM $tRNA^{AA}$. This mixture was incubated at 37 °C for 1

hour before a liquid-liquid extraction was performed using an equal volume of acid-phenol:chloroform, pH 4.5 with IAA (125:124:1). The bottom layer was removed, and 0.1 reaction volumes of 3 M sodium acetate and 3 reactions volumes of ethanol added. The reaction mix was incubated at -80 °C overnight to precipitate the tRNA which was isolated via centrifugation at 17950 g for 10 minutes at 4 °C. The resultant pellet was reconstituted in 50 mM sodium acetate, pH 5.5 with 2 units of nuclease P1 which was incubated at 37 °C for 30 minutes. Following this, Antarctic phosphatase was added to the reaction in its respective buffer of 50 mM Bis-Tris-Propane-HCl pH 6, 1 mM $MgCl_2$ and 0.1 mM $ZnCl_2$. Again, this was incubated at 37 °C for 30 minutes and then quenched with the addition of methanol to a final concentration of 80%. Precipitation occurred via incubation at -80 °C for 15 minutes which was pelleted via centrifugation at 17950 x g for 10 minutes at 4 °C. The supernatant was transferred to be dried under a stream of nitrogen and the residue was reconstituted in LC-MS grade water for analysis.

The extracted reaction mixture was injected onto a HSS T3 column at 40 °C (2.1 x 100 mm, 1.8 µm, Waters Acquity) connected to a Waters ACQUITY UPLC liquid chromatography system. A flow rate of 0.3 mL min^{-1} was applied, and analytes were eluted using a gradient of two mobile phases: A – 0.1% formic acid in water and B – 100% methanol. The LC gradient was as follows: 1% B for 1 minute before starting a gradient to 60% B over 3 minutes, held at 99% B for 3 minutes to wash the column before re-equilibrating to 1% B. Ions were detected using a Xevo G2-XS QTof mass spectrometer with an electrospray ionisation source at a capillary voltage of 3 kV. During runs, the source temperature was 100 °C and the desolvation temperature was 250 °C. The cone gas flow was 50 L/Hr and the desolvation gas flow was 600 L/Hr. Lockspray (Leucine Enkephalin) was infused at 10 µL min^{-1} every 10 s and mass correction was averaged over three 1 second scans.

2.6.2 Analysis of Methylated tRNA

Methylation of tRNA catalysed by TrmK was prepared by Dr Pamela Sweeney using 100 mM HEPES pH 7.5, 3 mM $MgCl_2$, 50 µM *S*-adenosyl methionine (SAM), 20 µM *Sa*TrmK and 50 µM of the desired tRNA – $tRNA^{Leu}$, A22C-$tRNA^{Leu}$ or A11-$RNA^{18mer\ 115}$. The reaction was incubated for 20 °C for 2.5 hours, after which 0.1 reaction volumes of 3 M sodium acetate and 3 reaction volumes of ethanol were added to the mixture. Samples were incubated overnight at -80 °C to precipitate the tRNA which was separated via centrifugation at 17950 x g for 10 minutes at 4 °C. The resultant pellet was reconstituted in 50 mM sodium acetate pH 5.5, whereupon 100 units of nuclease P1 was added to a total volume of 50 µL. The reaction was incubated at room temperature for 10 minutes and subsequently quenched by the addition of cold methanol to a final concentration of 80%. Samples were incubated at -80 °C for 15 minutes and centrifuged for 10 minutes at 17950 x g, 4 °C. The resultant supernatant was collected, dried under nitrogen and the leftover resultant was resuspended in LC-MS grade water. All

samples including controls were subjected to this protocol to ensure continuity of tRNA hydrolysis and extraction.

Analysis was conducted using a Waters ACQUITY UPLC liquid chromatography system with a Xevo G2-XS QTof mass spectrometer. Two mobile phases were prepared using LC-MS grade solvents: A - 0.1% formic acid in water and B – 0.1% formic acid in acetonitrile. The samples (10 µL) were injected onto an Atlantic Premier BEH C18 AX column (2.1 x 100 mm, 1.7 µm, Waters Acquity) at 40 °C, pre-equilibrated with the starting condition of 1% B. Using a constant flow rate of 0.4 mL min^{-1}, a gradient was applied from 1% to 99% B over 9 minutes and then the column was washed with 99% B for 2 minutes before returning to the starting conditions for the next injection. Ions were detected using an electrospray ionisation source in positive mode with a capillary voltage of 2.5 kV. The source and desolvation gas temperatures of the mass spectrometer were set at 120 °C and 500 °C respectively. The cone gas flow was set to 50 L/hr whilst the desolvation gas flow was set at 800 L/hr. Masses were recorded between 50 – 700 m/z under an MS function (i.e. no collision energy applied). An external mass correction was applied using Leucine Enkephalin as a standard whose signal was collected every 10 s, averaging 3 scans of 1 s each time during the run.

2.7 Methods for Chapter 5

2.7.1 Crystallisation of *Parcu*CDPS

Initial attempts to crystallise *Parcu*CDPS in the original storage buffer conditions were unsuccessful, therefore a second purification was performed where a different dialysis buffer was used - composed of 50 mM citrate pH 6.5, 150 mM NaCl, 2 mM DTT and 2 mM EDTA. This buffer system was employed after performing differential scanning fluorimetry with the protein to identify conditions which gave optimal proteins stability. Preliminary crystallisation trials were performed using a range of commercial screens including JCSG Plus™, ProPlex™, the BSC (Basic Chemical Space) screen, PACT Premier™ and Morpheus®. Crystallisation screens and protein were dispensed into CrystalMation Intelli-Plate 96-3 low-profile plates (Hampton Research) via a Gryphon crystallisation robot (Art Robbins Instrument). 70 µL of mother liquor was added to the reservoir well and 3 concentrations of protein were prepared between 5 – 20 mg mL^{-1} which were dispensed in a 1:1 ratio (protein to reservoir, 0.3 µL total) to create a sitting drop vapour diffusion crystallisation system. For storage, plates were sealed using ClearVue™ sheets and incubated at 20 °C where crystal growth progress was monitored daily. Plates were viewed using a SteREO Discovery.V8 microscope (ZEISS). Crystallisation hits were subsequently optimised by creating small grid screens using a Scorpion Screen Builder liquid handling robot (Art Robbins Instrument). Additional sitting drop crystallisation plates were prepared

using the new optimisation screen and the most successful protein concentration from the initial plate.

2.7.2 X-ray data collection and processing

Crystals suitable for data collection were cryoprotected in mother liquor containing 15% ethylene glycol before fishing with a range of loops from 30 to 200 µm (Molecular Dimensions). Crystals were flash frozen in liquid nitrogen following successful cryoprotection. Datasets were collected either using an in-house source or an external beamline. The in-house setup composed of a Micromax-007 HF system (Rigaku) with an X-ray beam at 1.541 Å wavelength coupled to an X-stream 2000 cryogenic crystal cooler system (Rigaku) and a Saturn 944+ CCD detector (Rigaku). Data images collected from the in-house source were indexed and integrated using *MOSFLM* and scaled using AIMLESS[116]. Crystals were sent externally to Diamond Light Source, Oxford and data was collected using beamlines I03 or I04. The dataset was processed internally using ISPyB automatic pipelines such as Xia2[117] with XDS and XSCALE or using DIALS[118]. Only the datasets with suitable overall statistics (completeness, resolution, I/σI, $cc_{1/2}$, Rmerge and Rmeas) were used to solve the structure.

Initial attempts to solve the structure of *Parcu*CDPS-wild type (WT) using molecular replacement (MR) models from known CDPS structures were unsuccessful. Thus, single wavelength anomalous dispersion (SAD) phasing was employed using iodide ions. Apo crystals were dosed with potassium iodide in mother liquor (containing 15% ethylene glycol cryoprotectant) up to a final concentration of 0.5 M for one hour[119]. The anomalous signal of iodine is more pronounced at longer wavelengths therefore a dataset could be collected using the in-house source[120]. Images were collected by oscillating 0.1° in the Ω axis at two different angles: x angle of 0° for 720 images and x angle of 30° for a further 720 images. The diffraction images were indexed and integrated using *MOSFLM*[121] and scaled using AIMLESS[122] (CCP4 suite) to generate a dataset for experimental phasing using Phenix AutoSol[123] with the settings: atoms = "I"; wavelength: 1.54 Å and sites = 25. The software successfully located 26 iodide sites with the following statistics: FOM = 0.43; BAYES_CC = 56.3 ± 14.7; R-factor = 0.41 and SKEW = 0.26. The experimental data and heavy atom sites were then used for automated building with Phenix autobuild[124]. This new model of *Parcu*CDPS-WT was finally refined using Phenix.refine and WinCOOT for manual model building[125,126].

For mutant variants of *Parcu*CDPS, the solved *Parcu*CDPS-WT could be used as the model structure for molecular replacement through PHASER[127]. Briefly, the reflection file in the form of mtz; an ensemble model (here, the structure of apo *Parcu*CDPS); the expected sequence of the protein and the number of copies found in the asymmetric unit (ASU) – this number was calculated using the Matthews coefficient from XTriage – were used as inputs for PHASER[128,129]. Solutions with acceptable

statistics – Log likelihood gain, LLG > 40 and translation function z-score, TFZ > 8 - were examined in *Coot* (Crystallographic Object-Oriented Toolkit) to ascertain the reliability of the model[126]. Manual refinement was performed in *Coot* followed by rounds of refinement using Phenix.refine and corrections by PDB-redo[125,130]. Final models were confirmed by considering the validation statistics such as R_{free}, R_{work}, Ramachandran outliers and $RMS_{(bonds)}$ and $RMS_{(angles)}$.

2.7.3 Site-directed mutagenesis of *Parcu*CDPS

Primers were designed to change the sequence to the desired amino acid using an online service – NEBase changer tool (Table 2.21 for primer sequences). The NEB Q5 mutagenesis protocol was adopted here[131]. Briefly, the initial PCR step was performed following the method previously described in Section 2.5.1 and the resultant products were analysed on an 1% (w/v) agarose gel before continuing with the protocol. If the gel showed a band relating to the correct DNA molecular weight, then the kinase ligase digest was performed at room temperature for one hour using the quantities shown in Table 2.20.

Table 2.20 Reaction components for kinase ligase digest reaction.

Reagent	Stock concentration	Final concentration	Volume (µL)
PCR product	-	-	1
T4 Ligase Buffer	10X	1X	1
Dpn1	20,000 units/mL	1000 unit/mL	0.5
T4 Polynucleotide kinase	10,000 units/mL	750 units/mL	0.75
T4 ligase	400,000 units/mL	30,000 units/mL	0.75
Water	-	-	To 10 µL

The post-digestion product (10 µL) was transformed into DH5α (NEB) competent cells (Section 2.5.2) and insertion of the correct mutation was confirmed through sequencing (Section 2.5.4) before transforming the DNA into BL21 (DE3) competent cells (NEB).

Table 2.21 Site-directed mutagenesis primers as designed by online NEBase changer tool.

Primer	Sequence	T_M (°C)
Y55F Forward	5'-GTAATTATCTTCATCGCCGATTC-3'	57
Y55F Reverse	5'-ATATTCCTTTGTGTGAAGC-3'	57
Y55V Forward	5'-TGTAATTATCGTCATCGCCGATTCC-3'	56
Y55V Reverse	5'-TATTCCTTTGTGTGAAGC-3'	56
Y189F Forward	5'-TTCGAGGCGTTCGTGTACCCTTACAAAAC-3'	65
Y189F Reverse	5'-CAGGACTCCACGGGCCTT-3'	70
Y189L Forward	5'-TTCGAGGCGTTGGTGTACCCTTACAAAAC-3'	65
Y189L Reverse	5'-CAGGACTCCACGGGCCTT-3'	70
E174A Forward	5'-GAATTGCCGGCACTGATGGTC-3'	62
E174A Reverse	5'-CTCCAAGATATAGCGCCC-3'	63
E174H Forward	5'-GGAATTGCCGCACCTGATGGTCC-3'	61
E174H Reverse	5'-TCCAAGATATAGCGCCCG-3'	65
E174L Forward	5'-GGAATTGCCGCTACTGATGGTCC-3'	61
E174L Reverse	5'-TCCAAGATATAGCGCCCG-3'	65
D58N Forward	5'-CTACATCGCCAATTCCATCCATG-3'	62
D58N Reverse	5'-ATAATTACATATTCCTTTGTGTGAAG-3'	59
D58A Forward	5'-TACATCGCCGCTTCCATCCATG-3'	63
D58A Reverse	5'-GATAATTACATATTCCTTTGTGTGAAG-3'	60
E171Q Forward	5'-TATCTTGGAGCAATTGCCGGAAC-3'	62
E171Q Reverse	5'-TAGCGCCCGAACTTGTTA-3'	65
E171A Forward	5'-ATCTTGGAGGCATTGCCGGAAC-3'	63
E171A Reverse	5'-ATAGCGCCCGAACTTGTT-3'	65
Y167F Forward	5'-CAAGTTCGGGCGCTTTATCTTGGAG-3'	63
Y167F Reverse	5'-TTAATTTCGTTGTTATTAAAAGTGCGTTTCTC-3'	57
Y167A Forward	5'-CAAGTTCGGGCGCGCTATCTTGGAG-3'	57
Y167A Reverse	5'-TTAATTTCGTTGTTATTAAAAGTGCGTTTCTC-3'	57
E174N Forward	5'-GGAATTGCCGAACCTGATGGTCC-3'	61
E174N Reverse	5'-TCCAAGATATAGCGCCCG-3'	65
E174Q Forward	5'-GGAATTGCCGCAACTGATGGTC-3'	63
E174Q Reverse	5'-TCCAAGATATAGCGCCCG-3'	65
S26A Forward	5'-GGTCGCCATCGCACTTGGAAAC-3'	65
S26A Reverse	5'-5CCAATATTGTACTTCTTCAACTCAATG-3'	62
S26C Forward	5'-GTCGCCATCTGCCTTGGAAACAAATG-3'	60
S26C Reverse	5'-CCCAATATTGTACTTCTTCAAC-3'	59

2.7.4 Differential scanning fluorimetry

Protein samples were mixed with SYPRO™ Orange Protein Gel Stain in a previously determined ratio of 10X SYPRO™:10 µM protein. To assess protein stability in different buffer conditions, 10 µL of this mixture was mixed with 10 µL from a commercial buffer screen – the Durham pH, Salt or Osmolyte

screen in a 96-well plate. Thermal stability experiments were performed in triplicate using the same ratio of protein to SYPRO™ dye and made to a final volume of 20 µL with the respective protein storage buffer. Using a Stratagene Mx3005p instrument, thermal denaturation was monitored by increasing the temperature from 25 – 95 °C in 1 °C per minute increments. To calculate the melting temperature of the sample, normalised fluorescence was plotted against temperature. The data were fitted to a Boltzmann sigmoidal function (Equation 2.2) where Bottom refers to the lowest value, Top is the largest and V50 is the median of those two values.

$$y = Bottom + \frac{(Top - Bottom)}{1 + e^{\frac{V50 - x}{Slope}}} \qquad \text{Equation 2.2}$$

2.8 Methods for Chapter 6

2.8.1 *Ndas*CDO DNA Sequences

Cyclodipeptide oxidases (CDOs) are composed of two subunits found next to each other in the genome. It is known that these two genes have an overlapping region where the end of subunit A encodes for the beginning of subunit B if the open reading frame is shifted by one base pair. The CDO of interest in this project was selected from *Nocardiopsis dasonvillei* (details in Table 2.22) with subunit A and B being called *Ndas*1146 and *Ndas*1147 respectively (table 2.23 for protein sequences). The DNA sequences of both genes are given below, and the overlapping region is highlighted by the nucleotides in blue.

Table 2.22 Details of the two cyclodipeptide oxidase genes cloned into the pRSF-DUET plasmid.

Protein	Genome assembly	Gene name	NCBI Reference	UniProtKB entry
*Ndas*1146	GCA_000092985.1	D7B1W6_NOCDD	WP_013152194	D7B1W6
*Ndas*1147		D7B1W7_NOCDD	WP_013152195.1	D7B1W7

***Ndas*1146**

ATG GAC ACA GGT TCG AGC GAG CCG GAT GCG AAC CGG TGC CCC TCT CAG CGG TCA TCA CAC GCC CTA CAG ACC CTG ACC ACC CGC CGT GCC GTA CGC GCC TTC GCC GAC CGG CCG GTG GAC GAC TCC CTC CTC GAC CCC ATG CTG GAC GCC ATG CTC GCC GCC CCC TCG GCG TCC AAC AAG CAG GCG TGG GCC TTC GTC GCC GTC CGC GAG CGG CGG GCG CTG AGG CTG CTG CGC GCC TTC TCC CCC GGA ATC ATC GAA CTC CCG CCC CTG GTC GTG GCG GCC TGC TTC GAC CGC TCC CGT GCC GTG GGG GGC TCA GGC AAC TCC ACG GAC TCC GGG GAC TCC TGG GAC GAG GGC ATG CTC TGC GTC GCG ATG GCG GTG GAG AAC CTC CTC CTG GCG GCC CAC TGC CTG GGG CTG GGC GGA TGC CCG TCC GGG AGC TTT CGG AGG

GGC CCC GTC CGC AGG CTC CTG GGC CTG CCC GAC CAC CTG GAA CCC CTG CTC CTG GTT CCG ATC GGG CAC CCC GCC CGG CCA CTC GCA CCC GCA CCC CGA CGA GAC CGG AAC GAG GTG GTC AGC CAT GAG CGC TGG GGA ACC

***Ndas*1147**

ATG AGC GCT GGG GAA CCT GAG GTC CGA CAG GTC GGC GAG GAA CTC CTC CTG CTC GCC GCC TAC CTG CTC AGC AGC GGC CGC GGC CTG CTG GAC GAG CCA CGG CAG TAC GGC ACG TTC CGC TGC CTG GAC GCC GCC CGG CGC GTC CTC GCC CTC GCG GCC GGA ACC GGC CCG CAC CAC CCC GAA CTC GAC GCC CTG CGC GGT CGG ATG GAC GAC GTC ATG TGC GGG CCG ATG GGC GAC CAC GAA CTG GAC ACC CTG CTC GAC CAG ATG TGC GAG CGG CTG GCA ACC GTC CTG GAG GAT CCC GAT GTC ATC TCC GAC TGA

Table 2.23 Protein sequences and parameters of *Ndas*1146 and *Ndas*1147.
Molecular weight (Da) and Extinction coefficient (M^{-1} cm^{-1}) were both calculated by ProtParam[98]. The hexa-histidine tag introduced onto the C-terminus of *Ndas*1146 is highlighted in red.

Protein	Sequence	Molecular weight (Da)	Extinction coefficient, ε280 (M^{-1} cm^{-1})
*Ndas*1146	MDTGSSEPDANRCPSQRSSHALQTLTTRRAVRAFADRPV DDSLLDPMLDAMLAAPSASNKQAWAFVAVRERRALRLLR AFSPGIIELPPLVVAACFDRSRAVGGSGNSTDSGDSWDEG MLCVAMAVENLLLAAHCLGLGGCPSGSFRRGPVRRLLGL PDHLEPLLLVPIGHPARPLAPAPRRDRNEVVSHERWGTGS SHHHHHH	22029	16500
*Ndas*1147	MSAGEPEVRQVGEELLLLAAYLLSSGRGLLDEPRQYGTFRC LDAARRVLALAAGTGPHHPELDALRGRMDDVMCGPMG DHELDTLLDQMCERLATVLEDPDVISD	11448	2980

2.8.2 Small scale expression and purification of *Ndas*CDO

Due to time constraints, a plasmid - pRSF-DUET *Ndas*CDO - was designed to feature two related CDO genes expressed by IPTG induction and ordered from Genscript (Appendix VI). Before use, the plasmid was transformed into DH5α competent cells (Section 2.5.2) from where the DNA was extracted and sequenced (Section 2.5.4). Protein expression was initially tested by transforming the plasmid containing the two necessary CDO subunits into BL21(DE3) competent cells (Section 2.5.2). A single colony from the BL21(DE3) transformation was used to inoculate 10 mL of LB supplemented with kanamycin (50 μg mL^{-1}) which was incubated overnight at 37 °C. The following day, 50 mL of LB media with kanamycin (50 μg mL^{-1}) was inoculated from the starter culture (10 mL) and grown at 37 °C until the OD_{600} had reached 0.6. IPTG was used to induce protein expression at a final concentration

of 0.5 mM, and the cells were split into 4 x 10 mL fractions in 50 mL falcon tubes. This was to allow for different incubation temperatures and time lengths to be investigated – 37 °C for 3 hours; 25 °C for 3 hours; 25 °C overnight and 16 °C overnight. To harvest the cells post-growth, the media was centrifuged at 761 x g for 10 minutes to pellet the cells. Initial expression trials used common purification buffers - 50 mM HEPES, 250 mM NaCl and 20 mM imidazole for lysis and 50 mM HEPES, 250 mM NaCl and 300 mM imidazole for elution. The pellet was resuspended in 500 µL of lysis buffer and lysed via sonication for 7 s at 25% amplitude (Sonic Dismembrator, Fisherbrand). This was performed three times with the sample being placed on ice in between each round to avoid over-heating. Cell debris was isolated via centrifugation at 17950 x g for 10 minutes at 4 °C. The supernatant was transferred to a well containing MagneHis™ Ni^{2+} particles (Promega) and expression was investigated using a BioSprint15® instrument (QIAGEN) on the Ni-NTA setting. The relevant samples post-elution were analysed via SDS-PAGE (Section 2.5.6) to confirm which condition appeared best for protein expression.

2.8.3 Large scale expression and purification of *Ndas*CDO

Large scale expression involved first preparing a 70 mL culture composed of 70 mL of LB media supplemented with kanamycin (50 µg mL^{-1}). This culture was inoculated by scraping a previously prepared glycerol stock of pRSF-DUET *Ndas*CDO and growing overnight at 37 °C with shaking at 180 RPM. Following this, 10 mL of overnight culture was used to inoculate 1 L of terrific broth (TB) containing kanamycin (50 µg mL^{-1}). Cells were grown at 37 °C with shaking at 180 RPM until the OD_{600} was between 0.6 and 0.8 whereupon protein expression was induced with 0.5 mM IPTG. The cells were incubated overnight at 25 °C before being harvested via centrifugation at 6900 x g for 15 minutes at 4 °C. Cells were resuspended in lysis buffer composed of 100 mM Tris pH 8, 200 mM NaCl, 20 mM imidazole, 10% (v/v) glycerol and stirred to homogeneity in the presence of lysozyme and DNase I at 4 °C. Cells were lysed at high pressure using a cell disrupter set to 30 kpsi three times and the lysate was separated by centrifugation at 17600 x g for 30 minutes at 4 °C. The resultant supernatant was filtered through a 0.8 µm membrane and then loaded onto 5 mL HisTrap FF nickel column pre-equilibrated in lysis buffer. Proteins were eluted in steps of 10%, 20% and 100% buffer containing 100 mM Tris pH 8, 200 mM NaCl, 300 mM imidazole, and 10% (v/v) glycerol. Fractions of interest were confirmed by SDS-PAGE (section 2.5.6) before collecting samples together for dialysis into 50 mM Tris pH 8, 200 mM NaCl, 10% (v/v) glycerol overnight at 4 °C. The following day, the protein was concentrated using an Amicon Stirred Cell (Merck) equipped with a 30 kDa MWCO ultrafiltration disc (Millipore). This was used to concentrate protein to 10 mL which was then transferred to a Vivaspin® 6 centrifuge unit with a 100 kDa MWCO PES membrane (Sartorius). Due to the nature of this protein complex, the exact molecular weight and extinction coefficient are unknown. Therefore, the

absorbance of the flavin co-factor at 450 nm was measured and this was used to calculate the concentration of active sites (as opposed to concentration of total protein). For storage, the protein was aliquoted, flash frozen in liquid nitrogen and placed at – 80 °C.

2.8.4 Generation of oxidised cyclodipeptides by *Ndas*CDO

Experiments to investigate the substrate specificity of *Ndas*CDO were performed by preparing a 50 µL reaction containing 50 mM Tris pH 8, 300 mM NaCl, 30 µM of the desired CDP and 1 µL of *Ndas*CDO. The reaction was incubated overnight at room temperature and extracted using cold methanol to a final concentration of 80%. The samples were incubated at -80 °C for 15 minutes and centrifuged at 17950 x g for 10 minutes to pellet any precipitate. The supernatant from this was transferred to a new Eppendorf which was subsequently dried under nitrogen. The resultant residue was reconstituted in 50 µL of LC-MS grade water.

All samples were analysed using a Waters ACQUITY UPLC liquid chromatography system coupled to a Xevo G2-XS QTof mass spectrometer equipped with an electrospray ionization (ESI) source. 1 µL of sample was injected onto the appropriate column and ran at 40 °C. All histidine or tryptophan containing compounds were injected onto a HSS-T3 column (2.1 x 100 mm, 1.8 µm, Waters Acquity) whilst non-polar CDPs such as cFP or cLP were run on a BEH C18 column (2.1 x 100 mm, 1.7 µm, Waters Acquity). CDPs were separated from the mixture using a gradient mobile phase from 1% B to 50% B where the two mobile phases consisted of A - 0.1% formic acid in water and B – 0.1% formic acid in acetonitrile at a flow rate of 400 µL min^{-1}. The capillary voltage was set at 2.5 kV in positive ion mode. The source and desolvation gas temperatures of the mass spectrometer were set at 120 °C and 500 °C respectively. The cone gas flow was set to 50 L/hr whilst the desolvation gas flow was set at 1000 L/hr. An MS^E scan was performed between 50 – 700 m/z where function 1 employed MS analysis whilst function 2 applied a collision energy ramp from 15 to 30 V to perform MS/MS fragmentation. In addition, a lockspray signal was measured and a mass correction was applied by collecting every 10 s, averaging 3 scans of 1 s each using 50 pg mL^{-1} of Leucine Enkephalin dissolved in water:acetonitrile (50:50) and 0.1% formic acid as a standard (556.2771 m/z, Waters).

2.8.5 Steady State Kinetics with *Ndas*CDO

Spectrophotometric assays were used to monitor the oxidation of the carbon-carbon bonds of selective cyclic dipeptides. The pH and temperature were varied to conclude that 50 mM Tris pH 9 at 30 °C was the most optimal reaction condition. All experiments therefore used this buffer system hereafter with varying concentrations of commercial cyclic dipeptide prepared in a 96 well half-area UV-star® microplate with a transparent bottom (Greiner Bio-one). The reaction was initiated by the simultaneous addition of *Ndas*CDO to all necessary wells – the final concentration was varied to give

a measurable rate, dependent on substrate. All assays were measured using a POLARstar plate reader (Omega, BMG Labtech) with the internal incubator pre-set to 30 °C. Before the first measurement was taken, the plate was shaken in a double orbital at 300 rpm to ensure homogeneity of enzyme within the reaction mixture. Measurements were taken every 9 s for 30 minutes by monitoring wavelengths specific for the cyclic dipeptide undergoing CDO-catalysed oxidation (Table 2.24).

Table 2.24 Wavelengths monitored during steady state kinetic analysis of CDO-catalysed oxidation of various CDPs.

Cyclodipeptide substrate	Wavelength (nm)
cWW	352
cWY	352
cWF	282
cHF	358
cFP	312
cHP	270
cLP	260

All kinetic data analysis was performed using GraphPad Prism 9.4.1 (GraphPad Software, San Diego, California) for Windows. Simple linear regression was applied to the linear region of the raw data to yield gradients which could be used for Michaelis-Menten analysis. To determine the enzyme's K_M, the enzyme velocity was measured at different substrate concentrations. Plotting these variables allowed the data to be fit to the Michaelis-Menten equation where *v* represents enzyme velocity and S for substrate concentration (Equation 2.3).

$$v = \frac{V_{max}S}{(K_M + S)} \qquad \text{Equation 2.3}$$

2.8.6 pH-rate profile of *Ndas*CDO

The initial reaction velocities were measured in the presence of varying cFP concentrations at a series of defined pH values to investigate the pH dependence of the kinetic parameters. A constant *Ndas*CDO concentration of 0.1 µM was used alongside a substrate concentration range between 0.125-10 mM to give a saturation curve where the data were fitted to Equation 2.3. A mixed buffer composed of 400mM MES, Tris, CAPS and KCl was prepared allowing for a pH range between 6-13. Upon addition to reaction, the buffer was diluted to 200 mM final. All measurements were performed in triplicate and the data were analysed in GraphPad Prism 9.5.1. The data were fit to two models: 1acid-1base (Equation 2.4) or 2acid-2base (Equation 2.5). *k* is the kinetic constant – can be either k_{cat} or k_{cat}/K_M – *C* is the pH-independent value of *k*, pK_a describes acidic behaviour whilst pK_b describes the basic.

$$\log k = \log \frac{C}{\left(\left(1 + \left({}^{10^{-pH}}/_{10^{-pK_a}}\right)\right) + \left(1 + \left({}^{10^{-pK_b}}/_{10^{-pH}}\right)\right)\right)} \quad \text{Equation 2.4}$$

$$\log k = \log \frac{C}{\left(\left(1 + \left({}^{10^{-pH}}/_{10^{-pK_a^2}}\right)\right) + \left(1 + \left({}^{10^{-pK_b^2}}/_{10^{-pH}}\right)\right)\right)} \quad \text{Equation 2.5}$$

2.8.7 Purification of cFG using high performance liquid chromatography

A large scale reaction was prepared from 1 µM *Ndas*CDO and 500 µM cFG in a buffer containing 50 mM Tris, pH 9 and 200 mM NaCl. The reaction was incubated overnight at 30 °C before being extracted with cold methanol to a final volume of 80%. The solution was incubated at -80 °C for 15 minutes, centrifuged at 17950 x g for 10 minutes and the supernatant dried under N_2. The resultant residue was reconstituted in 200 µL of LC-MS grade water. High performance liquid chromatography (HPLC) was performed using a Shimadzu HPLC system coupled to a UV-vis detector equipped with a Waters BEH C18 column (2.1 x 100 mm, 1.7 µm, Waters Acquity). The column was heated to 40 °C during runs and 50 µL was injected each time. Two mobile phases were prepared: A - water + 0.1% trifluoroacetic acid and B – 100% acetonitrile. The analytes were separated using a gradient mobile phase from 1% B to 100% B over 20 minutes using a flow rate of 1 mL min^{-1} whilst monitoring the absorbance at both 257 nm and 297 nm. method collected the peak at 14.25 minutes which was suspected to be oxidised cFG (cF*G) – this was lyophilised following isolation and then resuspended in LC-MS grade water (50 µL).

2.8.8 Electron paramagnetic resonance

5mM cFG was added to 100 µM *Ndas*CDO and the reaction mixture was frozen using liquid nitrogen. Continuous wave EPR spectra were obtained at 120 K with a Bruker EMX 10/12 spectrometer running Xenon software and equipped with an ELEXSYS Super Hi-Q resonator at an operating frequency of ~9.50 GHz with 100 kHz modulation. Temperature was controlled with an ER4141 VTM Nitrogen VT unit (Bruker) operated with liquid nitrogen. CW spectra were recorded using a 20 mT field sweep centred at 336 mT, a time constant of 40.96 ms, a conversion time of 20 ms, and 1000 points resolution. An attenuation of 10.0 dB (20 mW power) and a modulation amplitude of 0.2 mT were used. CW spectra were phase- and background-corrected using the Xenon software.

Chapter 3 - Functional characterisation and cyclodipeptide production using *Para*CDPS and *Parcu*CDPS

The majority of the following results from this chapter have been published as part of a manuscript: Sutherland, E., Harding, C.J., and Czekster, C.M., 2022, *Active site remodelling of a cyclodipeptide synthase redefines substrate scope*, Communications Chemistry, **5**(101).

3.1 Introduction

The cyclodipeptide synthase family is relatively new, having first been classified in 2009[50]. Hence, there are still numerous unanswered questions about these enzymes, particularly concerning their substrate specificity and recognition pathways. High throughput assays to identify products from supernatants from cultures expressing CDPSs have previously identified 76 of the 210 possible cyclodipeptides using only proteinogenic amino acids[68]. The majority of these products incorporated aromatic or hydrophobic amino acids whilst polar or charged amino acids were rarely used as substrates. Intriguingly, two amino acids are yet to be utilised by CDPSs, lysine and aspartate, with an arginine accepting CDPS only being discovered in 2018[62]. Research has shown that the majority of CDPS enzymes can produce multiple CDP products, possessing substrate promiscuity at varied levels. Typically, one major product and several minor products are observed where one amino acid is constant throughout. Most of the characterised CDPSs to date are predominantly cFX- or cWX-producing CDPSs (where X can be any amino acid) and indeed, a lot of putative CDPSs share high sequence similarity with these known enzymes, indicative of their potential to also yield the same products[132]. Consequently, the rapidly expanding family of CDPSs would benefit from a quick and reliable method to determine the substrate specificity and potentially uncover new CDPSs with unique activities.

Considering cyclodipeptide synthases are known to exhibit a degree of sequence similarity alignments of known CDPSs with putative CDPS genes could be useful to determine activities without the need to individually purify them[64]. This alignment methodology did prove to be effective at predicting the major product of 80% unknown CDPSs, however any minor products remained unconfirmed[68]. The sequence diversity of P2 alongside its rather large surface area are partly to blame for this caveat. Recently, Canu and colleagues proposed that sequence similarity networks (SSNs) using a mixture of previously characterised and newly proposed CDPSs could indicate potential activities of 80% from a set of 775 sequences[68]. It should be noted that this analysis did suggest more than 60% of putative CDPSs share the same specificities to CDPSs with easily predictable amino acids – cYY; cWW; cCC; cGG; cGE; cAA; cAE; cLL; cLE and cIL. Additionally, rare activities, such as cPX-, cLX-

and more pertinent to this research cHX-synthesising CDPSs, were more difficult to predict from the current CDPS set, therefore certain groups remain at a disadvantage. Canu estimated that 13% of the 775 CDPSs could not have predicted activities based on sequence alignment alone[68]. This further emphasises the importance of thoroughly characterising CDPSs, especially unique sequences, which could be easily overlooked by current prediction tools.

The first study of a large subset of CDPSs discovered *Para*CDPS (reported as CDPS17 in the original publication), the first to originate from bacteroidetes[64]. *In vivo* activity investigations highlighted its interesting ability to synthesise cyclo(L-His-L-Phe). The activity of *Parcu*CDPS was also identified *in vivo* by Gondry *et al.* (referred to as CDPS90) who observed the presence of both cyclo(L-His-L-Glu) and cyclo(L-His-L-Pro) from a culture of *E. coli*[62]. Neither of these proteins were not investigated further, therefore the molecular determinants for histidine specificity remained undetermined until the work described herein.

This chapter details preliminary work with the CDPSs of interest to probe their natural substrate selection, initially using a minimal media assay which was confirmed in an *in vitro* reaction system with purified enzymes. Different sources of tRNA were prepared in-house and investigated for optimal yield of cyclic product from the CDPSs. It was initially hypothesised that creating simplified routes to tRNA would facilitate CDP production whilst not greatly impacting yield when compared to using *in vitro* transcribed tRNA. Small molecule substrates were used to explore the minimal unit required for recognition by a CDPS. Additionally, these compounds proved useful for investigating the binding order of substrates for both *Para*CDPS and *Parcu*CDPS. Overall, the following experiments expanded the characterisation of these rare CDPSs enzymes which had previously been neglected.

3.2 Cloning, expression, and small-scale purification of *Para*CDPS and *Parcu*CDPS

3.2.1 Cloning of CDPS genes

Genes encoding *Para*CDPS (DXC95_17590) and *Parcu*CDPS (O210_OD1C00001G0241) were amplified from the gBlocks™ purchased from IDT and successfully cloned into a pJ411 expression vector following the procedure in Section 2.5.1. The insertion of the amplified DNA as well as the absence of any unintentional mutations were confirmed by Sanger sequencing (Figure 3.1).

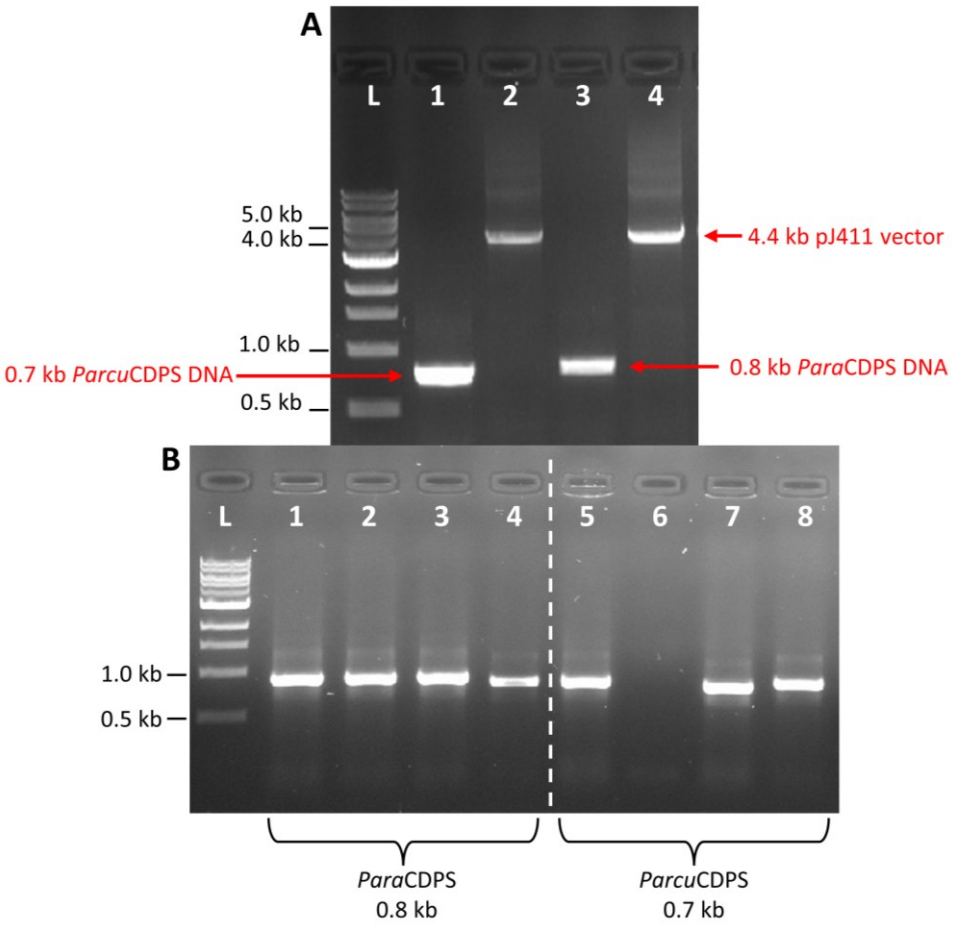


Figure 3.1 Cloning of CDPS genes.
A - Agarose gel electrophoresis of DNA fragments after PCR. Lanes are labelled as follows: L – DNA ladder; 1 – *Parcu*CDPS; 2 – pJ411 vector with *Parcu*CDPS overhang; 3 – *Para*CDPS and 4 – pJ411 vector with *Para*CDPS overhang. **B** – Colony PCR after transformation of Gibson Assembly product. Lanes are as follows: L is DNA ladder; 1-4 are *Para*CDPS colonies and 5-8 are *Parcu*CDPS colonies.

3.2.2 Small scale expression trials

Small scale expression and purification trials for both enzymes as detailed in Section 2.5.5 were performed to determine optimal growth conditions for protein expression. It was evident that growing at 16 °C overnight and using the 'common' buffer conditions featuring 50 mM HEPES pH 7 and 200mM NaCl yielded the highest levels of *Parcu*CPDS, with a band at the correct molecular weight, ≈ 28 kDa (Figure 3.2A). On the other hand, *Para*CDPS was not present in the elution fraction using the same buffer system containing 50 mM HEPES pH 7, 200 mM NaCl and 200 mM imidazole (Figure 3.2B) – this was potentially the result of several factors including unsuccessful cell lysis using sonication, lack of binding to the magnetic Ni-NTA beads or protein instability in the buffers used. Hence to explore protein stability, a diverse buffer screen was executed, kindly shared by Dr Tracey Gloster's lab (University of St Andrews) – the details of which are found in Section 2.5.5. This experiment revealed that *Para*CDPS was indeed more soluble when lysed in a select few buffer conditions highlighted in Figure 3.2C. The lane numbering relates to Table 2.11 detailing the respective buffer conditions for

each. From this experiment, future purification methods used 20 mM HEPES pH 7 and 150 mM NaCl (lane 1 on Figure 3.2C).

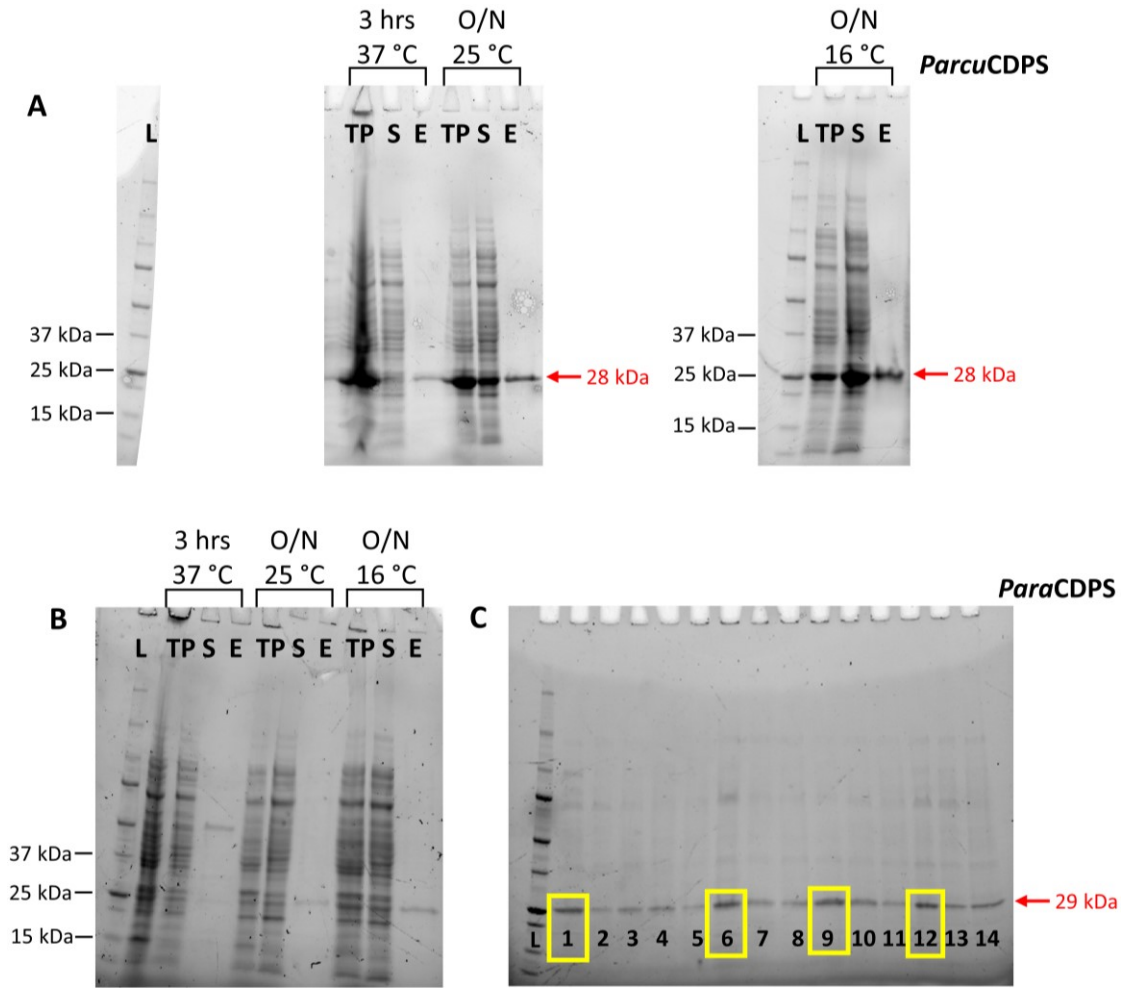


Figure 3.2 Small scale expression and purification trial of *Para*CDPS and *Parcu*CDPS.
A – SDS-PAGE analysis of *Parcu*CDPS split over two gels, hence two ladders are shown. The well labelling is as follows: L - protein standards ladder; TP – total protein; S – soluble fraction after cell lysis and E – elution from nickel affinity beads. Samples were loaded according to their growth conditions: 37 °C for 3 hours (3 hrs, 37 °C); 25 °C overnight (O/N 25 °C) and 16 °C overnight (O/N 16 °C). **B** – SDS-PAGE analysis of *Para*CDPS, the well labelling is the same as Figure A. **C** – SDS-PAGE analysis of buffer screen expression trial of *Para*CDPS grown at 16 °C post-induction. Lane numbering correlates to Table 2.11 where each buffer component is described. Expression conditions of interest are highlighted in yellow boxes: lanes 1; 6; 9 and 12.

3.2.3 Purification of *Para*CDPS

Large scale purification of *Para*CDPS was successfully performed as stated in Chapter 2 (Section 2.5.7). SDS-PAGE analysis of the elution fractions indicated protein of the expected molecular weight in fractions with high imidazole content (Figure 3.3B). Additionally, these fractions corresponded to the elution profile observed in the FPLC UV trace which was monitoring 280 nm throughout the run (Figure 3.3A). The hexa-histidine tag on the C-terminus was unable to be removed due to the absence of an appropriate cleavage site, thus further purification of the fractions from the first nickel-affinity

column was performed via size exclusion chromatography. From the size exclusion chromatography, it was evident that this protein existed as a dimer, calculated by using a calibration curve of known protein standards on the Superdex 200 16/60 gel filtration column (inset on Figure 3.3C). Protein identity and purity were verified further by intact protein mass spectrometry (Figure 3.4E). This result also highlighted the lack of the starting methionine residue as well as a missing histidine residue from the supposed hexa-histidine tag. It is possible that a methionine aminopeptidase from *E. coli* cleaved the starting methionine post-translationally, which is known to be crucial for protein stability and function[133]. Closer inspection of the primers used to amplify the pJ41 vector for cloning with ParaCDPS revealed there was one missing histidine thus leading to a penta-histidine tag instead of the typical hexahistidine tag. The histidine tag was present purely for purification purposes, therefore no further modifications were performed to rectify the missing residue, especially given that protein was successfully expressed and purified.

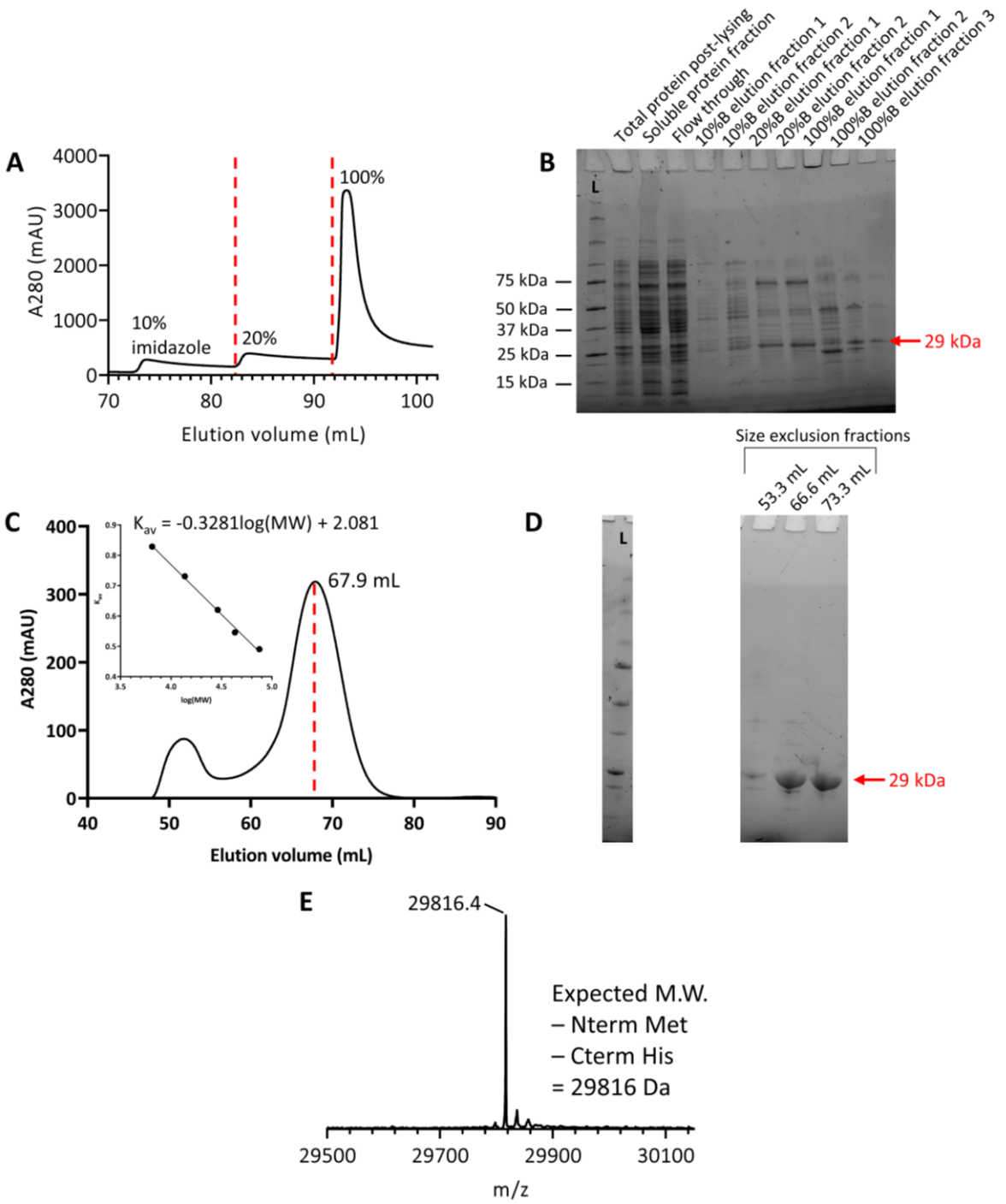


Figure 3.3 Large scale purification of *Para*CDPS.
A – UV absorbance trace of nickel-affinity chromatography of *Para*CDPS collected from the FPLC. The red dotted lines indicate the elution volume at which the imidazole concentration was increased from 10% (30 mM) imidazole to 20% (60 mM) and finally to 100% (300 mM). **B** – SDS-PAGE analysis of fractions from nickel-affinity chromatography. The expected molecular weight of *Para*CDPS is 29 kDa and expression is seen in fractions with high imidazole concentration as denoted by the red arrow. **C** – UV absorbance trace of size exclusion (Superdex 200 16/60) of dialysed *Para*CDPS fractions from nickel-affinity purification. The inset graph shows the calibration curve of standards to give a straight line with the equation: K_{av} = -0.3281log(MW) + 2.081. **D** – SDS-PAGE analysis of *Para*CDPS fractions post-elution after size exclusion chromatography. **E** – Intact protein mass spectrometry of major peak from size exclusion purification highlighting the loss of the starting methionine and a histidine from tag to give a final mass of 29816 Da.

3.2.4 Purification of *Parcu*CDPS

Using the optimal conditions determined by a small scale expression experiment*, Parcu*CDPS was successfully produced (Section 2.5.7 for details). Nickel-affinity chromatography allowed the purification of this protein which was confirmed via SDS-PAGE analysis (Figure 3.4B). Size exclusion chromatography was necessary to further purify the protein sample from contaminants identified in the SDS-PAGE. Size exclusion chromatography confirmed *Parcu*CDPS existed as a monomeric species (Figure 3.4C) and intact mass spectrometry analysis highlighted the presence of the starting methionine, unlike *Para*CDPS (Section 3.4E). Literature suggested that the methionine aminopeptidase present in *E. coli* is less efficient when the subsequent residue after the starting methionine is large or acidic[133] - which is the case for *Parcu*CDPS where the second residue is glutamate.

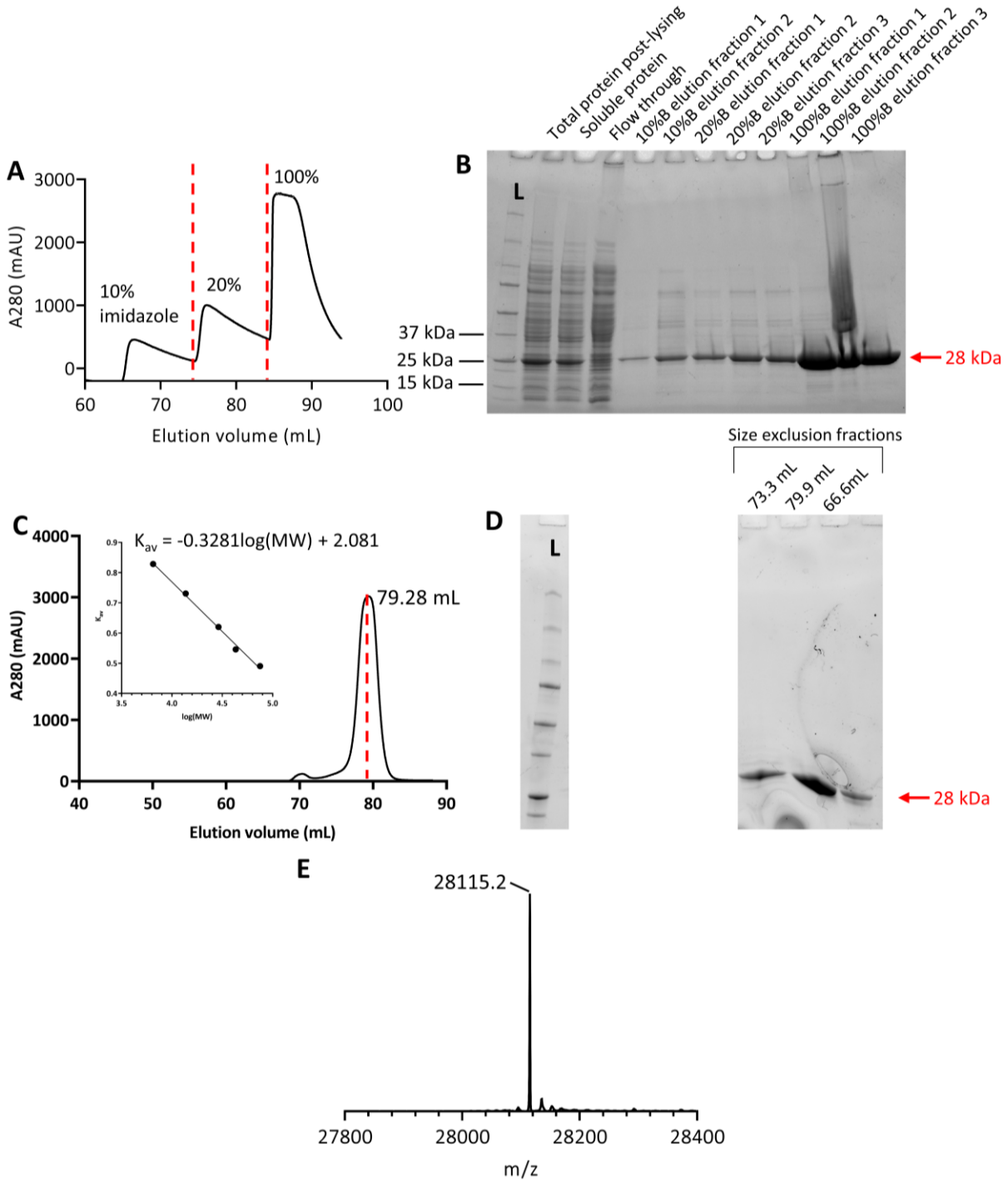


Figure 3.4 Large scale purification of *Parcu*CDPS.
A – UV absorbance trace of nickel-affinity chromatography of *Parcu*CDPS, as collected by the FPLC. The red dotted lines indicate the elution volume at which the imidazole concentration was increased from 10% (30 mM) imidazole to 20% (60 mM) and finally to 100% (300 mM). **B** – SDS-PAGE analysis of fractions from nickel-affinity chromatography. The expected molecular weight of *Parcu*CDPS is 28 kDa and overexpression is observed in the 100% B fractions, highlighted by the red arrow. **C** – UV absorbance trace of size exclusion (Superdex 200 16/60) analysis of *Parcu*CDPS after nickel-affinity purification. The inset graph shows the calibration curve of standards to give a straight line with the equation: K_{av} = -0.3281log(MW) + 2.081. **D** – SDS-PAGE analysis of *Parcu*CDPS fractions post-elution after size exclusion chromatography. **E** – Intact protein mass spectrometry of major peak from size exclusion purification highlighting the presence of purified *Parcu*CDPS.

3.3 Functional characterisation of *Para*CDPS and *Parcu*CDPS

3.3.1 *In vivo* investigations of *Para*CDPS and *Parcu*CDPS

Owing to previous research, the activities of *Para*CDPS and *Parcu*CDPS were proposed to be histidine-specific, yielding only cHF from *Para*CDPS and a mixture of cHP and cHE from *Parcu*CDPS[62,64]. Initial work therefore focused on confirming this substrate selectivity with a quick and effective *in vivo* method. This protocol was first published by Gondry *et al.* who expressed CDPSs of interest in a minimal media environment and analysed the subsequent supernatant which was hypothesised to contain cyclodipeptides produced enzymatically by the related CDPS[50]. Thus, we performed a similar experiment with *Para*CDPS and *Parcu*CDPS as well as a control containing no CDPS gene using minimal media as detailed in Section 2.5.11. Given that a large number of CDP products were possible from a single experiment, a targeted approach to HRLC-MS data analysis was adopted using Skyline by the MacCoss Lab (Section 2.5.12 for details). A list containing all possible cyclodipeptides from the combination of two natural amino acids was created featuring the expected singly protonated exact molecular weight which Skyline then extracted all possible products from each raw MS data file [134]. The control experiment allowed for the identification of truly unique peaks by direct comparison and further investigation of the mass deviation in ppm was performed to ascertain if the peaks were 'real' as opposed to false positives (Figure 3.5).

The major peak in the control cells was cyclo(L-Gly-L-Met) which was absent in the two other cultures containing plasmids with CDPS genes. This was not expected nor could its disappearance in the other experiments be explained as neither CDPS was known to accept glycine or methionine as a substrate – in fact no CDPS has been found to catalyse the formation of cGM to date, therefore there must be another factor dictating the production of this CDP by *E. coli*[68]. The major peak from cells containing *Para*CDPS appeared to be cEN which would be a unique ability as no CDPS is currently known to generate this product. Upon closer inspection, this peak was also observed in the control, therefore cEN was excluded from future analysis and the focus shifted to cHF which is in agreement with the previously detailed activity for this enzyme[62]. The presence of cHL was disregarded here, however future analyses could investigate the possibility of *Para*CDPS generating a second CDP which is known to be a common characterisation of this enzyme family. For *Parcu*CDPS, cHE was suggested to be the major product, with a peak corresponding to the expected m/z at 3.86 minutes which was absent in the control. Moreover, this experiment also highlighted the known minor product, cHP, however this was not as convincing as cHE and potentially would've been overlooked had it not been for previous research by Gondry *et al*[62]. A more thorough analysis of the data was performed by directly extracting the expected mass for the aforementioned cyclodipeptides from both the control

and reaction for comparison using MassLynx (Waters, Figure 3.6). This verified the initial hits from Skyline as the observed masses were within the allowed ppm deviations indicating this was a true signal and not some artefact.

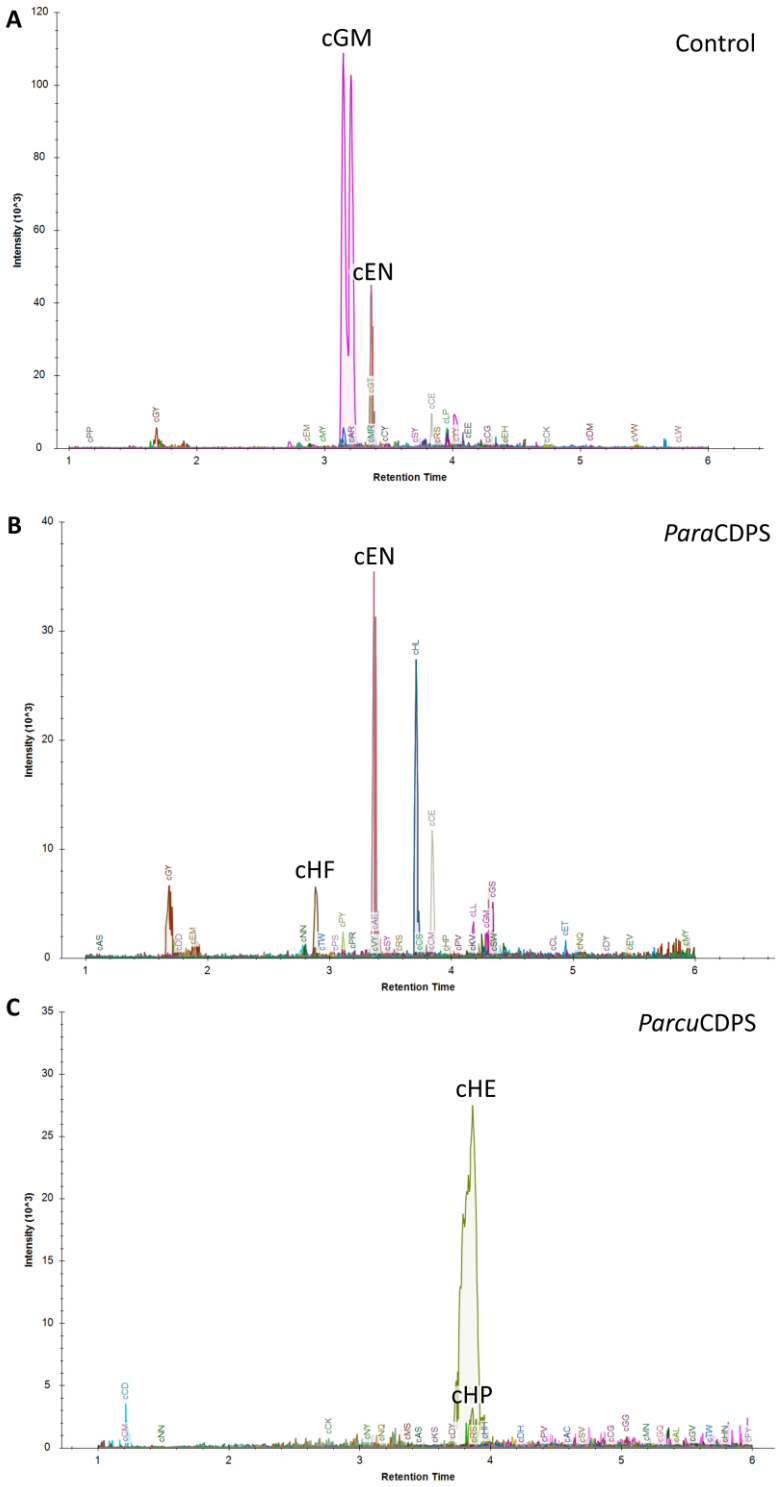


Figure 3.5 Skyline analysis of CDPS substrate selectivity *in vivo*.
Results from Skyline after peak matching to a transition list of >200 possible cyclodipeptides. **A** shows all peaks from the control cells containing an empty plasmid. Graphs **B** and **C** are from the cells grown expressing the gene encoding *Para*CDPS or *Parcu*CDPS respectively. Graphs were produced in Skyline where general editing and colour coding are restricted.

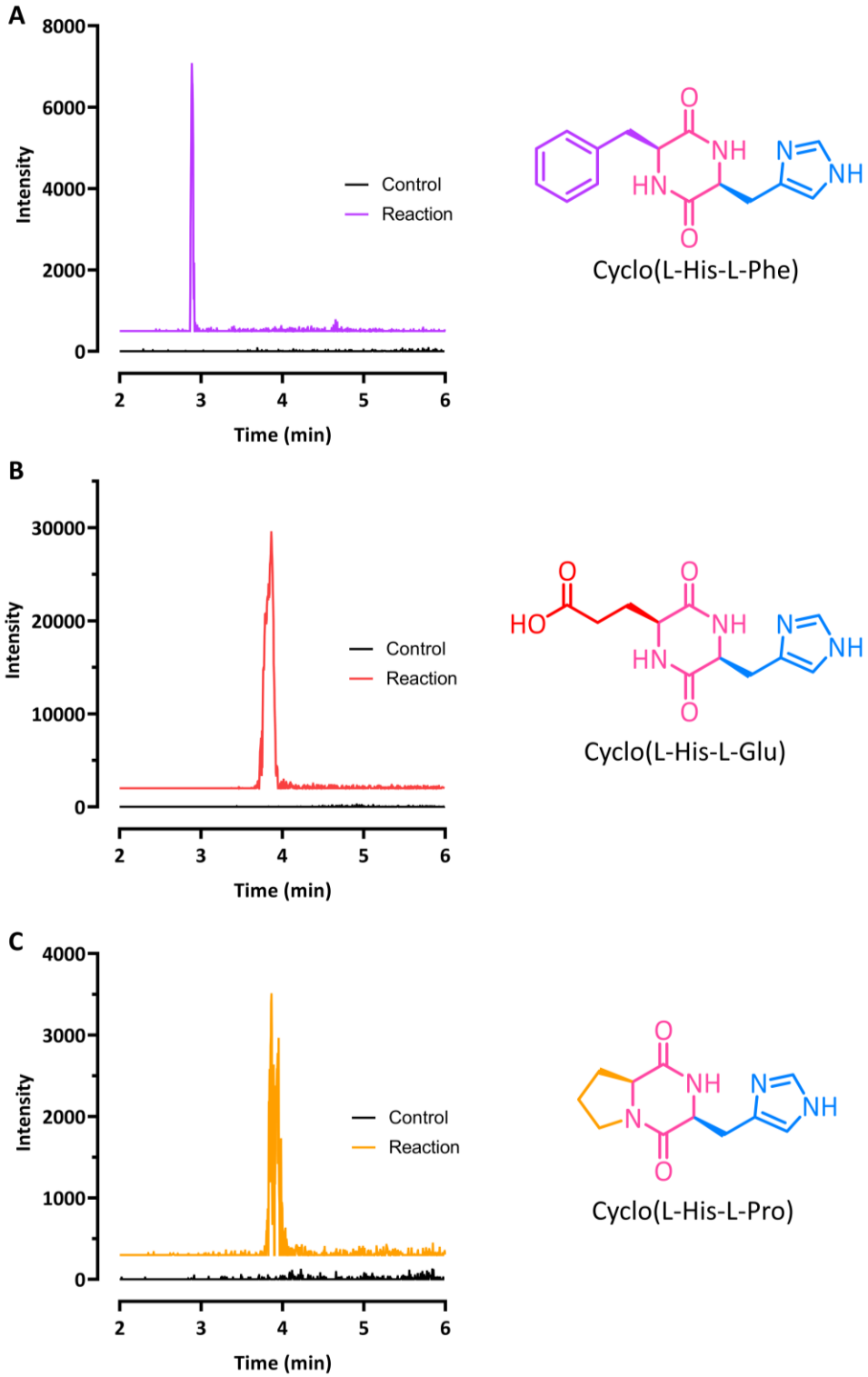


Figure 3.6 LC-MS chromatograms of CDPs from *in vivo* minimal media assay.
Each reaction is plotted individually with the empty plasmid control on the same graph for the 3 CDPs of interest (structures drawn on the right). Figure **A** shows cHF from *Para*CDPS with the reaction plotted in purple. *Parcu*CDPS produced both cHE (**B**) and cHP (**C**) drawn in red and orange respectively.

Generally, the activities observed here correlate to the reported activities for *Para*CDPS and *Parcu*CDPS. The data analysis performed was simplified by using Skyline and the results interpreted confidently by employing a strict mass deviation window for each peak (Section 2.5.12 for details).

This method, however, is not always practical and generates cyclodipeptides within a complex environment with several contaminants which could potentially mask smaller yields of peptides or complicate future purification attempts. Therefore, it was necessary to generate all reaction components individually, including tRNA and aminoacyl-tRNA synthetases, for use in a defined reaction. Employing a purified system would simplify the resultant product mixture thus allowing for controlled parameters which would improve the final yield and ease potential purification routes.

3.3.2 *In vitro* investigations of *Para*CDPS and *Parcu*CDPS

Isolation of tRNA was first required to validate potential CDPS activity *in vitro*, therefore the well-known established method of *in vitro* transcription was performed. Each respective tRNA was successfully produced following the method detailed in Section 2.5.13.1 and the expected molecular weight was confirmed using an RNA denaturing gel (Figure 3.7).

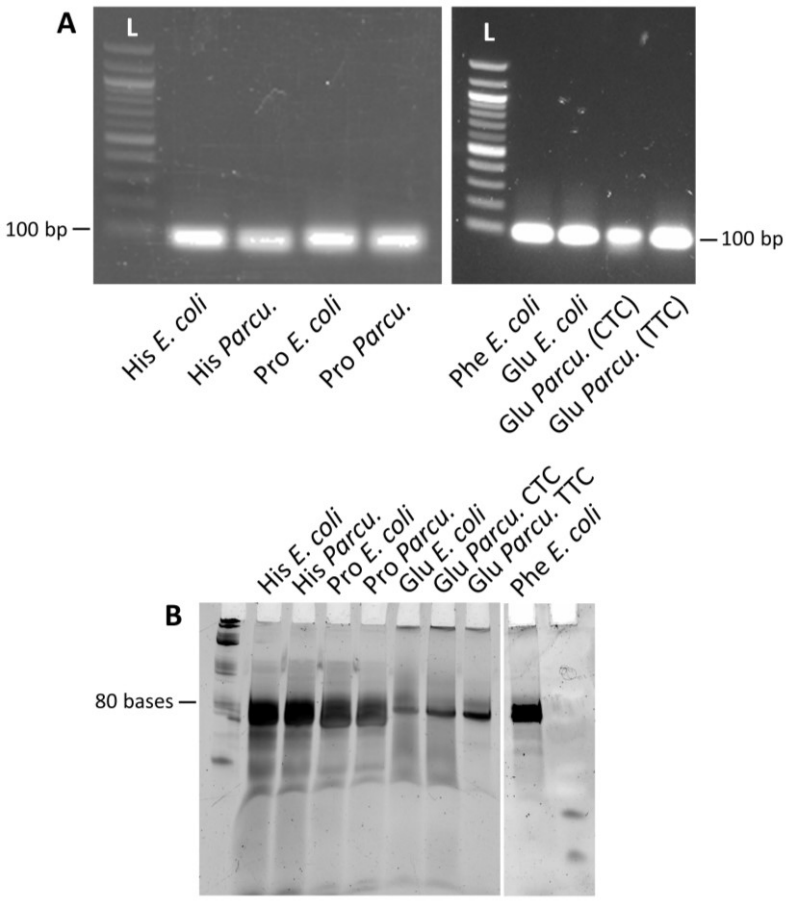


Figure 3.7 *In vitro* transcription of tRNA.
The anticodon sequence, if relevant, is shown in brackets. **A** – Agarose gel containing samples after the amplification of DNA sequences relating to the required tRNA. **B** – RNA denaturing gel after performing *in vitro* transcription. The RNA standards ladder appeared poorly resolved but the molecular weight can be inferred to be as expected – approximately 80 bases for all (Table 2.17 for exact molecular weights of each).

Following the successful purification of the desired tRNA, *in vitro* assays were designed to investigate the capabilities of *Para*CDPS and *Parcu*CDPS as catalysts for the synthesis of cyclodipeptides. Purified components including the necessary amino acids (histidine with either phenylalanine or proline), aminoacyl-tRNA synthetases (aaRS), the relevant tRNAs and the desired CDPS (full details in Section 2.5.15) were incubated together overnight at room temperature before being quenched and the CDP extracted following Section 2.5.18. This CDPS *in vitro* activity assay had been developed in the Czekster lab and tested with other members of the CDPS family to confirm the robust nature of the method. Reaction products were subsequently identified using LC-MS as described in Section 2.5.18 and the results are depicted in Figure 3.8. The presence of histidine initially posed a chromatographic issue where products were not binding to a standard reverse phase C18 column. Therefore, multiple protocols were tested using different columns, before a HSS T3 column (Waters) displayed enhanced retention of polar compounds due to the presence of a trifunctional C18 alkyl stationary phase with T3 end-capping. This column was implemented for subsequent analysis of all histidine containing CDPs.

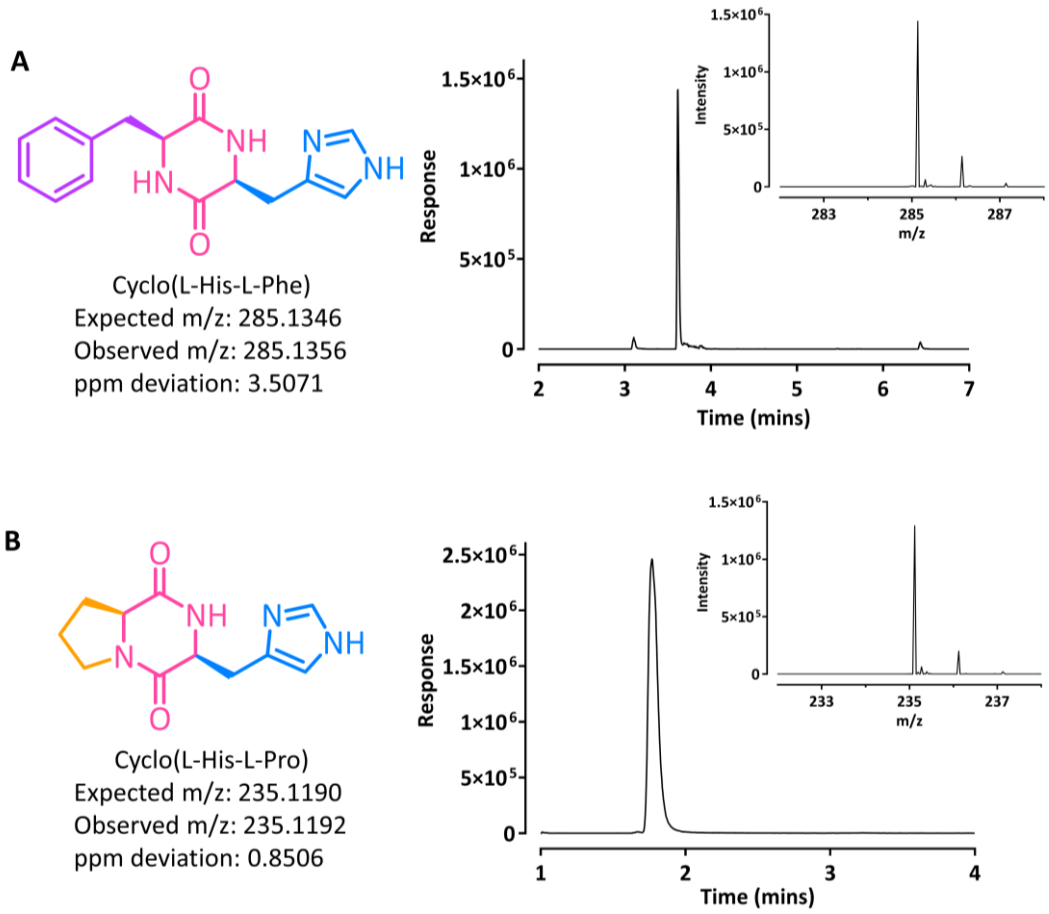


Figure 3.8 Verification of CDPS *in vitro* activity using LC-MS.
CDPSs were capable of yielding their expected CDP using *in vitro* transcribed tRNA for **A** – *Para*CDPS and **B** – *Parcu*CDPS. The extracted ion chromatogram for each is shown with the mass chromatogram as the inset (upper right of each EIC). CDP products were extracted following an overnight reaction at room temperature with 5 µM final of CDPS enzyme (Section 2.5.15 for full experimental conditions).

Whilst this method of tRNA purification was robust, it was also labour intensive, requiring the purification of each individual tRNA of interest. Consequently, alternative methods were sought which would give a higher yield of tRNA in a shorter timeframe. Firstly, an adapted cell-free protein synthesis system from *E. coli* was created, composed of all the molecular requirements to undergo coupled transcription-translation processes – known hereafter as an "S30 extract"[110]. This was hypothesised to be advantageous for the study as there would be no need to separately purify the aminoacyl-tRNA synthetases or tRNA, hence this could be used in a high-throughput manner for investigating substrate scope of unknown CDPSs. Secondly, a simpler approach containing all tRNAs from *E. coli's* protein expression machinery was established - labelled as the 'tRNA pool'.

Thus, the cyclodipeptide yield was compared using the three sources of tRNA with both CDPSs of interest here (Figure 3.9). All three methods worked successfully to give product but with varying yields. Our initial hypothesis centred on the S30 extract being the best alternative due to the ease of reaction with all the necessary components present within the mixture except ATP and amino acids. However, it was found that the high complexity of this mixture, possibly due to the presence of endogenous proteins, RNA, and ribosomes from *E. coli* was detrimental to overall product yield, resulting in large background noise. Overall, it was evident that using the tRNA pool yielded the most product after converting peak area to concentration using a calibration curve for quantification. Therefore, this new method of utilising all tRNA available from *E. coli* is a simpler and facile route that could be of benefit to areas outside of this realm - more specifically for enzymes which also utilise tRNAs as substrates such as aa-tRNA dependent transferases[135].

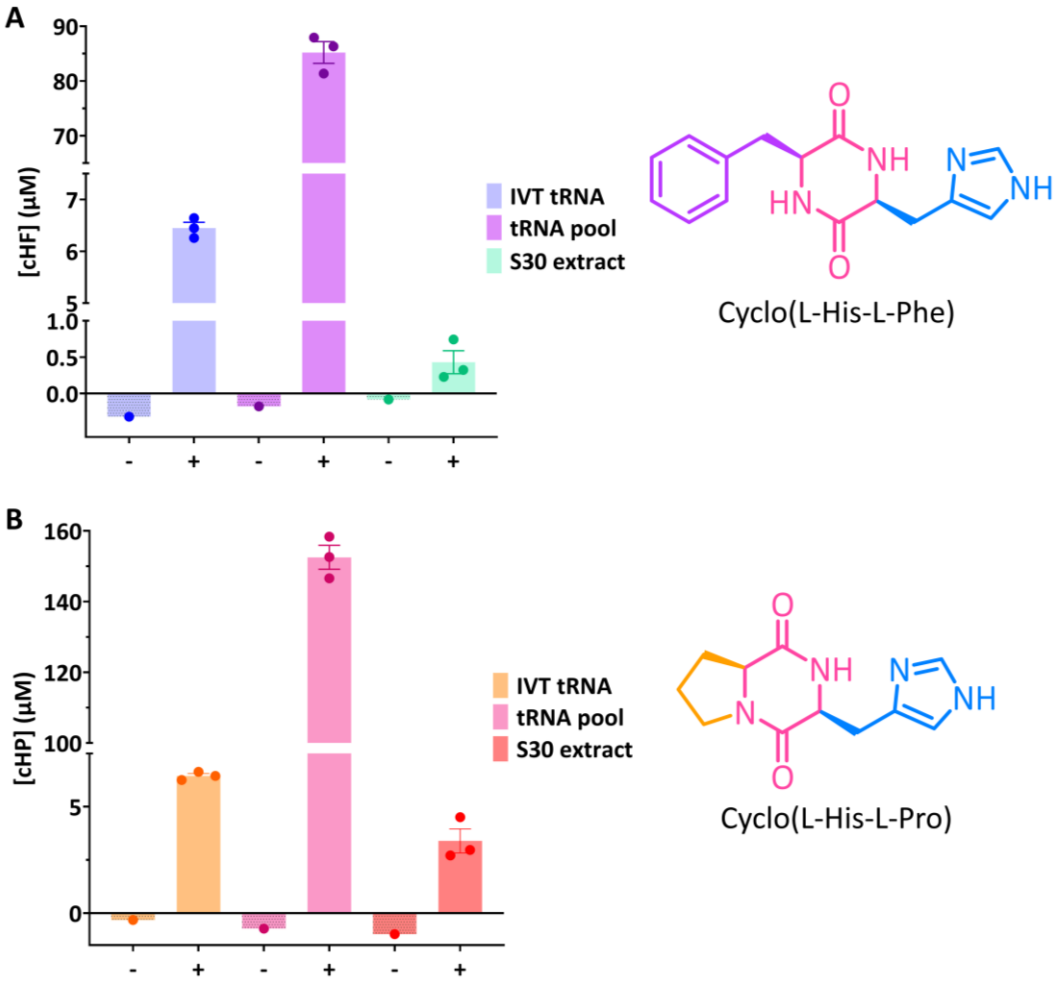


Figure 3.9 Product yield comparison between different tRNA sources.
Controls and reactions are signified by a minus (-) or plus (+) sign respectively. **A** – Quantification of cHF using *Para*CDPS with different sources of tRNA: *in vitro* transcribed (IVT) tRNA (blue); tRNA pool (purple) and S30 extract (green). **B** – Quantification of cHP using *Parcu*CDPS with different sources of tRNA: *in vitro* transcribed (IVT) tRNA (orange); tRNA pool (pink) and S30 extract (red). All LC-MS experiments were performed in triplicate with each data point represented by a circle with the standard error of the mean shown as error bars.

3.4 Use of alternative small molecule substrates

Recent work from our group showcased the use of an activated amino acid substrate to circumvent the need for aminoacylated tRNA. This research revealed the minimal substrate requirements for recognition by a CDPS involved an amino acid coupled to a dinitrobenzyl (DBE) group via an ester bond[67]. However, this is not the first time a dinitrobenzyl ester has been used to activate amino acids. Previously, Murakami detailed the use of amino acid-DBE (aa-DBE) substrates by an aminoacyl-tRNA synthetase-like RNA enzyme, known as Flexizyme[136]. Therefore, it was hypothesised these substrates would also work with cyclodipeptide synthases. It should be highlighted that prior attempts using aa-DBE as substrates for CDPSs only required the use of one activated amino acid, given that the CDPS

produced a symmetrical cyclic product. Thus, it was expected that these histidine specific CDPSs would have additional barriers to overcome in order to yield product.

Both CDPSs investigated here use two different amino acids as substrates i.e. *Para*CDPS uses histidine and phenylalanine whilst *Parcu*CDPS uses histidine and proline. Therefore, His-DBE, Phe-DBE and Pro-DBE were synthesised following a published method by Peacock *et al*[114]. Considering both aa-tRNA and aa-DBE substrates, various combinations of different substrates could be used with the CDPSs - for example, one aa-DBE could be used with one aa-tRNA substrate where the respective amino acid and its cognate aaRS were present in the mixture (Figure 3.10). On the other hand, reactions using solely aa-DBEs with a CDPS would not require the use of any tRNA containing elements, which also including ATP used for aminoacylation (section 2.5.17 for reaction details).

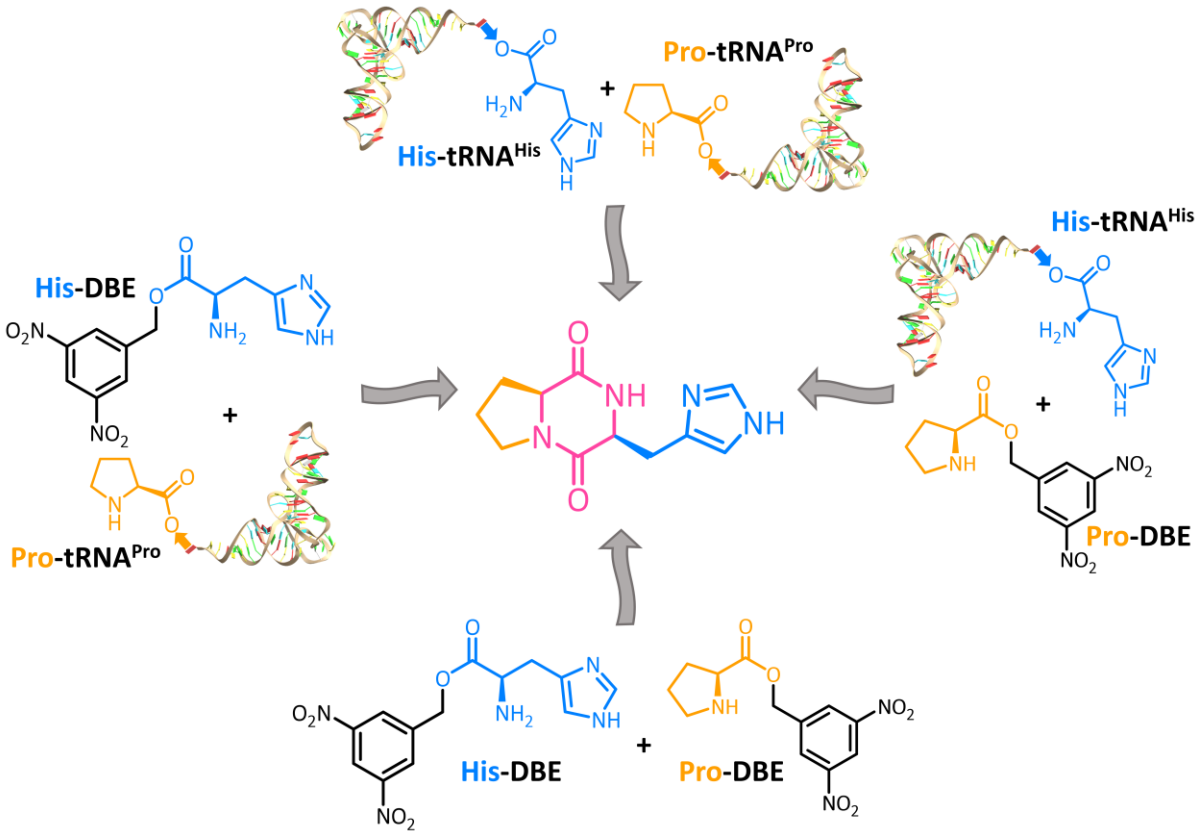


Figure 3.10 Combinatorial arrangement of substrates to yield cyclodipeptides.
Reaction scheme using cHP as an example to showcase the four combinations available when using aa-tRNA in combination with aa-DBE substrates.

LC-MS analysis of the reactions was performed as described in Section 2.5.18 and product quantification was accomplished using a calibration curve. It was observed that CDP formation was overall minimal when using two aa-DBE substrates by both CDPSs. However, *Para*CDPS and *Parcu*CDPS displayed differing trends when given a combination of the substrates. For *Para*CDPS, more product was formed when using Phe-DBE and His-tRNA^{His} whereas for *Parcu*CDPS, more product was observed

with His-DBE and Pro-tRNA^{Pro} (Figure 3.11). This result was initially puzzling, and more information was required to interpret the observation.

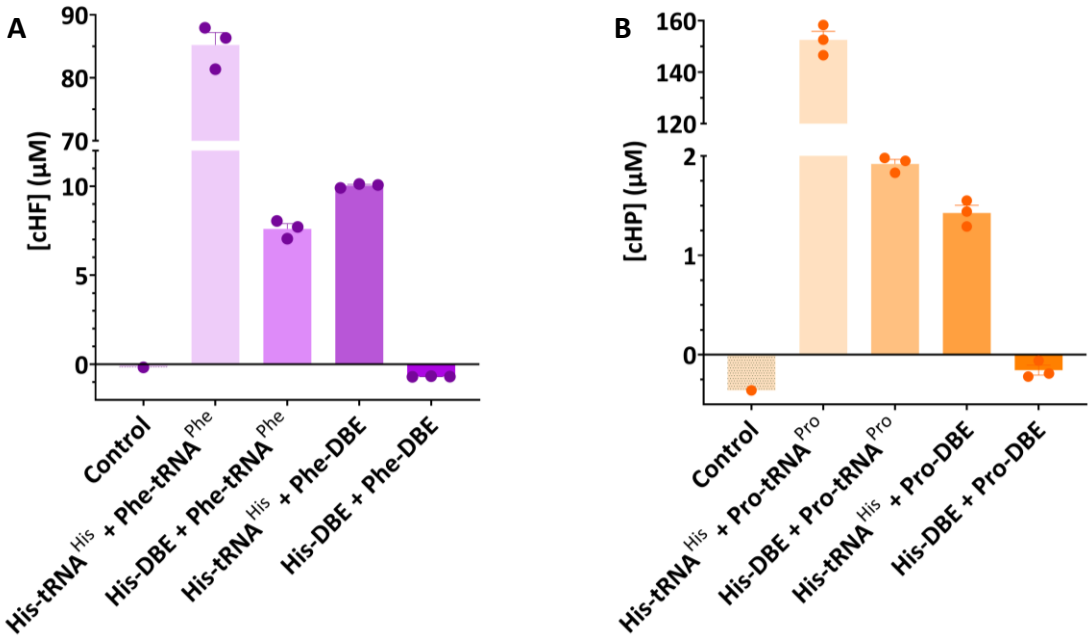


Figure 3.11 Quantification of CDP yield from the different substrate arrangements.
Reactions catalysed by *Para*CDPS are shown in purple (**A**) and reactions by *Parcu*CDPS are in orange (**B**). Experiments were performed in triplicate with each data point represented by a circle alongside the standard mean of error (SEM) as calculated by Prism (GraphPad).

Computational docking studies previously suggested that aa-DBE substrates could be productively bound into the active site binding pocket (P1) of the CDPS to generate the first aminoacyl-enzyme intermediate[67]. However, the second substrate binding was hypothesised to be poor due to the shallow nature of P2 allowing the substrate to sample multiple conformations, mostly of an unproductive nature. Additionally, the aa-DBE molecule was subject to hydrolysis in solution therefore it was more likely that the substrate would degrade before it encountered a productive binding conformation for enzymatic turnover. Thus, the overall product yield when using aa-DBEs was proven to be lower in comparison to the natural aa-tRNA substrates.

3.5 Kinetic analysis of product formation using minimal substrate unit

As shown by the reaction mechanism of CDPSs, there are three main chemical energy barriers to overcome for product formation: acylation, peptide bond formation and finally cyclisation. Acylation refers to the first half-reaction whereby the amino acid is attacked by the catalytic serine to generate the aminoacyl-enzyme intermediate. The energy barrier to cyclisation was expected to remain constant regardless of which combination of substrates were used and can only be crossed once the second acylation i.e. peptide bond formation has occurred and the enzyme promotes spontaneous cyclisation. To investigate the influence of these steps on product formation, a time-course

experiment was performed in duplicate using a combination of substrates with each CDPS. By using aa-DBE as a substrate, the energy barrier for either acylation or peptide bond formation was expected to increase, thus slowing the reaction rate which would be highlighted by following the production of CDP over time (Figure 3.12).

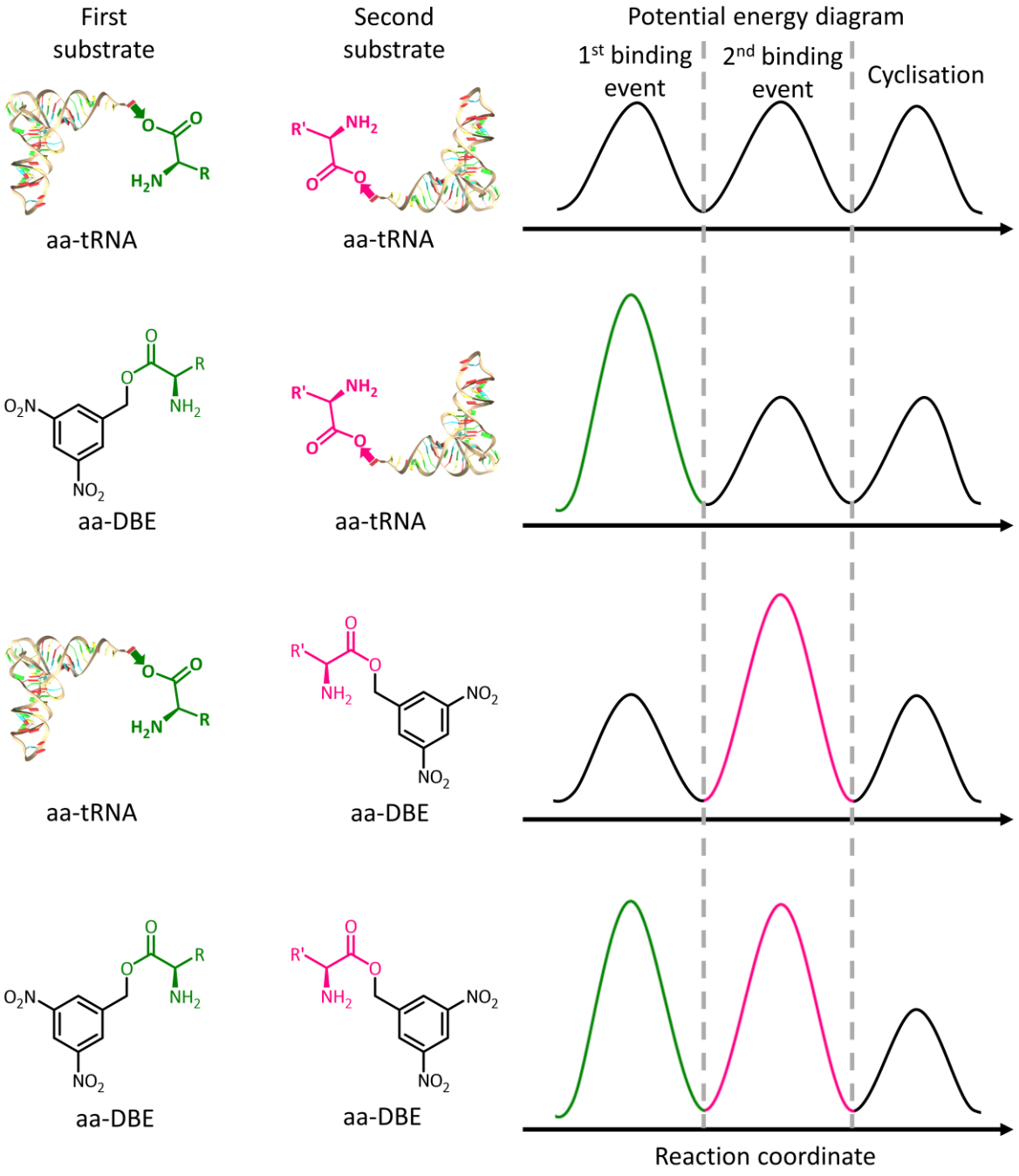


Figure 3.12 Energy barriers to product formation.
Schematic diagram highlighting the hypothetical change in energy when using different substrates. The first substrate is drawn in green and the second in pink. When aa-DBE is used, the energy barrier increases in comparison to aa-tRNA. The cyclisation step however remains constant throughout.

The rates of CDP formation were estimated by fitting the data to an exponential function in GraphPad Prism 9.4.1 (Figure 3.13). For *Para*CDPS, the rates of cHF formation using different substrates were as follows: His-tRNA^{His} + Phe-tRNA^{Phe} = 0.0024 min^{-1}; His-DBE + Phe-tRNA^{Phe} = 0.0022 min^{-1}; His-tRNA^{His} + Phe-DBE = 0.0048 min^{-1}, and His-DBE + Phe-DBE was too low to accurately fit. For *Parcu*CDPS, the rates of cHP formation were: His-tRNA^{His} + Pro-tRNA^{Pro} = 0.05 min^{-1}; His-DBE + Pro-tRNA^{Pro} = 0.00002 min^{-1} and His-tRNA^{His} + Pro-DBE and His-DBE + Pro-DBE were both too low to accurately fit. It should be noted that the rate when using two aa-tRNAs was not solely the rate of reaction as the tRNA has to first be aminoacylated by an aaRS before any step of the reaction can proceed. Therefore, this rate is not reflective of the cyclodipeptide synthase alone and any information inferred from this may not be accurate. Regardless, the aa-DBE substrates revealed a potential difference in mechanism between *Para*CDPS and *Parcu*CDPS. In the first hour of reaction, the rate of reaction for *Para*CDPS with Phe-DBE appeared to be double the rate when using two aa-tRNA substrates whereas the rates of reaction for *Para*CDPS were more comparable between the different combinations of substrates. This experiment was performed in duplicate and repeated three additional times however the trends reported remained consistent. Regardless, large standard errors of the mean are denoted by the error bars on each respective data point highlighting that this style of discontinuous assay is inherently flawed. A possible improvement to the experimental design could involve the generation of aminoacylated tRNA before the addition of CDPS to ensure the rate being measured relates only to the reaction catalysed by the CDPS. Overall, the time course highlighted that whilst each CDPS faces the same energy barriers, they differ in how rate limiting each step is. Thus, this experiment confirmed that further research concerning the binding order of these CDPSs was necessary to confidently assume the exact order in which the substrates are interacted with.

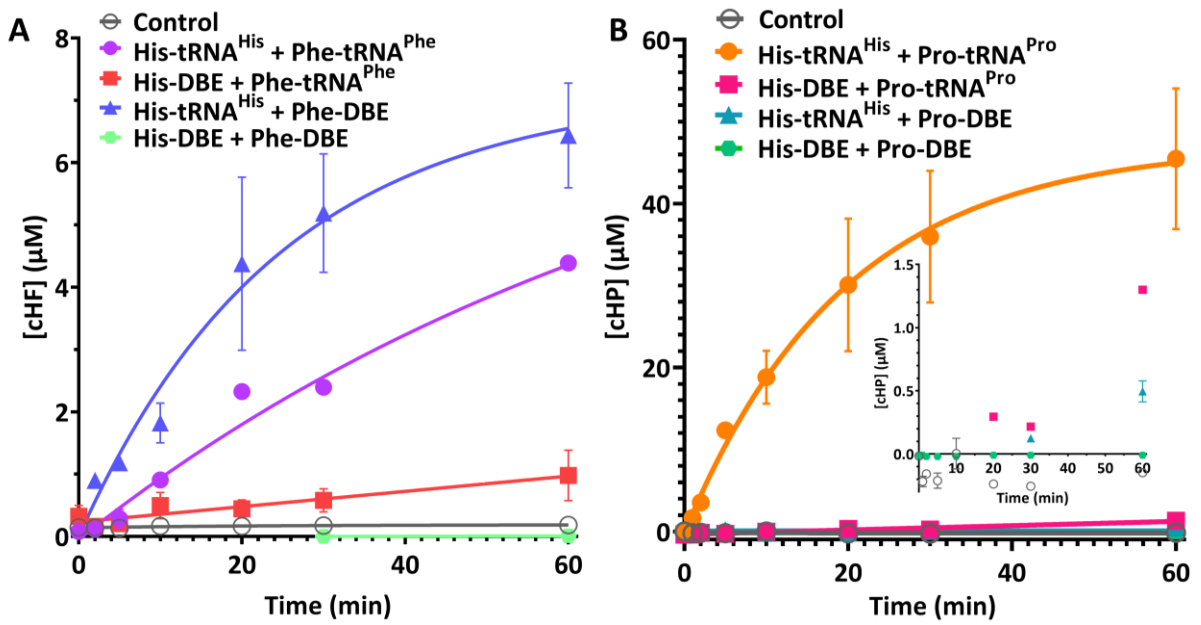


Figure 3.13 Cyclodipeptide formation monitored over time.
Time course data using *Para*CDPS (**A**) and *Parcu*CDPS (**B**) fitted to a single exponential equation. The inset on graph **B** highlights the poor levels of product formation during the first hour of reaction using aa-DBE substrates – two aa-tRNA substrates have been omitted here for clarity. Reactions were performed in duplicate, and the standard error of the mean (SEM) is shown from GraphPad Prism 9.4.1.

3.6 Investigation of substrate binding order

Minimal substrate units are doubly advantageous as alternatives to tRNA-based substrates and probes for substrate binding to P1. By removing the requirement to purify aminoacylated tRNA, a single aa-DBE could be used to bind to the active site of the CDPS thereby trapping the enzyme in the first acylation step if no second substrate was present. The two CDPSs in question use a mixture of two substrates, therefore it was hypothesised that only one would bind to P1 and this mass change would be observed by performing intact protein mass spectrometry.

Indeed, this hypothesis proved true revealing that *Para*CDPS accepted phenylalanine in P1 whilst *Parcu*CDPS accepted histidine first (Figure 3.14). The expected intact mass of wildtype *Para*CDPS, as shown previously, was 29816 Da and a mass change relating to the addition of phenylalanine (+147 Da) led to an observation of 29963 Da. No mass change relating to histidine (*Para*CDPS + His = 29953 Da) was detected for *Para*CDPS. On the other hand, a mass shift was only seen when incubating *Parcu*CDPS with histidine (*Parcu*CDPS + His = 28252 Da) and not with proline proving this CDPS must accept histidine first. P1 is known to be more restricted in size owing to its depth within the enzyme, therefore a strict mechanism is imposed on substrate acceptance in P1. Conversely, P2 is larger and wider on the surface of the enzyme with a less tightly regulated binding network. By accepting

histidine in different pockets, the recognition system of these CDPSs must differ. This, of course, could only be confirmed by structural investigations of each CDPS.

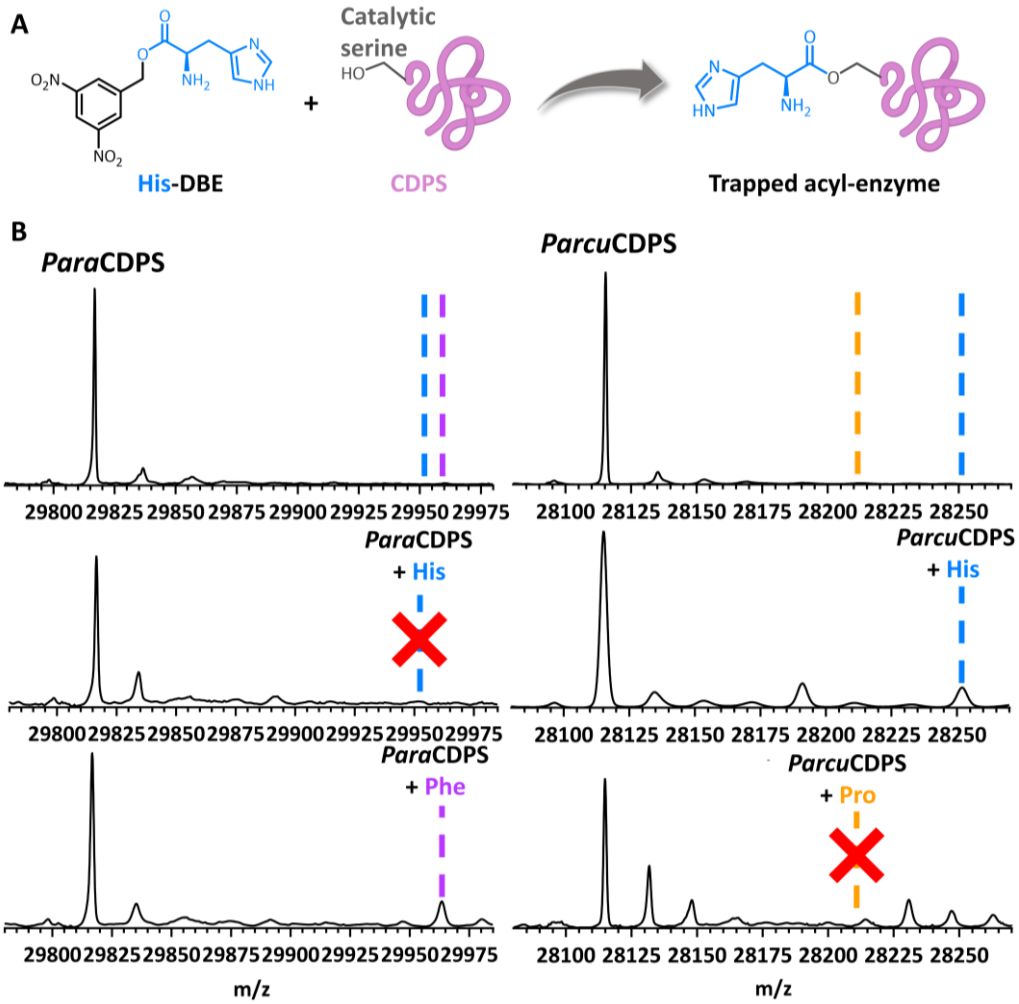


Figure 3.14 Investigation of CDPS substrate binding order.
A – Cartoon schematic detailing the formation of the trapped acyl-enzyme intermediate when given one aa-DBE substrate, His-DBE in this example. The protein is depicted in pink cartoon form with the catalytic serine drawn for clarification. **B** – Intact protein mass spectrometry of trapped acyl-enzyme intermediates. The graphs highlight that *Para*CDPS accepts phenylalanine first whilst *Parcu*CDPS accepts histidine. Each dotted line denotes the expected molecular weight of each enzyme-substrate complex: 29953 Da for *Para*CDPS + His; 29963 Da for ParaCDPS + Phe; 28212 Da for *Parcu*CDPS + Pro and 28525 Da for *Parcu*CDPS + His.

3.7 Concluding Remarks

Cyclodipeptide synthases use aminoacylated tRNA to yield cyclic dipeptides, however the nature of the accepted tRNA has been shown to be relaxed for CDPSs. Purified *in vitro* transcribed tRNA was the original substrate for these enzymes, but the results presented here highlighted CDPSs can use a mixture of tRNA from *E. coli* or even a cell extract containing tRNA and enzymes endogenous to *E. coli*. Moreover, the use of a mixture of tRNA avoids the need to individually purify and allows for ease of activity testing. The S30 extract would be a good initial experiment to investigate the substrate

selectivity of unknown CDPSs but high background signals mean that further experiments are required to ensure false positives are avoided. In general, more product was formed when using an excess of the tRNA pool which replaced the need for the laborious *in vitro* transcription, thus all forthcoming experiments used exclusively the tRNA pool.

Research by the Czekster lab highlighted that CDPSs could recognise activated amino acids with a dinitrobenzyl leaving group (aa-DBE)[67]. Both CDPSs investigated were able to use a mixture of aa-tRNA and aa-DBE as substrates to synthesise their respective product. It should be highlighted that this unique combination of aa-DBE and aa-tRNA as substrates for CDPSs has not been identified before. This is therefore a simplified method of producing cyclic products and has the potential to bypass the limitations posed by aminoacyl-tRNA synthetases which typically have restricted substrate tolerance due to their vital role in protein synthesis. Additionally, these minimal substrate units are key tools for investigating the binding order of CDPSs. Trapped acyl-enzyme experiments discovered that *Para*CDPS and *Parcu*CDPS accept histidine in different pockets which suggests each enzyme imposes its own distinctive recognition mechanism for histidine. This difference was further emphasised by the differing rates of reactions observed during a one hour time course monitoring product formation. This work involving both the newly synthesised tRNA pool and minimal DBE substrates could be of relevance to other enzymes which use aminoacylated tRNA as substrates such as fem aminoacyl transferases from *Staphylococcus aureus* where fem stands for factors essential for methicillin resistance[137]. This family of proteins catalyse the formation of peptide bonds using aminoacylated tRNA as substrates to synthesise the key component of bacterial cell walls, peptidoglycan[138]. The identity of the amino acid loaded onto the tRNA for this reaction must be correct otherwise the peptidoglycan biosynthetic pathway will be directly affected which could be fatal to the cell, however there is no current research on the use of different sources of tRNA with fem aminoacyl transferases[139].

Chapter 4 - Determining the substrate scope of *Para*CDPS and *Parcu*CDPS

The majority of the results presented here have been published as part of a manuscript: Sutherland, E., Harding, C.J., and Czekster, C.M., 2022, *Active site remodelling of a cyclodipeptide synthase redefines substrate scope*, Communications Chemistry, **5**(101).

4.1 Introduction

4.1.1 Natural promiscuity of aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases have been studied for approximately 60 years, elucidating many characteristics of these enzymes such as their mechanism of action and vital editing pathways to correct misaminoacylation. It is accepted that aminoacylation occurs via two steps - initially the synthetase facilitates the formation of the covalent aminoacyl-adenylate intermediate (aa-AMP) alongside the release of inorganic pyrophosphate (PPi). Esterification then occurs either at the 2'- or 3'-OH position in the ribose of the adenosine of the tRNA to transfer the amino acid from aa-AMP to give aa-tRNA and AMP finally (Figure 4.1)[140]. The way in which aaRSs bind the initial tRNA body depends on which class they belong to: Class I aaRSs bind ATP in an extended conformation whereas class II bind the substrate in a bent conformation[141]. This difference in ATP binding modes contributes to the different hydroxyl group which is esterified by the incoming aminoacyl-adenylate intermediate.

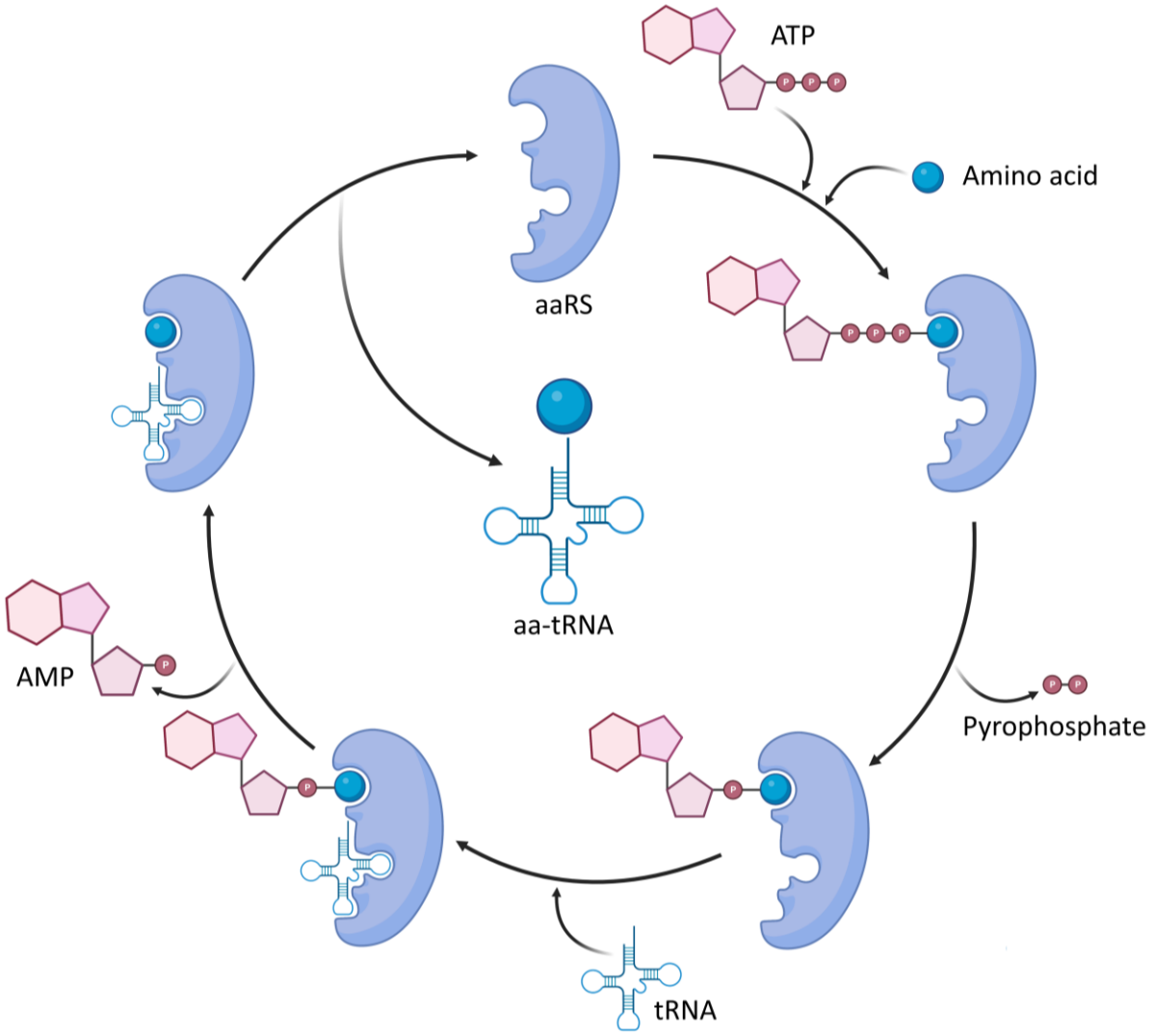


Figure 4.1 Aminoacylation of tRNA by aaRSs.
Reaction scheme depicting the use of aminoacyl-tRNA synthetase (aaRS) to generate aminoacylated tRNA via a two-step mechanism. Firstly, the aaRS catalyses formation of aminoacyl-adenylate followed by release of pyrophosphate. Secondly, a tRNA molecule binds to the synthetase and transesterification is facilitated by the synthetase to yield amino acid-tRNA (aa-tRNA) and AMP.

Numerous methods have since been developed to monitor the aminoacylation of tRNA which has been implicated in disease and human health. Initially, assays for this employed radioactive components such as labelled amino acids or tRNA. Labelled ^{14}C amino acids were first used to generate radioactive aminoacyl-tRNA that could be monitored via a scintillation counter, however this posed issues working at sub-K_M concentrations[142]. Alternatively, the tRNA body was altered by incorporating a ^{32}P label onto the 3'-adenosine end which was cleaved into mononucleotides whereupon the labelled amino acid-AMP could be separated from the native mononucleotides via thin layer chromatography (TLC)[143]. Radioactive assays were a valuable foundation to investigating aminoacylation, however the additional costs and procedures associated with these methods prompted the development of facile processes for general use.

The very nature of aminoacylated tRNA dictated that it was a difficult analyte composed of a labile ester bond and phosphate groups providing multiple charge states. Therefore, it became common practice to further derivatise this moiety for improved stability and binding. As previously described, Hartman *et al.* developed a method which both added a triphenylphosphonium derivatising group to aminoacylated-AMP and avoided any chromatographic difficulties by using MALDI for detection (Figure 4.2A)[144]. Near 200 commercially available non-canonical amino acids were screened via this procedure revealing that 92 were accepted by the aaRSs and subsequently charged onto the tRNA moiety. The synthetases only evolved to discriminate against other known natural amino acids found within a cell, thus the natural promiscuity of aaRSs was exploited for use with unnatural amino acids. Therefore, this new high throughput method using derivatised aa-AMP with MALDI analysis could pinpoint exact amino acid specificities for each aaRS under investigation. More recently, Gamper and Hou detailed the use of biotin and streptavidin to form a conjugate with tRNA which could be monitored by denaturing PAGE/7M urea gel[145]. This was also applied to non-proteinogenic amino acids including aa-DBE molecules using a ribonucleic acid (RNA) enzyme known as a flexizyme. This family of enzymes were first discovered whilst exploring the evolutionary path of aminoacyl tRNA synthetases and have since been shown to be effective enzymes for preparing aminoacylated tRNA as well as small activated aa-tRNA mimics for use in *in vitro* translation systems[146]. Overall, these two derivatisation methods eliminated the need for radioactive labels such as labelled pyrophosphate (^{32}PPi)[147] or tRNA labelled with ^{32}P at the 3'-terminal adenosine (Figure 4.2B)[143,148]. Radioactivity alone was not capable of directly confirming the incorporation of unnatural amino acids. Labelled pyrophosphate used excess ^{32}PPi to shift the reversible aminoacylation equilibrium thus reforming ATP with a ^{32}P label which was detected, whilst labelled tRNA could not distinguish the very nature of the amino acid, only the presence of aa-AMP. It should be highlighted that all previously mentioned methods to detect tRNA aminoacylation are indirect and at times convoluted multi-step processes. Clearly, it would be beneficial to develop of a facile method bypassing the need for additional derivatisation or radioactive labels.

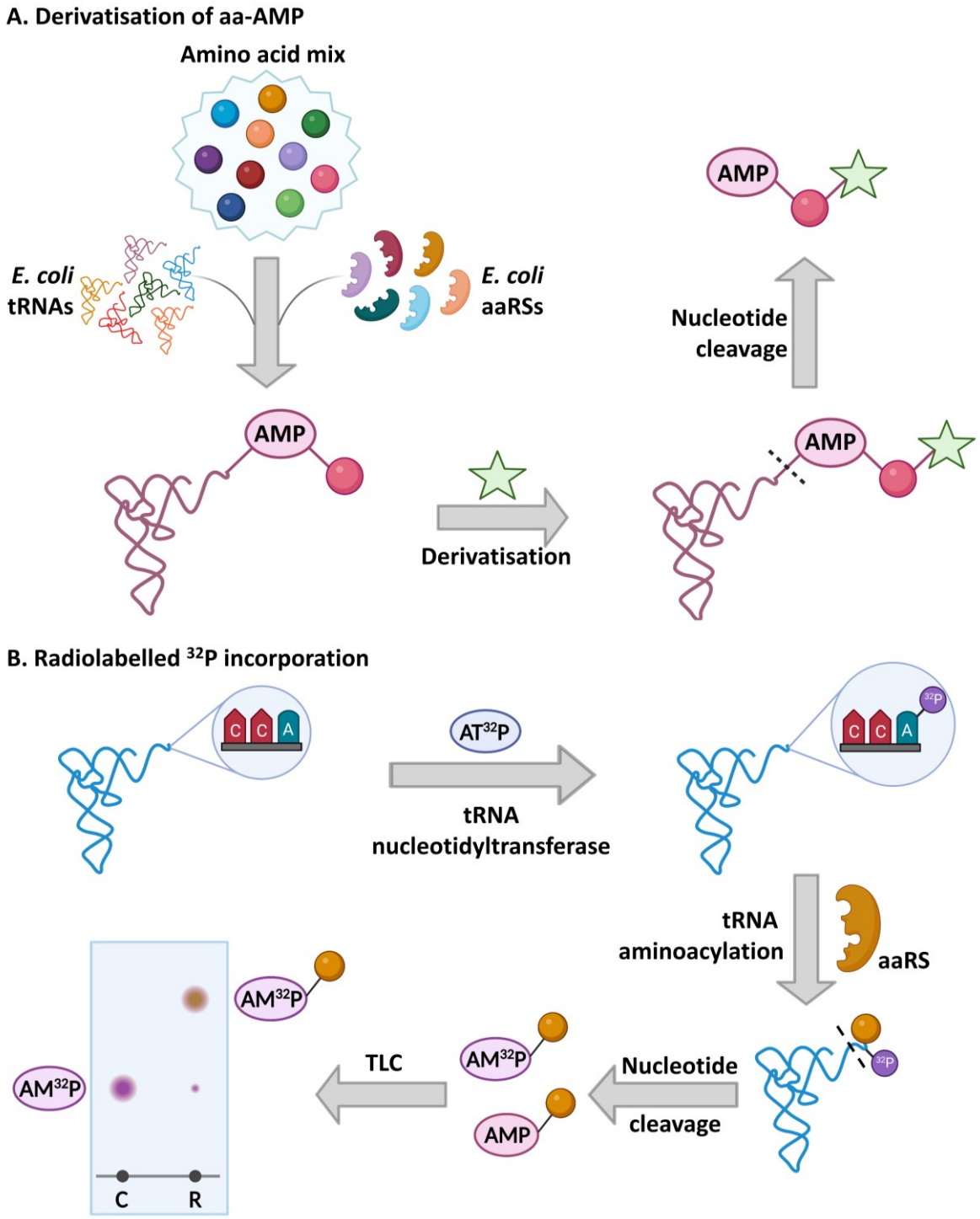


Figure 4.2 Methods to detect aminoacylated tRNA.
A - Schematic illustration showing the method developed by Hartman *et al.* to detect non-canonical amino acids conjugated to tRNA by aminoacyl-tRNA synthetases[144]. Amino acids are represented by coloured spheres and the triphenylphosphonium group as a green star. Cleavage into nucleotides (black dashed line) is catalysed by P1 nuclease. **B** – Schematic illustration depicting the use of ATP with a radiolabelled phosphate (^{32}P) which is transferred to the terminal adenosine of the tRNA before aminoacylation. Again, nucleotide cleavage is catalysed by P1 nuclease and the labelled aminoacylated-AMP is detected via thin layer chromatography (TLC)[149]. Graphic was created using BioRender.com.

4.1.2 Natural promiscuity of cyclodipeptide synthases

Originally, it was believed that the substrates of CDPSs would be a major limiting factor given that they are limited to the 20 proteinogenic amino acids from the cell. However, research introducing non-canonical amino acids into protein synthesis was also potentially applicable for CDPSs. The acceptance of non-canonical amino acids overcame one obstacle, but the second difficulty was postulated to originate from the CDPS natural selectivity. Yet, a pivotal study from 2018 detailed the generation of 198 CDPs containing non-canonical amino acids by using 10 different CDPSs[87]. This method employed auxotrophic strains of *E. coli* with plasmids containing previously characterised CDPSs. Initially, the cultures were supplemented with the necessary proteinogenic amino acid to complement the auxotrophy for growth. Following the induction of CDPS expression, the cells were grown in new media replaced with a non-canonical amino acid. Cyclodipeptides were detected from LC-MS analysis of the supernatant following cell lysis. In this particular publication, only certain analogues were tested including proline, phenylalanine, tyrosine, tryptophan and certain aliphatic derivatives. One shortcoming of this experiment was the fact that cells needed to first transport the unnatural amino acid inside the cell before it could be catalysed by the CDPS. Considering the nature of these amino acids, permeability into the cell could be less efficient than the natural counterparts, or simply not happen at all for some.

The following results presented in this chapter used the previously described tRNA pool with purified aaRSs to investigate the inherent promiscuity of *Para*CDPS and *Parcu*CDPS. We hypothesised that unnatural amino acids would be accepted substrates by these two CDPSs and any distinctive trends in the accepted substrates may highlight certain characteristics which might be essential for recognition. A method to directly detect aminoacylated tRNA was attempted, however aa-tRNAs were difficult analytes to observe via LC-MS with the methodology tested. Additionally, this method was further adapted to detect methylated tRNA which was included in a publication by Sweeney *et al*[115].

4.2 Generation of a cyclic dipeptide library using non-canonical amino acids

The promiscuity of *Para*CDPS and *Parcu*CDPS was tested by introducing commercially available non-canonical amino acids to the previously described *in vitro* experimental set-up using the tRNA pool as the source to generate aminoacylated tRNA with the respective aminoacyl tRNA synthetase (Section 2.5.15). It should be noted that the PheRS used here was the mutant PheRS-A294 which was previously shown to exhibit a wider binding pocket thus increasing its tolerance for unnatural substrates[150]. Reactions were analysed via LC-MS which allowed strict mass constraints to be imposed in order to confidently confirm the presence of a non-canonical product. Unique peaks – in comparison

to controls – were only considered if the extracted peak area was greater than 100 and the m/z mass deviation was within ± 5 ppm. Figure 4.3 shows all of the analogues tested with *Para*CDPS and Figure 4.4 for *Parcu*CDPS. LC-MS chromatograms for the accepted unnatural amino acids are displayed in Figures 4.5 and 4.6 whilst chromatograms for the rejected substrates can be found in Appendix II.

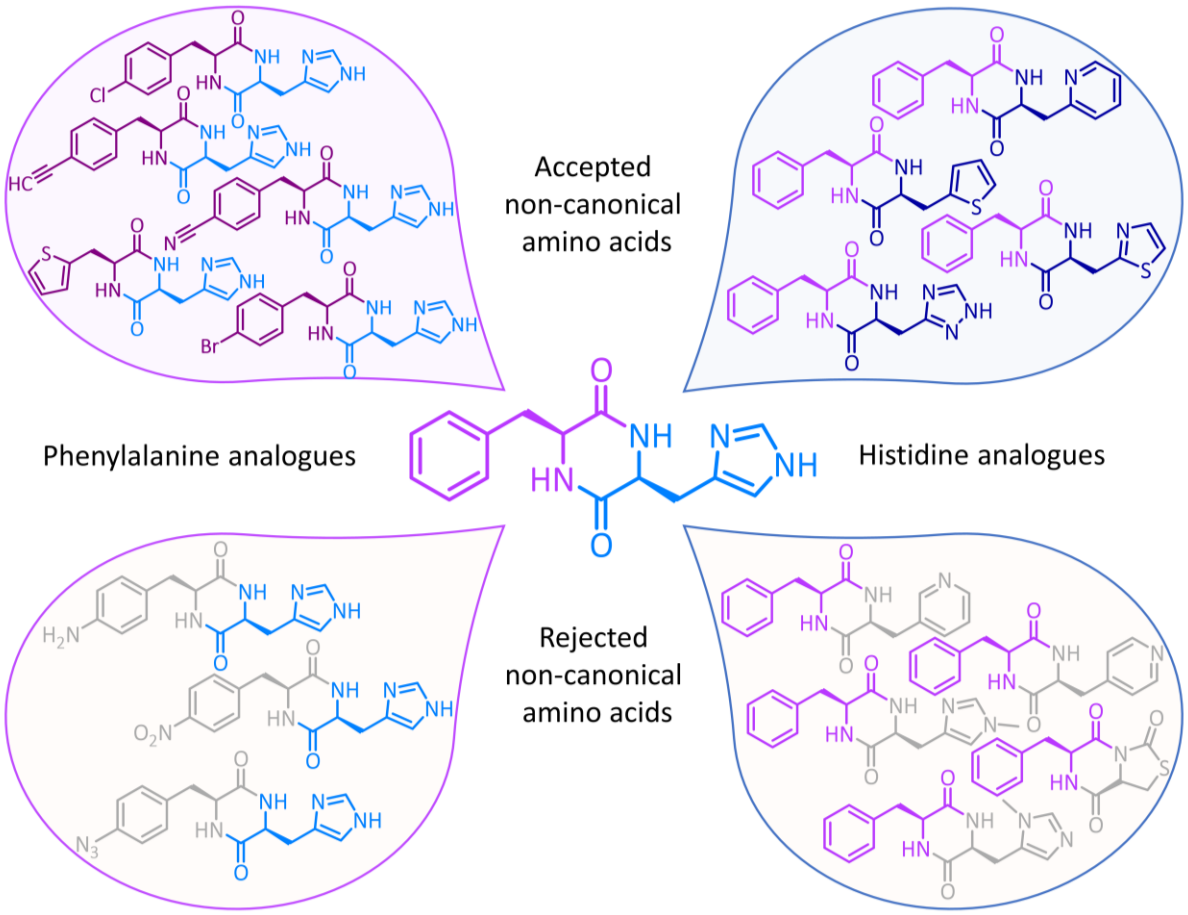


Figure 4.3 Catalogue of non-canonical amino acids tested with *Para*CDPS.
Using *Para*CDPS, 9 new cyclic dipeptide products were observed when replacing the proteinogenic substrate with a non-canonical amino acid (top two groups above cHF). Phenylalanine derivatives are shown in plum and histidine derivatives are in navy blue. Products which were not observed are drawn below cHF with the non-canonical amino acid shown in grey.

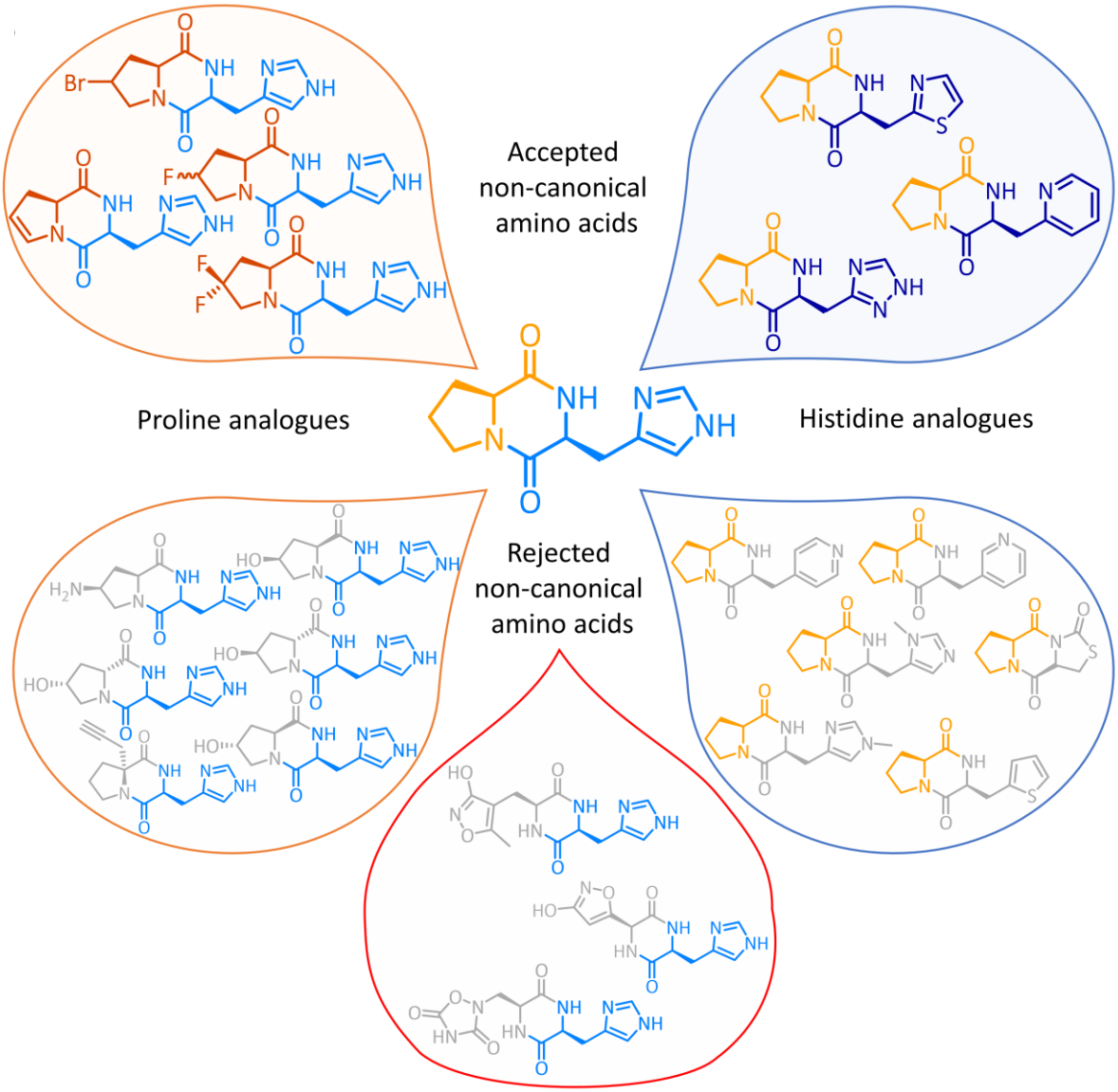


Figure 4.4 Catalogue of non-canonical amino acids tested with *Parcu*CDPS.
7 new cyclic dipeptides were generated using *Parcu*CDPS (top two groups above cHP). Proline derivatives are shown in dark orange and histidine derivatives are in navy blue. Products which were not observed are drawn below cHP with the non-canonical amino acid shown in grey.

The majority of unnatural amino acids tested were mentioned by Hartman in his extensive review of unnatural aaRSs substrates, with the exception of three: 4-bromo-L-proline; 4-amino-L-proline, and 4-amino-L-phenylalanine[86]. Results from mass spectrometric analysis highlighted that *Para*CDPS was particularly tolerant of various *para*-substituted halogenated phenylalanine molecules and larger derivations such as ethynyl and nitrile groups. Interestingly, *Para*CDPS was unable to turn over substrates where a nitrogen atom was directly bound to the benzyl ring, for example 4-nitro-L-phenylalanine or 4-azido-L-phenylalanine, both of which are known substrates for PheRS-A294G. This same hypothesis could extend to 4-amino-L-phenylalanine, however current literature contained no

mention of 4-amino-L-phenylalanine being accepted by PheRS, hence it remained unclear which enzyme was the deciding factor on substrate acceptance of this particular amino acid.

Focussing on histidine analogues, *Para*CDPS accepted modest changes to the imidazole ring including the introduction of an additional nitrogen or substitution by a sulfur atom. Moreover, 3,2-pyridyl-L-alanine was incorporated into the ring, however its isomers, 3,3- and 3,4-pyridyl-L-alanine were rejected. The same trend for histidine selection was observed when using *Parcu*CDPS except 3,2-thienyl-alanine was also rejected. These results indicated that the beta nitrogen of the imidazole ring may be necessary for recognition by a CDPS. For both enzymes, rejected histidine derivatives included methyl substituted histidine as well as L-oxothiazolidine-4-carboxylic acid. This indicated that both the position and hydrogen bonding capability of the imidazole nitrogens were vital for recognition as the presence of a methyl substituent may be interrupting the recognition by the CDPSs. Investigations of the substrate binding order (Section 3.6) highlighted that histidine was accepted in the first binding pocket (P1) of *Parcu*CDPS whilst it was accepted in the second pocket (P2) of *Para*CDPS. Thus, the observed similarities of the accepted unnatural amino acids cannot be solely attributed to the pocket interactions as they vary between the two CDPSs. Structural investigations of both CDPSs would aid our understanding of the otherwise imperceptible subtleties that allow these enzymes to accept unnatural amino acids as substrates.

Additionally, *Parcu*CDPS was shown to accept conservative changes to proline such as halogen atoms or a double bond within the proline ring. As stated previously, 4-bromo-L-proline was not confirmed as an accepted substrate for ProRS, however it was assumed that it would exhibit a similar chemical nature to 4-fluoro-L-proline which is a known substrate. Substituents such as hydroxyl or amine groups were rejected by *Parcu*CDPS despite being substrates for ProRS. It was hypothesised that these groups interrupted the hydrogen bonding network within the pocket by introducing a hydrogen bond donor group causing potential clashes of both a steric and/or electrostatic nature. *Parcu*CDPS was shown to also synthesise cyclo(L-His-L-Glu), however it was unable to incorporate any known glutamate derivatives into the CDP structure. Said glutamate derivatives differed greatly from the highly flexible glutamate as each were comprised of a five membered heterocycle. Thus, it can be inferred that these non-canonical substrates may be too bulky to enter the pocket of the CDPS for the reaction.

Overall, 16 new cyclic dipeptides containing unnatural amino acids were produced from two histidine accepting CDPSs. With the exception of cyclo(L-His-4-cyano-L-Phe)[151] and cyclo(L-His-3-(2-thienyl)-L-Ala)[152], there were no published synthetic routes to these products according to searches performed on Reaxys. Therefore, a small library of novel cyclic molecules was successfully generated

from *Para*CDPS and *Parcu*CDPS highlighting the advantages of exploiting the natural promiscuity of these enzymes.

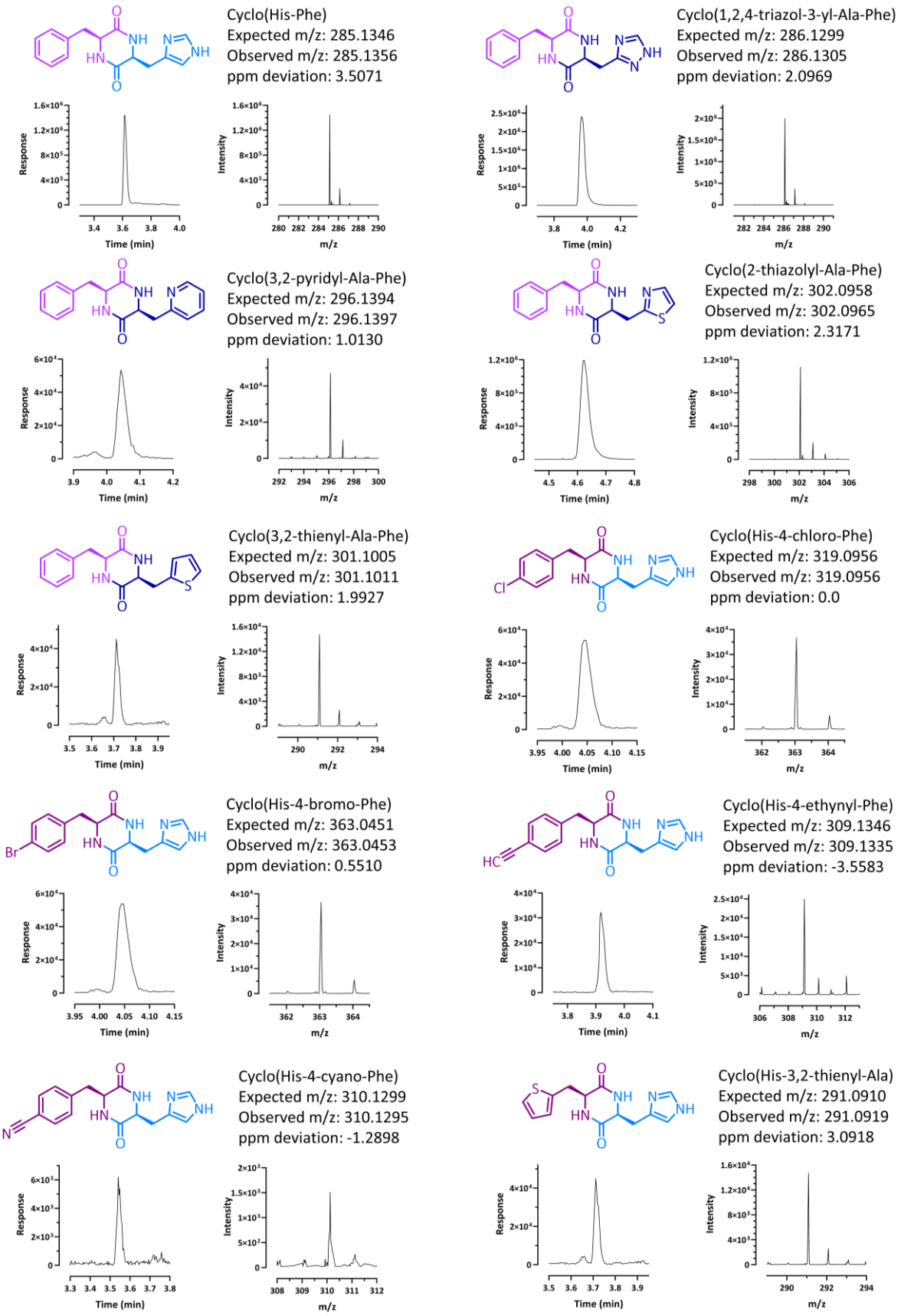


Figure 4.5 LC-MS analysis of accepted unnatural amino acids by *Para*CDPS.
The structure, extracted ion chromatogram and the ion count are given for all observed products from *Para*CDPS. Histidine analogues are drawn in dark blue and proline analogues in plum.

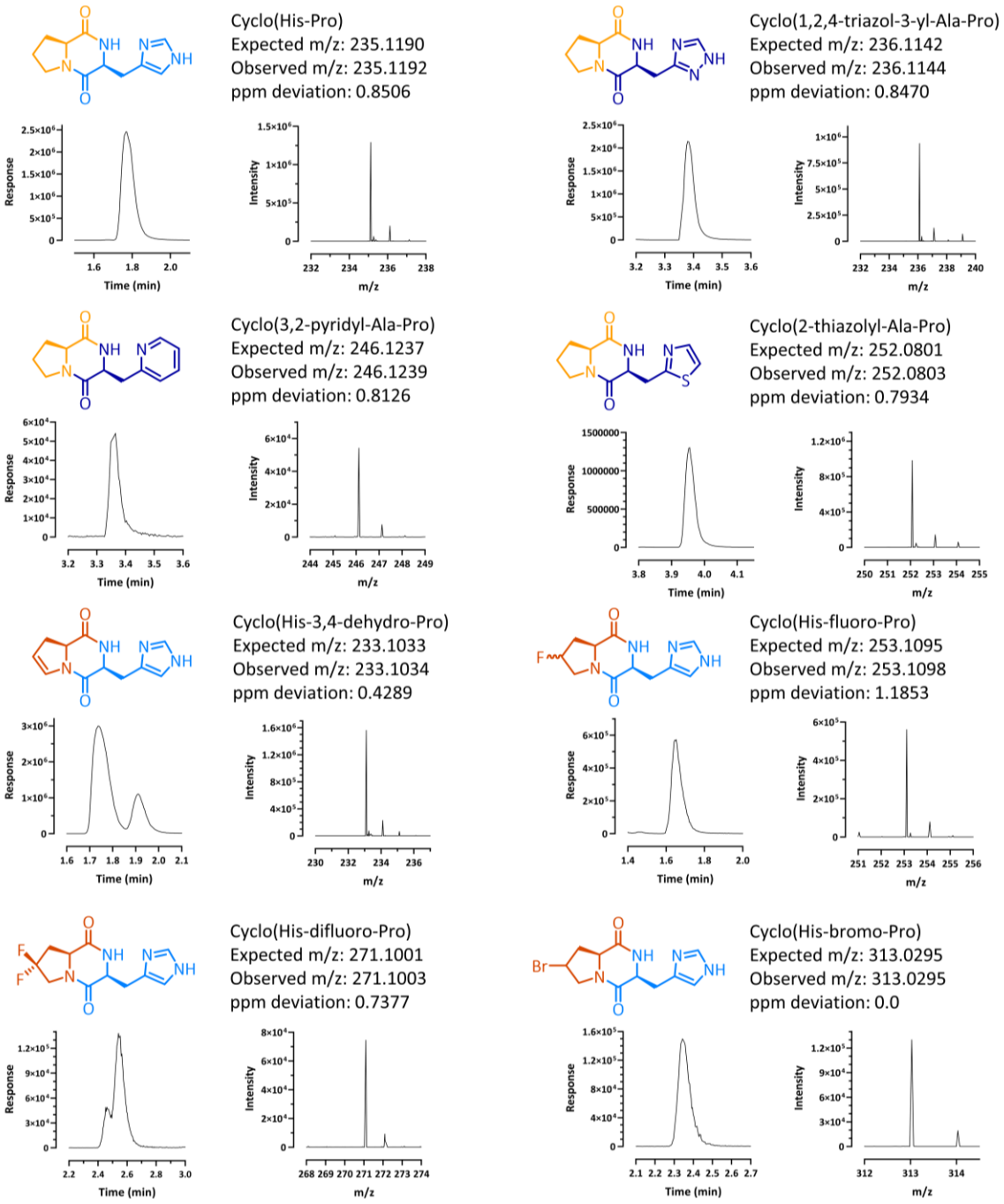


Figure 4.6 LC-MS analysis of unnatural amino acids accepted by *Parcu*CDPS.
The structure, extracted ion chromatogram and the ion count are given for all observed products from *Parcu*CDPS. Histidine analogues are drawn in dark blue and proline analogues in dark orange.

4.3 Development of a method to detect aminoacylated-tRNA

Using aminoacyl-tRNA synthetases with their natural amino acid substrates, the detection of aa-tRNA was attempted by establishing a direct LC-MS method. Two commercially available enzymes – nuclease P1 and Antarctic (acid) phosphatase - were utilised to first cleave the phosphodiester bond between the RNA nucleotides and then dephosphorylate individual nucleotides forming a mixture of nucleosides. It was hypothesised that one of the adenosines would be bound to the desired amino

acid if aminoacylation occurred which could be detected by the mass change relating to the amino acid (Figure 4.7). Similar methods had been published previously which also used nuclease P1 to cleave tRNA into its respective mononucleotides for analysis thus demonstrating this method was potentially applicable for the desired purposes here[153].

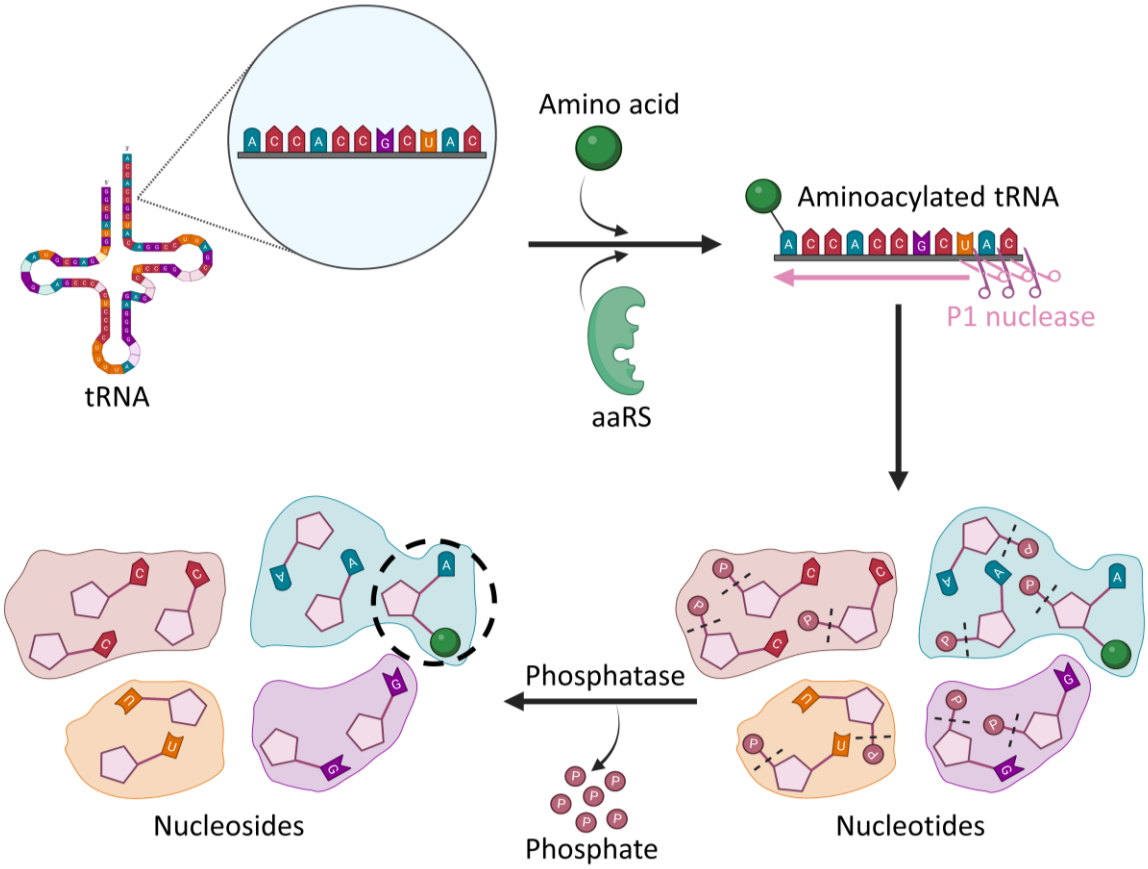


Figure 4.7 Derivatisation of aminoacylated tRNA for direct LC-MS detection.
Cartoon schematic describing the steps taken to prepare aminoacylated tRNA for analysis via mass spectrometry. The nucleobases are denoted by the first letter of their name: A – adenine (green); C – cytosine (red); G – guanine (purple) and U – uracil (orange). The tRNA was first aminoacylated by the desired amino acid followed by P1 nuclease-catalysed hydrolysis of the phosphodiester RNA backbone generating a mixture of nucleotides. These nucleotides were subsequently broken down to their respective nucleosides whereupon LC-MS analysis was performed, specifically looking for adenosine bound to an amino acid.

Preliminary trials used natural amino acid substrates with their respective aaRS to confirm proof of principle – for example histidine with HisRS, proline with ProRS and phenylalanine with PheRS (Figure 4.8). Strict mass allowances (± 5 ppm) were imposed on this analysis due to the nature of the metabolite and the high resolution possible from the instrument used (0.01 Da extraction mass window). For proline, it appeared that the adenosine species could be present indicated by the small ppm deviation of 3.5601 from the expected m/z values, however the signal to noise ratio majorly affected the chromatogram (Figure 4.8B). For histidine and phenylalanine however, it was impossible

to conclude if the amino acid-adenosine product was present due to poor signal quality and high ppm deviations beyond the accepted mass tolerances (Figures 4.8A and C).

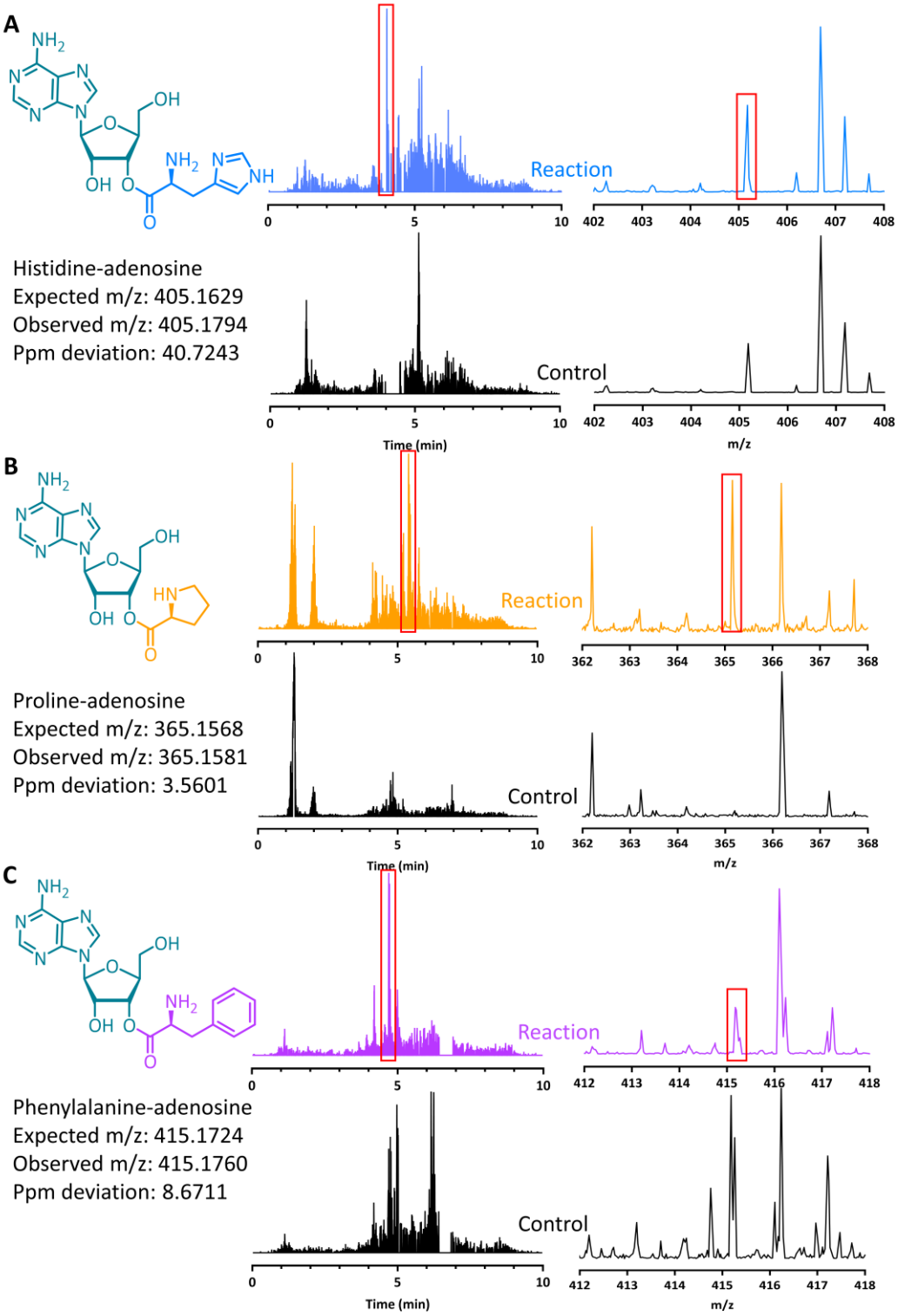


Figure 4.8 LC-MS analysis of aminoacylation using proteinogenic amino acids.
Following the two step manipulation of aminoacylated tRNA, LC-MS analysis was conducted to observed amino acid coupled to adenosine (aa-AMP). All controls are shown in black and each reaction is drawn in the corresponding colour of the amino acid: blue for histidine (A), orange for proline (B) and purple for phenylalanine (C). The mass chromatogram was extracted from the peak within the red box of the extracted ion chromatogram (EIC) and the exact masses are detailed to the left of graphs.

This assay was intended as a means to investigate the substrate specificity of the aaRSs when incorporating unnatural amino acids into the reaction. Therefore, some known accepted and unaccepted amino acids were selected for analysis using this method (Figure 4.9).

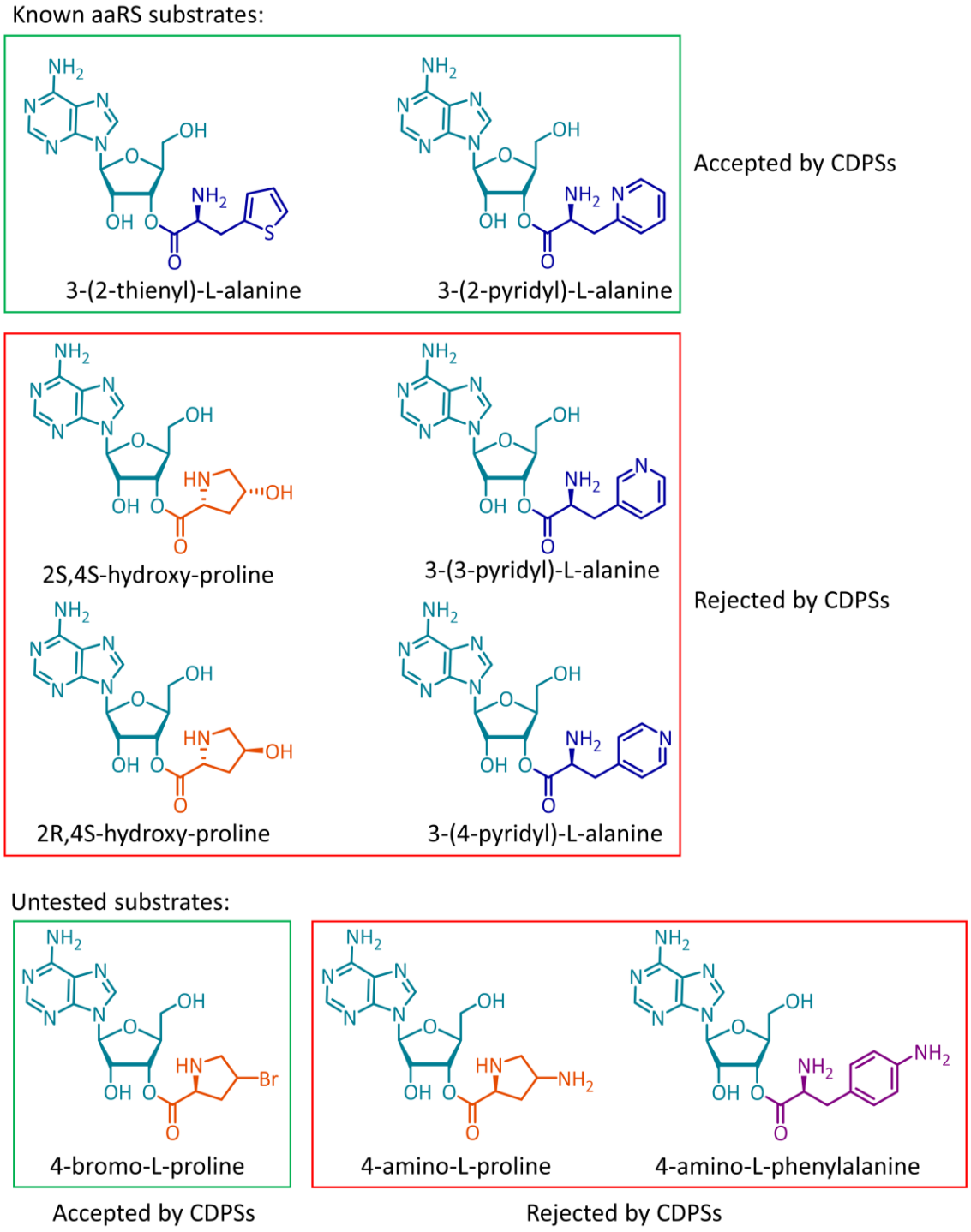
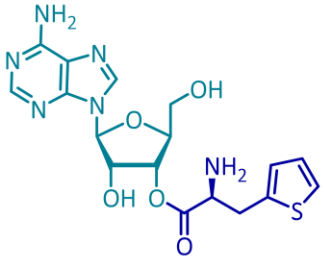
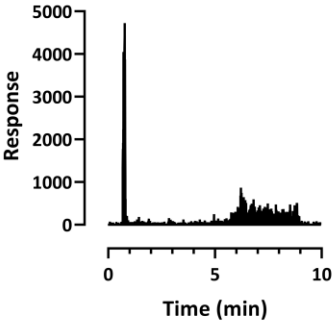


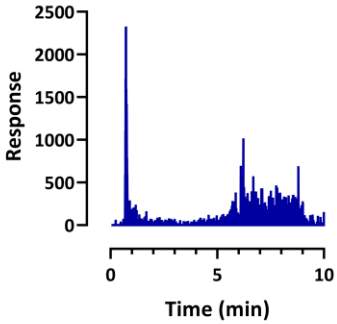
Figure 4.9 Non-proteinogenic amino acids tested for tRNA aminoacylation.
Structures of the expected nucleoside products when using unnatural amino acids. Known substrates of aaRSs were chosen to reflect both accepted and rejected amino acids by the CDPSs. The only three unconfirmed aaRS substrates were also included: two of which were also rejected by CDPSs for incorporation into the ring.

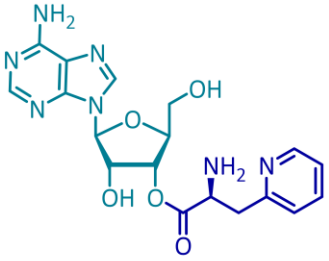
The three amino acids which were not established substrates for aaRSs were included - 4-bromo-L-proline, 4-amino-L-proline and 4-amino-L-phenylalanine - as well other substrates confirmed by Hartman such as 3,2-pyridyl-L-alanine and both diastereomers of 4-hydroxy proline. Unfortunately, the lack of sensitivity and increased background noise posed a significant issue for detecting the low concentrations, if any at all, of the elusive and unstable amino acid-adenosine product. Consequently, no products were confidently detected when using non-proteinogenic amino acids as substrates for aminoacyl-tRNA synthetases (Figure 4.10). Thus, it can be concluded that this particular method to detect aminoacylation was not successful and further optimisation was necessary by perhaps changing the enzymes responsible for cleavage to maximise the efficiency of resultant nucleotide production.



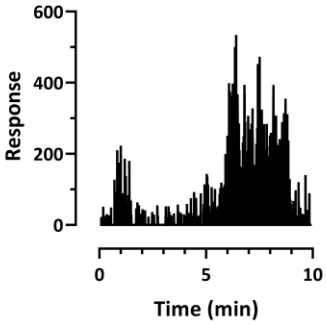
3-(2-thienyl)-L-alanine
Expected m/z: 421.1289

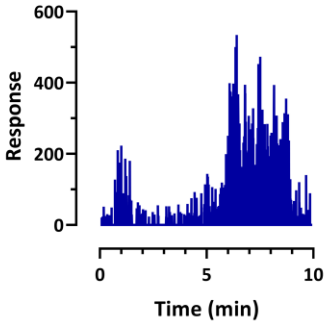


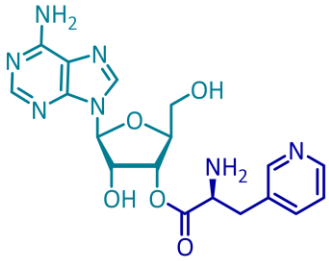




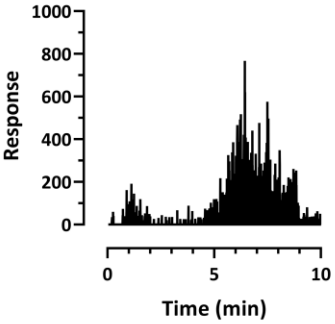
3-(2-pyridyl)-L-alanine
Expected m/z: 416.1677

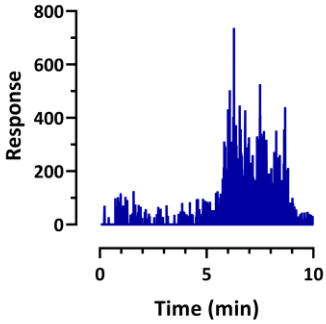


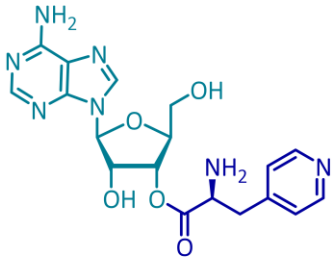




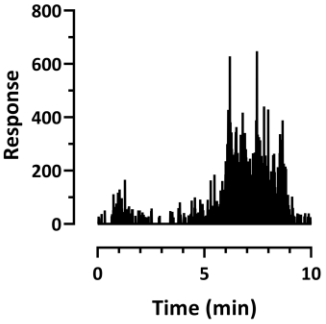
3-(3-pyridyl)-L-alanine
Expected m/z: 416.1677

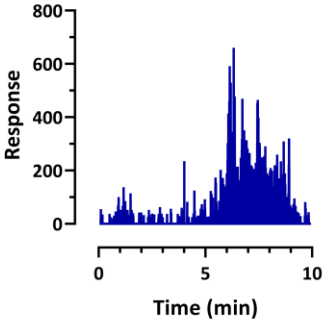


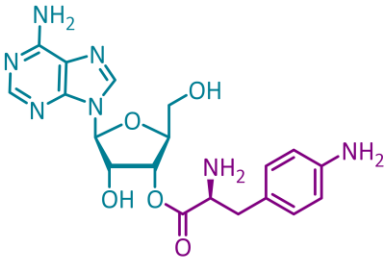




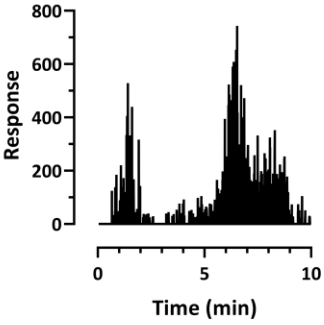
3-(4-pyridyl)-L-alanine
Expected m/z: 416.1677

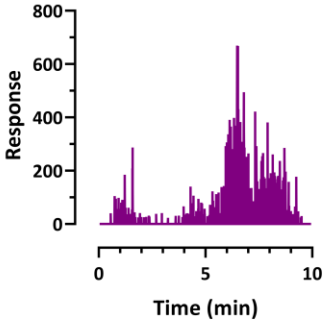






4-amino-L-phenylalanine
Expected m/z: 430.1833





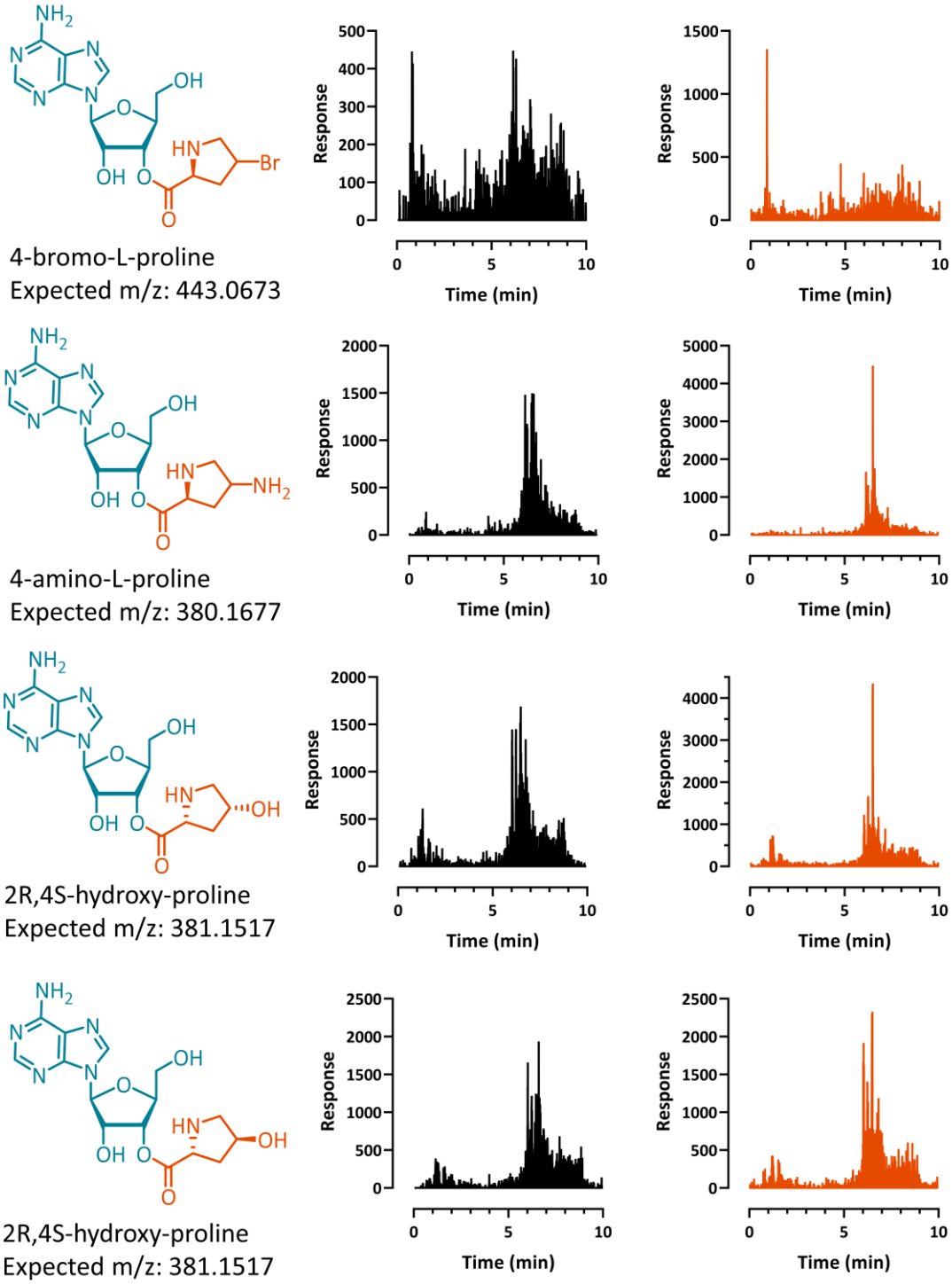


Figure 4.10 LC-MS analysis of aminoacylation assay using unnatural amino acids as substrate.
The expected aa-adenosine product is shown on the left with the m/z value underneath. Extracted ion chromatograms (within a 0.01 Da mass window) are on the right - controls are drawn in black and reactions are in the same colour as the amino acid derivative used: histidine in navy blue, phenylalanine in plum and proline in burnt orange.

4.4 Confirmation of methylated tRNA using a methyltransferase

When aminoacylated tRNA is not being diverted by cyclodipeptide synthases, it typically takes part in protein translation, acting as a substrate for the ribosome. The structure of tRNA is well established as a cloverleaf secondary structure which forms an L-shaped tertiary structure via interactions between the D (dihydrouridine) and T (TΨC) loops[154]. Modifications to tRNA are relatively common with current literature detailing 93 possible post-transcriptional modifications that can occur, but often their overall effect is not fully understood[155]. The most frequently found modifications are methylations to the base and/or ribose groups catalysed by methyltransferases, most of which use S-adenosyl-L-methionine (SAM) as a methyl group donor[156]. This modified nucleoside, known as 1-methyladenosine (m^1A), can be found at 4 different positions in tRNA: A9, A14, A22 and A58[157]. More specifically, the m^1A22 modification has only been observed in $tRNA^{Ser}$ and $tRNA^{Tyr}$ from *Bacillus subtilis*[158] and *Mycoplasma capricolum*[159] which both feature a long extra arm. In 2008, a methyltransferase was isolated from *Bacillus subtilis* which displayed selective methylation of the A22 position of tRNA, termed hereafter as TrmK (Figure 4.11)[156]. This enzyme, TrmK, has been observed in other Gram-positive bacteria, and shown to be vital for the survival and growth of *Staphylococcus aureus*[160] and *Streptococcus pneumoniae*[161]. It has been hypothesised that TrmK could be a likely candidate for antibiotic development against *S. aureus,* considering that there is no homolog for TrmK found in mammals[156].

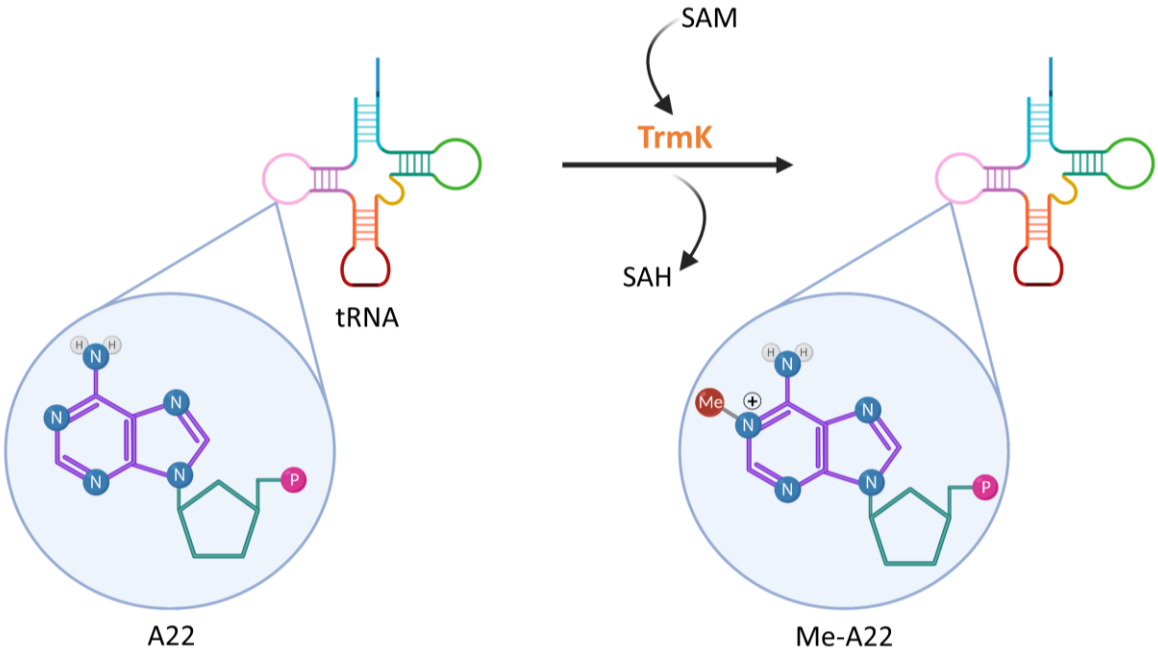


Figure 4.11 tRNA methylation catalysed by TrmK.
Reaction scheme highlighting the use of TrmK to methylate adenine at position 22 of tRNA using SAM as a methyl group donor. The ribose ring is denoted in green, the adenine base in purple and the phosphate group is shown as a pink circle. The addition of the methyl group to N^1 of the adenine ring is shown as a red circle.

Current literature has since focussed on developing a covalent inhibitor for TrmK of *Staphylococcus aureus* origin[115]. One proven method to investigate enzyme catalysed methylation uses a MTase-Glo reagent to convert the SAH by-product to ADP which only reacts with luciferase once converted to ATP[162]. This assay is known to be reliable and highly sensitive to SAH production for methyltransferase characterisation, however further validation of this modified tRNA was sought in the form of direct detection using LC-MS techniques. This research was completed in collaboration with Dr Pamela Sweeney and Dr Rafael Guimaraes da Silva (University of St Andrews)[115].

In brief, a method was developed in-house using $tRNA^{Leu}$ - known to contain an adenine at position 22 - with TrmK to yield a methylated form which was hydrolysed to its corresponding nucleotides using nuclease P1 to cleave the phosphodiester backbone of tRNA. N^1-methylated-AMP was detected using LC-MS to first separate the analyte from the mixture and then observe the expected mass with a Q-ToF mass spectrometer. A commercial standard of N^1-methylated AMP confirmed the expected retention time of the desired product which also allowed comparison to the reaction mixture where the peak was not as abundant (Figure 4.12A). By using a mutant tRNA version which replaced adenine 22 with a cytidine, LC-MS analysis proved that no methylation had occurred. Therefore, this method strengthened the argument that the position of the adenine is essential for methylation by TrmK. Furthermore, a truncated moiety resembling only the D arm of the tRNA was tested as a substrate for TrmK. Again, this was successfully detected using the newly developed assay, albeit with higher concentrations of substrate and enzyme necessary to give a confident signal (Figure 4.12B).

Therefore, it was demonstrated that digesting tRNA into nucleotides was a successful derivatisation method for direct observation of methylated tRNA by TrmK. This assay proved itself to be reliable and amenable to different forms of tRNA such as truncated versions, as long they remained substrates for the enzyme. This procedure was relatively simple and could be applied to investigate different modifications to tRNA, granted that the newly formed covalent bonds are stable at the acidic pH required by the nuclease.

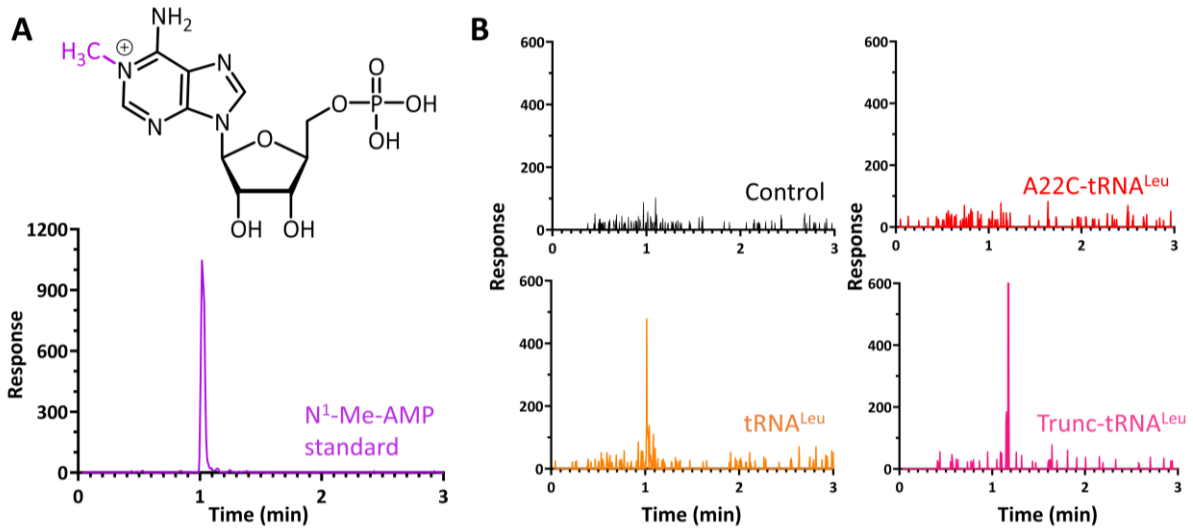


Figure 4.12 Detection of methylated-AMP via LC-MS.
A – The structure of N^1-methylated AMP and below, the extracted ion chromatogram observed from the commercial standard (purple). **B** – EICs of control (black), A22C-$tRNA^{Leu}$ (red), $tRNA^{Leu}$ (orange) and Trunc-$tRNA^{Leu}$ (pink). From this analysis, only a peak relating to the expected mass for N^1-Me-AMP was observed with tRNA featuring an adenine nucleobase at position 22.

4.5 Conclusion

Cyclodipeptide synthases are known as promiscuous enzymes, typically capable of yielding more than one product. This fact, however, was based solely on the 20 proteinogenic amino acids readily available within a cell. Herein, both histidine accepting CDPSs were exposed to unnatural amino acids for incorporation into the CDP ring using an *in vitro* experimental setup. The success of this experiment highlighted the unexploited diversity of accessible products using cyclodipeptide synthases. From *Para*CDPS and *Parcu*CDPS, 16 new cyclodipeptides were created - the majority of which had no previous known synthetic route. In future, these could be purified via high performance liquid chromatography (HPLC), especially since this enzymatic reaction has no difficult side-products to separate from the mixture. The isolation of these new molecules could lead to the discovery of new bioactive molecules which could be tested against a variety of disease. It is evident that the combinations of products available by using the 120 previously characterised CDPSs are endless and should be investigated to fully utilise the promiscuity of these enzymes.

Considering the use of unconventional amino acids with aminoacyl-tRNA synthetases, it would be beneficial to have a direct method of detection for confirmation of acceptance by these enzymes. Previous research had devised various methods, all of which required forms of derivatisation and some were unable to explicitly confirm the conjugation of an unnatural amino acid to tRNA. Therefore, a facile method was sought using LC-MS analysis of hydrolysed tRNA with the aim of observing the amino acid coupled to the terminal adenosine. Regrettably, this approach did not yield successful

results with the majority of expected products undetected. However, hydrolysis of tRNA was shown to be effective when detecting an alternative metabolite - methylated-AMP - therefore it can be inferred that amino acid-adenosine intermediates pose additional challenges by the presence of the labile ester bond and the overall nature of the amino acid substrates used here.

Chapter 5 - Structural characterisation of wild type *Parcu*CDPS and tailored engineering of key residues

The majority of the following results from this chapter have been published as part of a manuscript: Sutherland, E., Harding, C.J., and Czekster, C.M., 2022, *Active site remodelling of a cyclodipeptide synthase redefines substrate scope*, Communications Chemistry, **5**(101).

5.1 Introduction

Initial structural investigations of CDPSs focussed on some of the first known members of the family such as *Snou*-CDPS (PDB: 3OQV)[55], Rv2275 (PDB: 2X9Q)[54], and YvmC-Blic (PDB: 3OQH)[57]. Secondary structure comparisons highlighted that these CDPSs could be likened to class-Ic aa-tRNA synthetases, more specifically tyrosyl and tryptophanyl-tRNA synthetases[54]. It is highly probable that CDPSs and aaRSs evolved from a similar ancestor, however CDPSs have since developed unique features which distinguish them from aa-tRNA synthetases. For example, each CDPS possesses two solvent-accessible pockets for substrate binding, known hereafter as P1 and P2. The P1 pocket is located deep within the core of the CDPS enzyme and is typically narrower as a consequence – this pocket can be compared to a feature found in both TyrRS and TrpRS, used for binding the aminoacyl moiety[68]. However, aaRSs bear no similar feature to the CDPS P2 which is situated closer to the surface and is variable in size and shape thus giving these enzymes diverse substrate promiscuity.

Using the structure of *Snou*-CDPS complexed to a reaction intermediate mimic, Moutiez *et al*. postulated a reaction mechanism for CDPSs which included the use of 4 residues: S37; Y178; E182 and Y202 (numbering respective to *Snou*-CDPS)[56]. Subsequent research revealed that these residues were consistent throughout members of the family, regardless of origin. Specifically, S37 and Y202 were strictly conserved whilst N40, Y178, E182 and H203 were mostly conserved[64]. Around half of the CDPSs considered at the time contained an ‘N40 and H203’ pair whilst an ‘X40 and P203’ pair was also frequently observed (X being a non-conserved residue)[64]. Therefore, two distinct sub-groups were created, known as ‘NYH’ and ‘XYP’ according to the N40, Y202, H203 or X40, Y202, P203 sequences (numbering respective to *Snou*-CDPS). Interestingly, the first three published CDPS structures were inadvertently all members of the NYH family which gave some indication to the higher population of NYH CDPSs found in nature.

Considering most research centred on CDPSs from the NYH sub-family, Bourgeois and colleagues aimed to bring more clarity to the XYP classification by solving the structures of three CDPSs from this sub-group for comparison[65]. The subsequent research highlighted that whilst the overall topology was similar to CDPSs of the NYH class, structural divergence between the two groups was most prominent

in the first half of the Rossman fold, a known characteristic feature of all CDPSs. Significant differences were observed more specifically in the positioning of sheets β2 and β7, as well as the helices α1 which was typically well conserved in the NYH sub-class and α8 which appeared longer in XYP enzymes (Figure 5.1)[65]. This evidence corroborated a previous publication investigating the phylogenetic origins of the two CDPS branches suggesting that this structural divergence occurred early in the evolution of the cyclodipeptide synthase family[64].

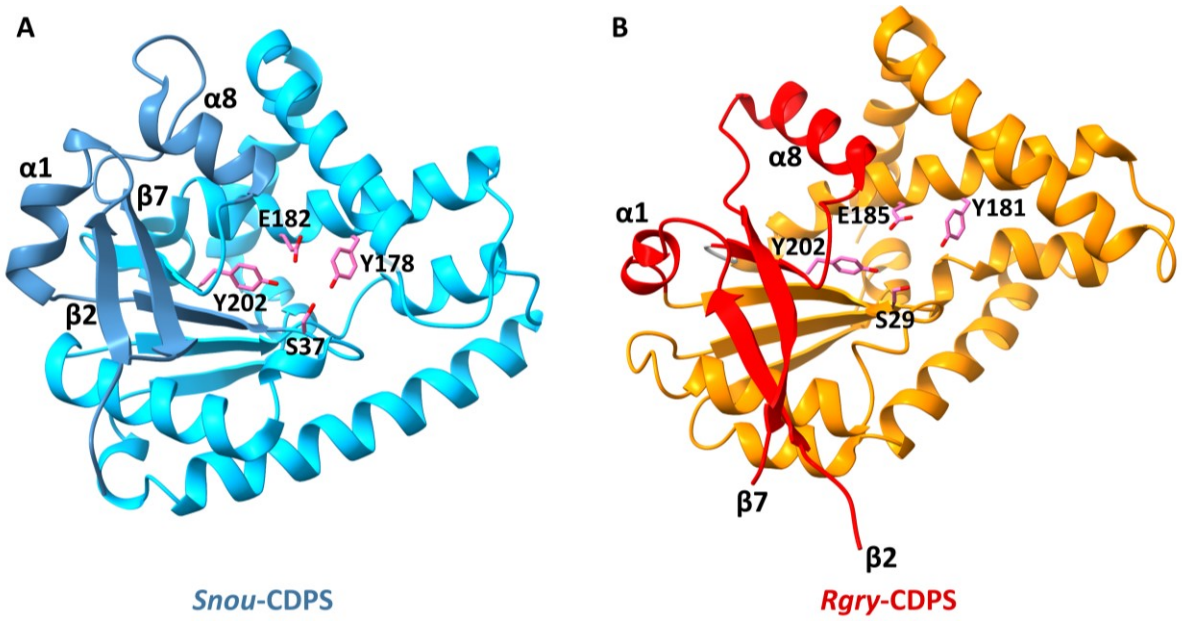


Figure 5.1 Structural comparison of *Snou*-CDPS with *Rgry*-CDPS.
A - *Snou*-CDPS is drawn in light blue with the first half of the Rossman fold in dark blue (PDB: 3OQV). **B** - *Rgry*-CDPS is in orange with the first half of the Rossman fold in red (PDB: 5MLP). The 4 catalytic residues of each protein are depicted in pink and labelled as well as the secondary structure elements of the first half of the Rossman fold for clarity.

Irrespective of the sub-classes, all CDPSs were still assumed to participate in the same catalytic mechanism to yield cyclic products (Reaction shown in Figure 1.4). Sequence alignments highlighted two strictly conserved catalytic residues, S37 and Y202 throughout the CDPS family. Additionally, the third catalytically relevant residue, E182, was present with only two exceptions possessing a glutamine replacement and lastly, Y178 was found in 96.7% of analysed sequences[65]. The importance of these residues was demonstrated by reduced product yield when using mutant CDPS variants from either NYH or XYP groups lacking one or multiple catalytic residues. Moreover, structural investigations revealed that position 40 was involved in stabilising the catalytic serine by main or side chain interactions regardless of whether the residue was a conserved N or variable X[65]. Further sequence analysis emphasised that NYH enzymes featured highly conserved residues with N40 present in 97% of sequences and H203 in 95%. On the other hand, enzymes from the XYP sub-group displayed more variability at position 40 where five different amino acids were found: Q – 38.5%; N – 25%; K – 13.5%; S – 15.4% and A – 7.7%[65]. Interestingly, P203 was strictly conserved in all but one CDPS where a

glutamine was observed. These two sequence features were posed to play parallel functions in each group of CDPSs but to confirm their differences, Bourgeois *et al.* converted an XYP member into an NYH and vice versa[65]. However, this experiment only resulted in poor enzymatic activity from the XYP-turned-NYH CDPS and complete loss of activity in the mutated NYH enzyme. Therefore, the NYH and XYP motifs were not interchangeable due to the complex interactions necessary for product formation. It should be noted that sequence alignments and site directed mutagenesis are not sufficient evidence that both families do participate in the catalytic mechanism but it is a widely accepted theory within the CDPS field.

The research presented here aimed to investigate the structural features of a histidine binding cyclodipeptide synthase, *Parcu*CDPS. Whilst this enzyme is part of the XYP family, stereotypical CDPS structures were unable to be used as models for molecular replacement indicating unexpected structural dissimilarity. This new structure allowed for the design of mutants to probe the key residues responsible for histidine recognition. In doing so, a new generation of *Parcu*CDPS variants were successfully engineered to generate new peptides including cLP and cFP.

5.2 Crystal structure of *Parcu*CDPS

Following purification of *Parcu*CDPS, crystallisation was attempted as described in Section 2.8. Numerous commercial crystallisation screens were used, however this failed to yield any crystalline material. Therefore, *Parcu*CDPS was subjected to a pH buffer screen trials using a commercial screen from Molecular Dimensions. Differential scanning fluorimetry (DSF) was performed as described in Section 2.7.4 and the raw data were analysed in GraphPad Prism 9.5.1. The results from this experiment highlighted that *Parcu*CDPS displayed enhanced stability in citrate buffers ranging from pH 5.5 – 6.5 with pH 6.5 being the most optimal (Figure 5.2). This particular buffer gave a melting temperature (T_M) of 39.06 °C whereas the current HEPES purification buffer used for *Parcu*CDPS had a T_M of 37.67 °C. This almost 2 °C increase could be essential to achieving crystalline material of *Parcu*CDPS therefore the protein was purified again into a different buffer composed of 50 mM citrate pH 6.5, 150 mM NaCl, 2 mM DTT and 2 mM EDTA.

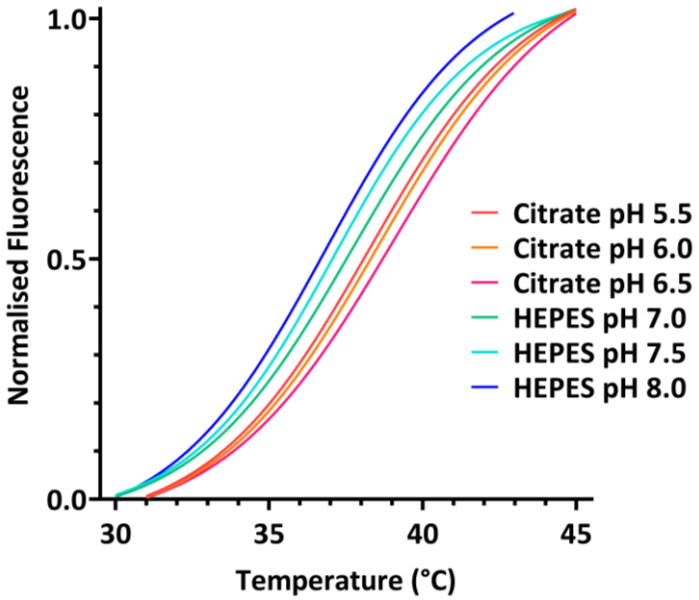


Figure 5.2 Differential scanning fluorimetry of *Parcu*CDPS.
The Durham pH screen was used to investigate the effect of buffer pH on the stability of ParcuCDPS. The optimal buffer was found to be citrate pH 6.5 which displayed a T_M of 39.06 °C. Melting temperatures were calculated using a sigmoidal fit employed by GraphPad Prism 9.5.1.

This new system yielded initial crystals in a condition containing 2.0 M ammonium sulfate and 0.1 M sodium acetate pH 5. These were improved by slightly altering each parameter until the optimal crystallisation condition was discovered – in this case, 2.27 M ammonium sulfate and 0.1 M sodium acetate, pH 5.17 (Figure 5.3).

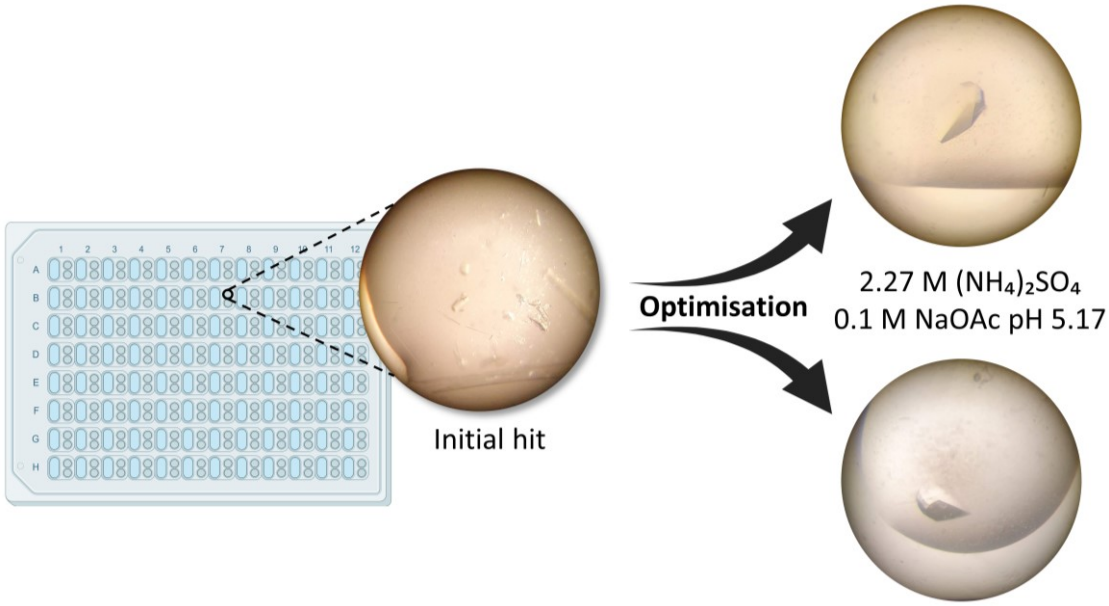


Figure 5.3 Optimisation of *Parcu*CDPS crystals.
Initial hits from the ProPlex™ screen were further optimised, uncovering 2.27 M ammonium sulfate and 0.1 M sodium acetate pH 5.17 to be the best condition for producing large crystalline material which could be used for diffraction purposes.

The resultant crystals diffracted to 1.9 Å resolution, however attempts at solving the structure using molecular replacement with other CDPSs from either NYH or XYP subgroup were unsuccessful. To overcome this issue, a second dataset was collected from iodine-soaked crystals using single wavelength anomalous diffraction (SAD) to create a model which could be used to solve the structure with the first dataset. Intriguingly, previous attempts to solve the first structures of XYP proteins also failed to use NYH-CDPSs as a model for molecular replacement and relied on selenomethionine to provide anomalous signal[65].

At the time of solving *Parcu*CDPS in March 2021, the novel structural prediction software AlphaFold (DeepMind/EMBL-EBI) did not include the predicted structure of *Parcu*CDPS within its database[163]. Therefore, a model could not be generated at that time to solve the structure of *Parcu*CDPS. Since then, AlphaFold has expanded to include almost every sequenced protein from the UniProt reference database. Consequently, Alphafold (in CCP4 Cloud)[163–165] predicted a model for *Parcu*CDPS based on structure and known homologs. However, a solution was not found when using this structure to perform molecular replacement (Phenix). On the other hand, when given *Snou*-CDPS to solve, Alphafold generated a relatively similar structure with an RMSD of 1.00 Å over 217 aligned residues when superimposed in WinCoot 0.9.4.1. Thus, it was clearly necessary to employ iodide-SAD phasing to formally solve the structure of *Parcu*CDPS WT.

Superficially, *Parcu*CDPS featured a Rossman fold which is characteristic of all CDPS structures to date (Figure 5.4A). Nine residues out of the total 230 were not traceable indicating a region of flexibility, therefore a small gap is present within the structure. Fortunately, this area does not appear to be part of the active site, which is composed of S26, Y167, E171 and Y191 - numbering respective to *Parcu*CDPS (Figure 5.3B). These residues are strictly conserved amongst the CDPS family and are commonly known for all the currently solved structures. It should be highlighted that Y167 points away from the active site, removing the stereotypical hydrogen bonding network observed in other CDPS structures. This residue is postulated to play the same role as Y178 in *Snou*-CDPS which stabilises the aminoacyl moiety formed from the binding of the first substrate to the catalytic serine (S26 in the case of *Parcu*CDPS). Moreover, an unusual feature was highlighted by the position of D58 whose side chain is located within the active site of *Parcu*CDPS (Figure 5.4B). This has not previously been noted for other CDPS structures and it has been hypothesised as an important residue with hydrogen bonding interactions possible between D58 and the catalytic residue S26.

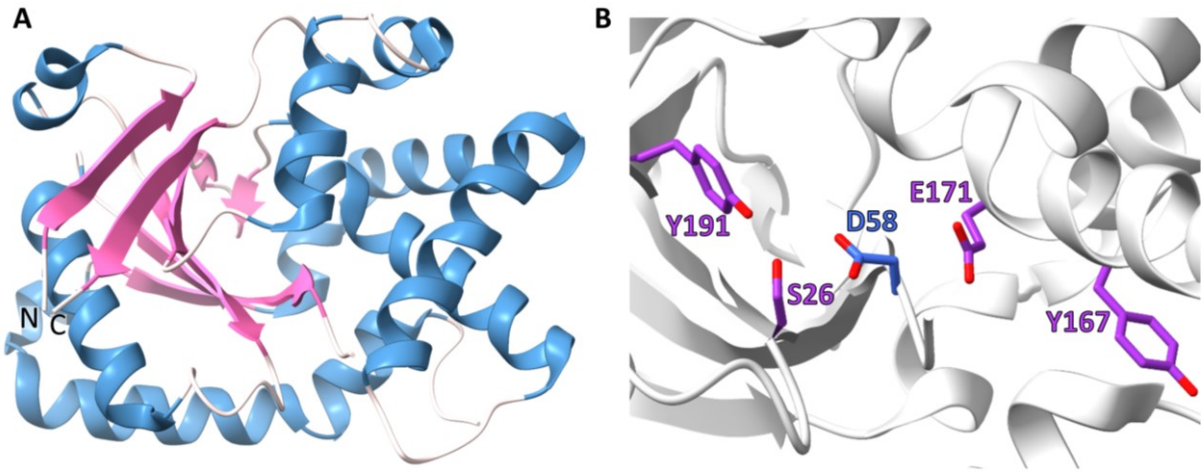


Figure 5.4 Structure of *Parcu*CDPS.
A – Overall structure of *Parcu*CDPS, depicted in ribbon form. Alpha helices are drawn in blue and beta strands in pink. **B** – Close up perspective of *Parcu*CDPS active site. Active site residues (S26, Y167, E171, and Y191) are shown in purple and a putative essential residue, D58, in blue.

Additionally, there are two pockets for substrate recognition - P1 which is found deep within the protein and P2 which is wider and more solvent accessible as denoted by Figure 5.5A. Previously, Gondry *et al.* predicted the sequence motifs of over 100 CDPSs based on the assumption that the positions of P1 pocket residues would be conserved amongst enzymes with the same specificity[62]. These predicted sequences were correct for *Parcu*CDPS where P1 is composed of G22, A24, Y55, A57, W112, E174, L175, and Y189 and P2 contains F144, E147, W148, K151, Y193, T195, and R196 (Figure 5.5B). The particular sequences of each pocket are rather unique for a CDPS where the most common pocket residues are typically uncharged amino acids such as valine, leucine, and phenylalanine[62].

The electrostatic environment of both pockets was investigated using PROPKA, which calculated the theoretical pK_a of each residue at pH 7.0 (Table 5.1)[166,167]. This highlighted that only E174 contributed to a net negative environment within P1 by having a predicted pK_a of 3.32 which is shifted from the typical 4.5 expected of this residue's side chain. Additionally, PROPKA highlighted a plausible sidechain hydrogen bonding interaction between E174 and Y55 with the predicted pK_a of Y55 to have shifted to 13.27, making this residue an even weaker acid unlikely to donate its proton in the reaction pH of 7. On the other hand, the pK_a values of P2 residues did not reveal any distinguishing trends unfortunately, and instead appeared to match the expected pK_a values. This could be a consequence of the rather large surface area of P2, stopping the pocket residues from being close enough to generate attractive/repulsive forces which would then perturb the overall pK_a. Further research into these pocket environments was aided by an online resource known as Computed Atlas of Surface Topography of proteins (CASTp) which measured the area and volume of any particular pocket on the surface or within a protein[168]. It was evident that P1 was situated deep within the CDPS core, close to the catalytic serine which is to be expected with the mechanism of action of these enzymes (Figure

5.5C and D). UCSF ChimeraX also calculated the electrostatic environment of this area to be negative which supports the prediction by PROPKA. The missing region from residue 61-70 spans across a section of P2 which limits conclusions that can be drawn from this analysis. However, electrostatic investigation revealed that this pocket contained positively charged areas in comparison to P1. The surface area of P2 was much larger than P1, unsurprisingly, which has been postulated to aid the promiscuity of these enzymes and allow the acceptance of very different substrates - for example glutamate and proline in the case of *Parcu*CDPS.

Table 5.1 Calculated pK_a values of *Parcu*CDPS pocket residues using PROPKA.
Residues in P1 are shaded in blue and residues in P2 are in red.

Residue	Expected pK_a	Calculated pK_a
Y55	10.00	13.27
E174	4.50	3.32
Y189	10.00	10.06
E147	4.50	4.50
K151	10.50	10.50
Y193	10.00	10.00
R196	12.50	12.50

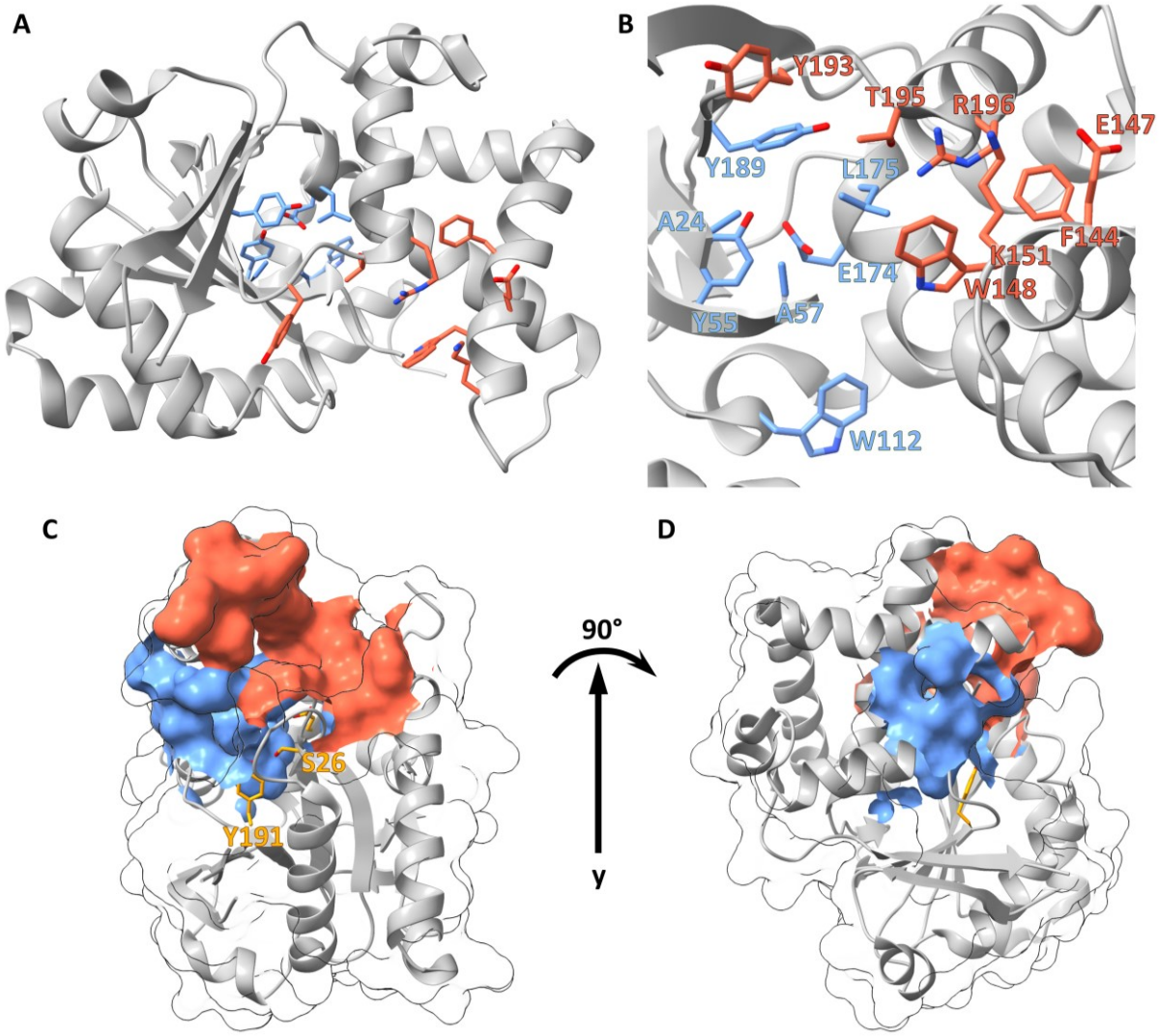


Figure 5.5 Binding pocket environments of *Parcu*CDPS.
A – Pocket residues of *Parcu*CDPS drawn in stick form. Residues from P1 are shown in blue whilst residues from P2 are in red. **B** – Close up view of pocket residues with labels (same colour coding as A). **C** – Pocket surfaces shown in blue for P1 and red for P2, as calculated by CASTp[168]. The active site residues are highlighted in orange to envision where the active site sits in relation to the two pockets. **D** – 90° rotation of structure to highlight the depth of P1 within the protein core.

Moreover, online tools were used to confirm structurally similar proteins based on *Parcu*CDPS: HHpred server (https://toolkit.tuebingen.mpg.de/tools/hhpred), the DALI server (http://ekhidna2.biocenter. helsinki.fi/dali) and the PDBeFold server (http://www.ebi.ac.uk/msd-srv/ssm). HHpred (MPI Bioinformatics Toolkit) used a single sequence - *Parcu*CDPS here - with the default parameters to search alignment databases for proteins with similar sequences[169,170]. Both DALI (Biocenter Finland)[171] and PDBeFold (European Bioinformatics Institute)[172] methods required the coordinates of *Parcu*CDPS: DALI compared the structure to all entries in the Protein Data Bank whilst PDBefold compared 187732 PDB entries to *Parcu*CDPS but only found 7 which were similar. All three

methods highlighted that *Shae*-CDPS was the most similar to *Parcu*CDPS, a NYH CDPS which contained a T instead of N at position 40 (Table 5.2)[65]. The sequence identity of *Parcu*CDPS when compared to the CDPS family remained low regardless of the member chosen. This finding corroborates our previous evidence that *Parcu*CDPS is a unique CDPS that utilises an amino acid substrate which is uncommon, with only two histidine-accepting CDPSs being characterised to date.

Table 5.2 Summary of the top hits from structural similarity searches.
Using three different methods - HHPred; Dali, and PDBeFold - with *Parcu*CDPS to highlight similar proteins for comparison. The RMSD (), $N_{aligned}$ and sequence identity values were calculated by each individual analytical method thus allowing for discrepancies between the values for the same protein.

Method	PDB ID	Name	Organism	RMSD (Å)	$N_{aligned}$	Sequence identity (%)
HHpred	6EZ3_A	*Shae*-CDPS	*Staphylococcus haemolyticus JCSC1435*		196	14
	6ZTU_A	BtCDPS	*Caldibacillus thermoamylovorans*		201	19
	2X9Q_B	Rv2275	*Mycobacterium tuberculosis H37Rv*		208	17
	3OQV_A	AlbC (aka *Snou*-CDPS)	*Streptomyces noursei*		200	18
	3OQI_A	YvmC-Blic with CHES bound	*Bacillus licheniformis DSM 13 = ATCC 14580*		203	17
Dali	6EZ3_F	*Shae*-CDPS	*Staphylococcus haemolyticus JCSC1435*	3.0	192	15
	2X9Q_A	Rv2275	*Mycobacterium tuberculosis H37Rv*	3.2	192	17
	3S7T_A	YvmC-Blic seleno-met form	*Bacillus licheniformis DSM 13 = ATCC 14580*	2.8	184	16
	3OQH_A	YvmC-Blic	*Bacillus licheniformis DSM 13 = ATCC 14580*	2.9	190	16
	6ZTU_A	BtCDPS	*Caldibacillus thermoamylovorans*	2.9	190	17
PDBeFold	6EZ3_F	*Shae*-CDPS	*Staphylococcus haemolyticus JCSC1435*	2.78	178	14
	3OQI_A	YvmC-Blic with CHES bound	*Bacillus licheniformis DSM 13 = ATCC 14580*	2.84	174	16
	5OCD_J	*Fdum*-CDPS	*Fluoribacter dumoffii NY 23*	2.95	181	12
	4Q24_A	*Snou*-CDPS-S37C bound to reaction intermediate mimic – ZPK	*Streptomyces noursei*	3.04	176	16

Using Secondary Structure Matching (SSM) Superpose from Wincoot 0.9.4.1, a global comparison of *Parcu*CDPS with the identified CDPSs from Table 5.2 further emphasised the unique regions of *Parcu*CDPS (Figure 5.6). All CDPSs contained a long continuous alpha helix (α4 or α5 in the case of BtCDPS), however in *Parcu*CDPS this helix is split into two parts, causing a change in direction due to the flexible amino acids present including glycine (G84) and arginine (R82). Another distinguishing element is β3 which curves into the active site of *Parcu*CDPS and effectively positions D58 within hydrogen bonding distance of S26 – this is not present in any other CDPS structure. Moreover, *Parcu*CDPS contains two small beta strands (β5 and β5') which are again absent from other CDPSs – these however do not feature any residue relevant for catalysis and may be more of a superficial difference rather than functional. Residues belonging to P1 of *Parcu*CDPS are found across β3 and α4, both of which are deep within the protein itself. The small size of this region emphasises the strict binding network required for substrate recognition by *Parcu*CDPS and aids our understanding as to why few unnatural histidine analogues were accepted. On the other hand, P2 residues are in helices 4 and 6 which both sit in different positions to other CDPSs leading to a rather large diffuse pocket environment. Coincidentally, this could explain why *Parcu*CDPS can accept two different amino acids in the same region as different bonding arrangements could be achieved within the one pocket. Overall, it is evident that whilst *Parcu*CDPS does feature a Rossman fold and two pockets, there are actually multiple details which are divergent from the archetypal CDPS structure. These are clearly contributing factors to the unique substrate selection of histidine which is rare amongst this family of enzymes.

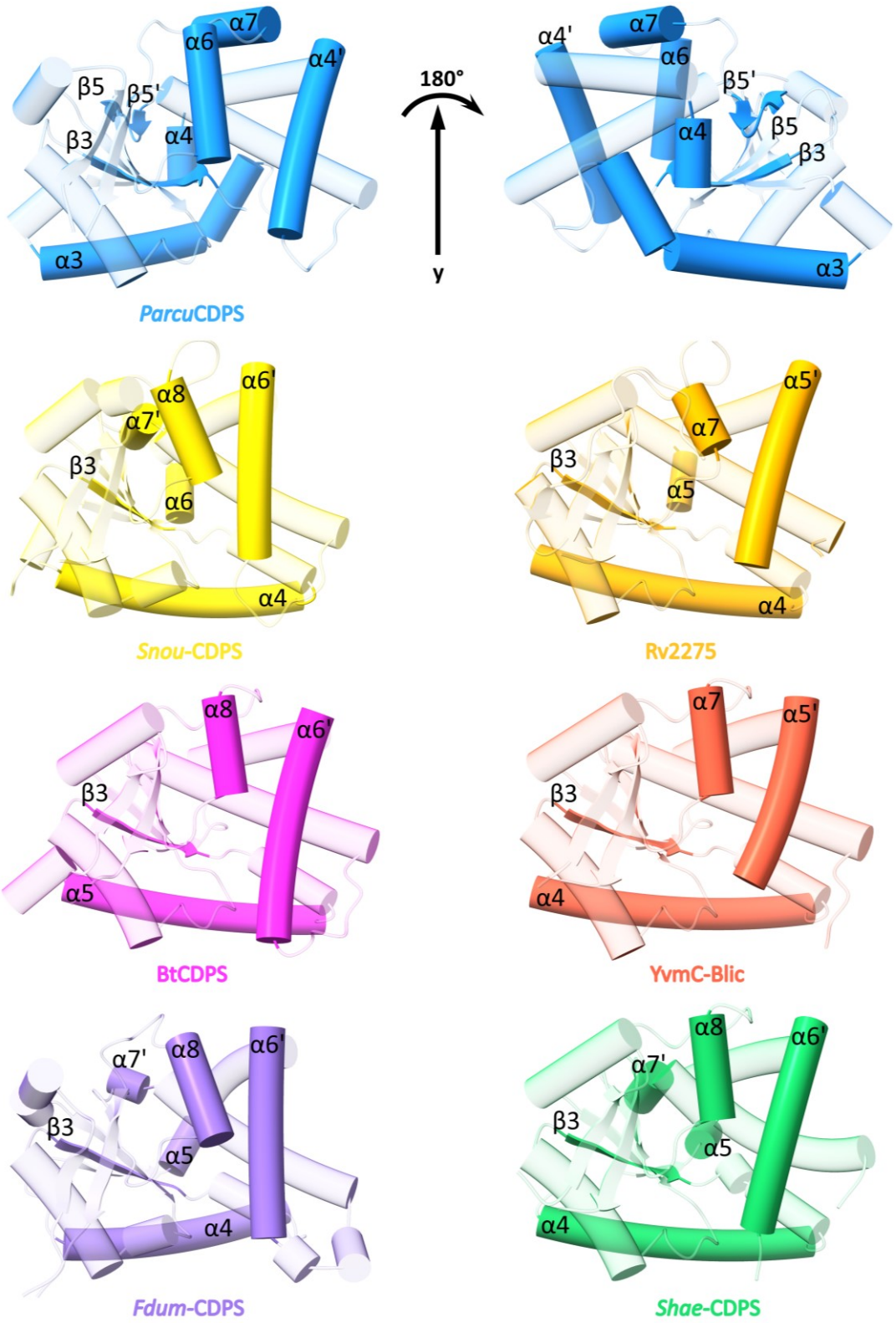


Figure 5.6 Structural comparison of *Parcu*CDPS.
*Parcu*CDPS (blue, PDB: 7QB8) drawn with tubes representing alpha helices and divergent features are shown in 100% opaque colour and labelled. A 180° rotation snapshot is also included to highlight the presence of β5 and α7, both of which are absent in the other CDPSs. Six CDPSs are compared here, chosen by three structure similarity servers (Table 5.2): *Snou*-CDPS (yellow, PDB: 3OQV); Rv2275 (orange, PDB: 2X9Q); BtCDPS (pink, PDB: 6ZTU), YvmC-Blic (red, PDB: 3OQH); *Fdum*-CDPS (purple, PDB: 5OCD) and *Shae*-CDPS (green, PDB: 6EZ3).

5.3 Active site mutations disrupt *Parcu*CDPS activity

To validate the hypothesis that *Parcu*CDPS would partake in the same mechanism as *Snou*-CDPS, the conserved catalytic residues – S26, Y167, E171 – were mutated and their activity measured using the *in vitro* reaction system (Figure 5.7). The proposed nucleophile, S26, was investigated in two ways: abolition of the nucleophilicity using an alanine and increasing the nucleophilicity by introducing a cysteine in its place. No activity was observed from the S26A mutant whilst S26C retained around 40% activity. It should be noted that these particular mutations have been performed with other CDPSs, where similar trends of lowered enzymatic activity were observed[55,65]. The S26C mutant, whilst being a better nucleophile, appears to form an activated thioester bond which is susceptible to hydrolysis before *Parcu*CDPS can effectively turn the acyl-enzyme complex into product. Furthermore, the intriguing role of D58 was explored by replacing this residue with an alanine to disrupt the potential hydrogen bonding interaction with S26. Additionally, the carboxylic acid functionality of aspartic acid was replaced by an amide group from asparagine thus removing the negatively charged environment which would attract the incoming histidine substrate. Both of these mutations severely hindered the activity of *Parcu*CDPS thus confirming the theory that D58 is essential for catalysis.

Together Y167 and E171 were postulated to position the dipeptidyl intermediate correctly for cyclisation, therefore these residues were also mutated to investigate their role in CDP formation. Both E171 mutants – E171A and E171Q - exhibited complete loss of activity, highlighting the need for the carboxylate group. Additionally, the Y167A mutant was unable to yield any measurable quantity of cHP product, however Y167F retained 17% of the original activity. This corroborates other research indicating that the π interactions of the benzene ring aid the substrate binding alongside the hydroxyl group hydrogen bonding of tyrosine[65]. Therefore, this evidence demonstrates that *Parcu*CDPS does act via the same catalytic mechanism proposed for other members of the CDPS family. Despite the clear structural differences, the four conserved catalytic residues adopt the same roles as in *Snou*-CDPS and are undoubtedly essential for enzymatic activity.

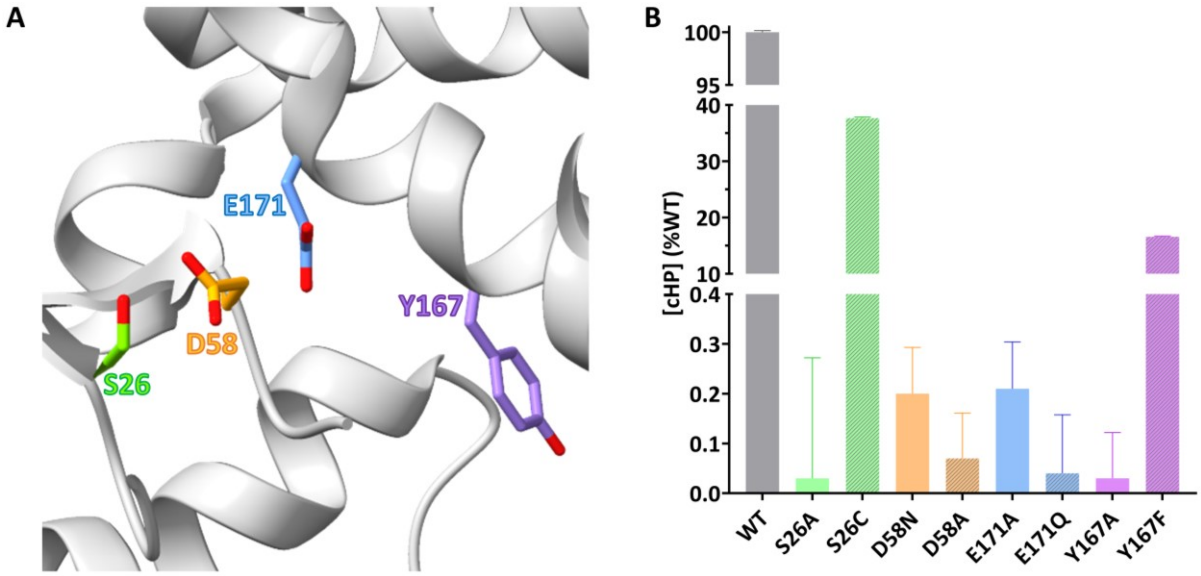


Figure 5.7 Active site mutants destroy enzymatic activity.
A – Structural depiction of the active site residues which were targeted for mutagenesis: S26 (green); D58 (orange); E171 (blue) and Y167 (purple). **B** – Assessment of enzymatic activity using the previously described *in vitro* reaction. The wildtype (WT) was classed as '100% activity' to which the mutant activity was calculated as a percentage of (% WT). The bar graphs are coloured respective to the residues in Figure **A** and the error bars represent the standard error of the mean (SEM).

5.4 Mutations within P1 greatly impact enzymatic activity

5.4.1 Comparison of *Parcu*CDPS and HisRS substrate recognition mechanisms

As proven by the acyl-enzyme intermediate experiment (results presented in Section 3.6), *Parcu*CDPS accepts histidine first in P1 using residues which are uncommon for a CDPS - G22, A24, Y55, A57, W112, E174, L175, and Y189. To determine which residues were of relevance for substrate recognition, the active site of HisRS from *E. coli*, another histidine accepting enzyme, was examined for similarities to *Parcu*CDPS. In 1995, a structure was solved where the histidyl-adenylate (His-AMP) substrate was bound to the HisRS active site, allowing for the identification of essential residues for recognition[173]. Arnez *et al.* revealed the intricate network of interactions employed by HisRS to anchor the substrate in place, and more specifically a Thr-Glu-Tyr triad present for histidine recognition (Figure 5.8). The carboxylate group of E131 interacts directly with the τ-nitrogen from the imidazole ring by using T85 and Y288 to orient the Glu side chain into the optimal configuration for interaction[174]. Additionally, Y264 is responsible for anchoring the π-nitrogen using hydrogen bonds, therefore there are two distinct molecular requirements that must be satisfied by the substrate for successful recognition by HisRS.

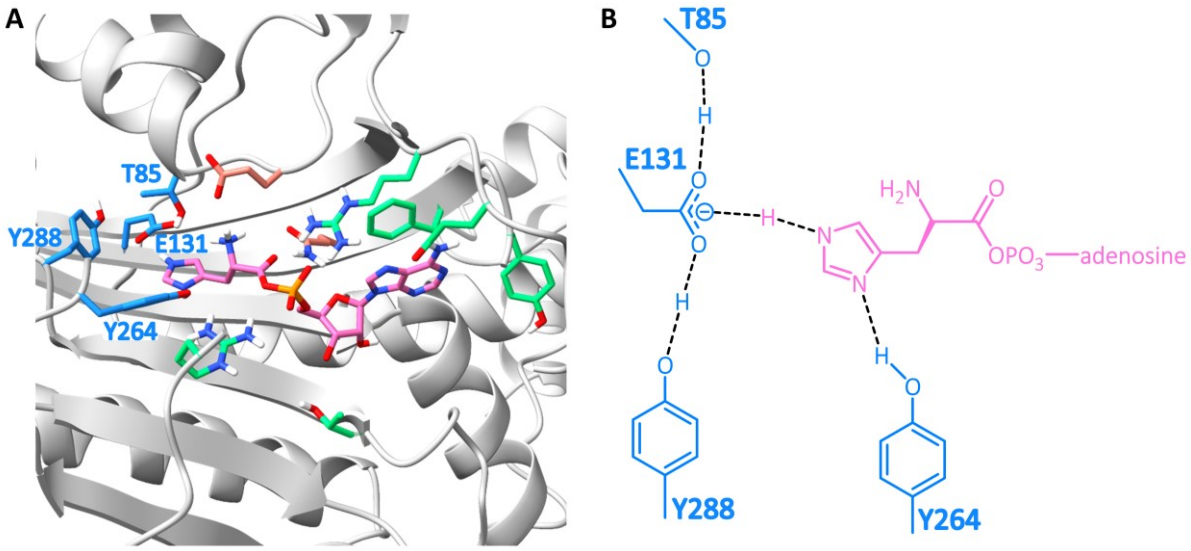


Figure 5.8 Histidine recognition network of HisRS.
A – With the ligand, histidyl-adenylate, bound to HisRS (PDB: 1KMM)[174], a distinct network of intramolecular interactions responsible for substrate binding are highlighted. The residues are coloured depending on which region of the ligand they interact with: blue for histidine side chain, orange for peptide backbone and green for AMP. **B** – Simplified view of the hydrogen bond network employed by HisRS to recognise histidine. T85, E131 and T288 form a triad of interactions with the τ-nitrogen of the imidazole ring whilst Y264 interacts solely with the π-nitrogen.

Comparison of *Parcu*CDPS pocket residues revealed a similar hydrogen bonding network that could be part of the histidine recognition mechanism (Figure 5.9). Where HisRS featured E131 aided by Y288, *Parcu*CDPS contains E174 and Y55 within hydrogen bonding distance of each other. Additionally, Y264 of HisRS points in a different direction to specifically interact with the π-nitrogen which is mirrored by Y189 in *Parcu*CDPS. Thus, it can be anticipated that histidine would interact with Y55, E174 and Y189 from the CDPS in a similar fashion as observed in HisRS.

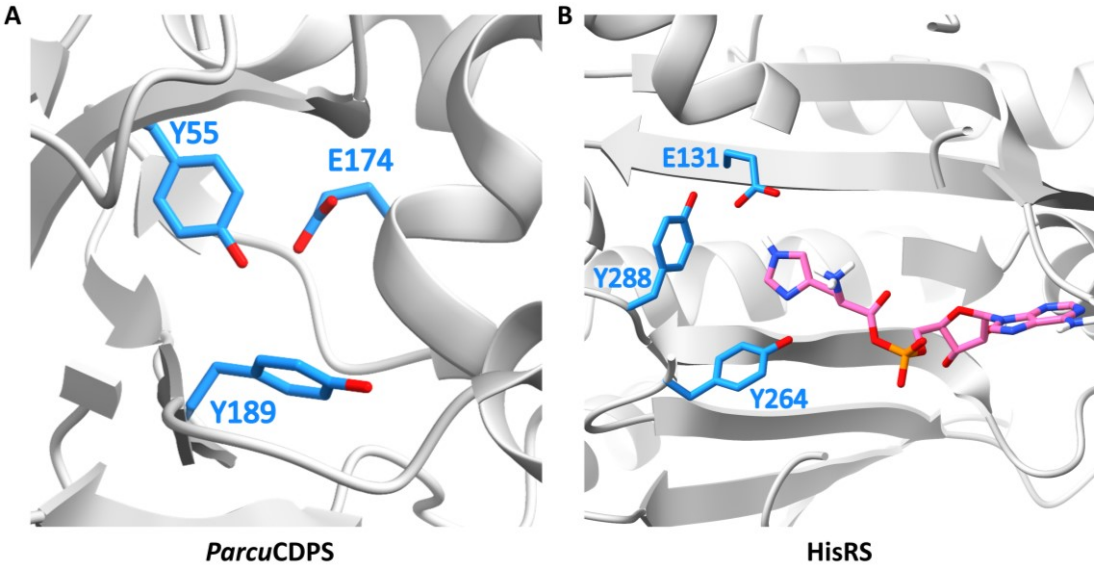


Figure 5.9 Comparison of histidine recognition sites in *Parcu*CDPS and HisRS.
A – Pocket residues of *Parcu*CDPS postulated to interact with histidine. **B** – Active site of *Ec*-HisRS showing only the residues responsible for histidine recognition. This clearly highlights similarities between both of these active sites which recognise histidine.

5.4.2 Rational engineering of *Parcu*CDPS pocket residues

Site directed mutagenesis was performed on Y55, E174 and Y189 to generate a series of single mutants initially, whose activity was tested using the same *in vitro* experiment monitoring the production of cHP (Figure 5.10). The hydrogen bond capabilities of all three residues were interrupted by replacing with either an alanine or a phenylalanine, creating Y55F, E174A and Y189F variants. It was envisioned that replacing E174 with histidine (E174H) would reverse the overall polarity of the pocket and introducing glutamine with an amide functionality (E174Q) would remove the anionic carboxylate group and interrupt the hydrogen bonding network. Three mutants were also prepared using structural comparison to convert each position into the corresponding *Snou*-CDPS residue – Y55V, E174L and Y189L.

In short, it was apparent that these residues were essential in histidine recognition as over 80% of the activity was lost in each mutant. The most successful variant was Y189F which produced 12% of the wild-type activity *in vitro*. This may hint at the less important role of π-nitrogen interactions when the τ-nitrogen interactions remain intact. Replacing E174 with alanine, histidine or glutamine resulted in a 99% decrease in activity whilst E174L maintained around 7% of the original wild-type activity. Again, both mutations of Y55 destroyed the production of cHP despite the presence of the E174 residue, therefore it can be assumed that both Y55 and E174 are vital for the recognition of histidine in *Parcu*CDPS.

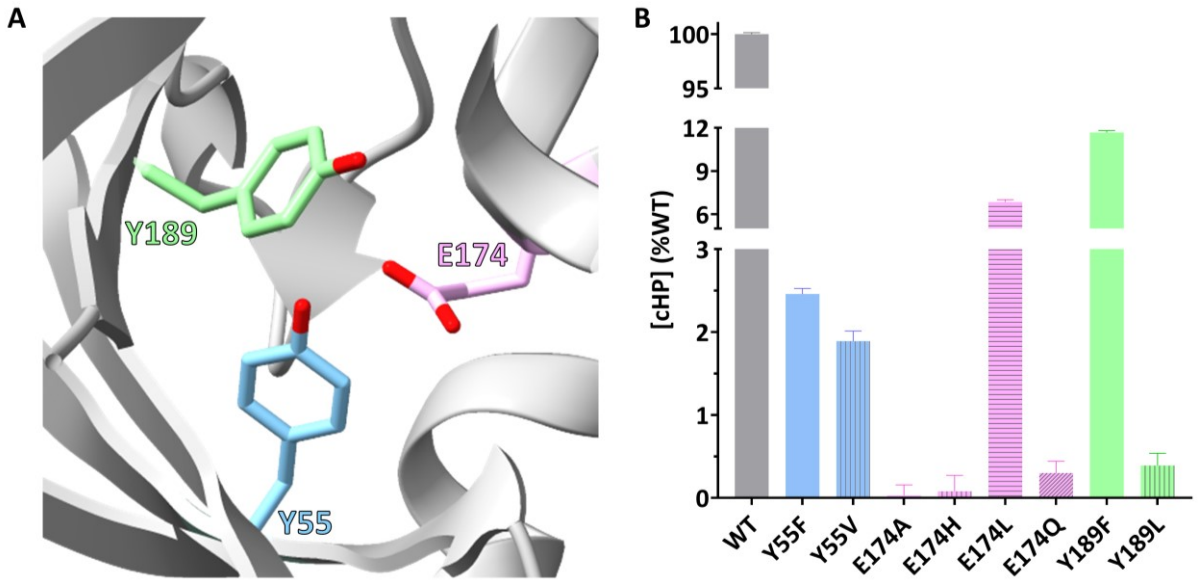


Figure 5.10 *Parcu*CDPS P1 mutations destroy cHP activity.
A – Structure of *Parcu*CDPS with the pocket residues of interest drawn in stick form: Y55 (blue); E174 (pink) and Y189 (green). **B** – Comparison of *Parcu*CDPS mutant *in vitro* activity to the wildtype, displayed as a percentage. Again, the colours correlate to the residues in **A** and the error bars represent the standard error of the mean (SEM) from the triplicate measurements for each mutant.

With the aim of broadening our understanding of the binding site of histidine, six double mutants and one triple mutant were created – Y55F + Y189F; Y55F+E174A; Y55V+Y189L; Y55V+E174L; Y189F+E174A; E189L+E174L and Y55V+E174L+Y189L. They were tested in the same manner as the previous mutants to quantify the activity in relation to wildtype *Parcu*CDPS. These mutants highlighted the essential roles of all three residues as not one was able to produce even 1% of cHP produced by the wildtype (Figure 5.11). As expected, the triple mutant displayed the least amount of activity owing to the lack of interaction opportunities for the incoming histidine substrate. These experiments confirmed the vital roles of Y55, E174 and Y189 and highlighted that not one single residue was entirely responsible for recognition but rather an intricate balance of interactions between the three.

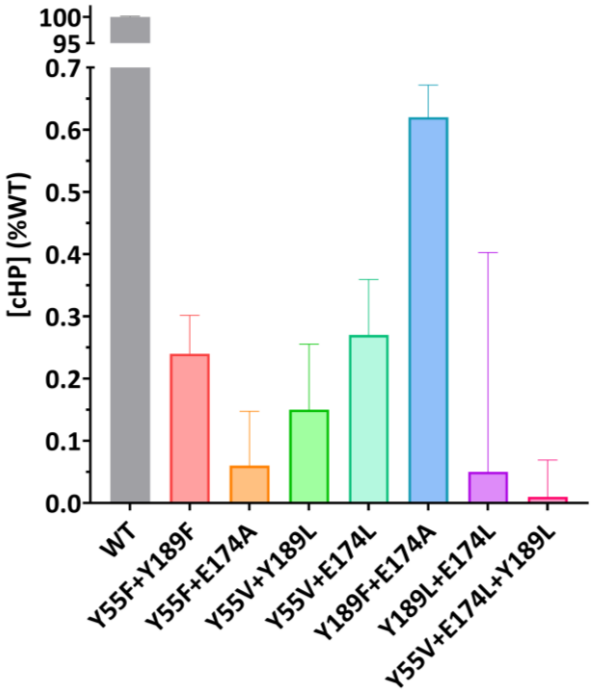


Figure 5.11 Multiple *Parcu*CDPS mutations destroy cHP production.
Production of cHP *in vitro* using double and triple mutants of *Parcu*CDPS, calculated as percentage of wildtype activity. All mutants failed to produce more than 1% of the original wildtype product formation. The experiments were performed in triplicate and the error bars represent the standard error of the mean (SEM) as calculated by GraphPad Prism 9.5.1.

5.4.3 *Parcu*CDPS mutants exhibit altered substrate specificity

Although *Parcu*CDPS mutants were no longer capable of recognising histidine, it was then hypothesised that their substrate recognition capacity had instead been altered to favour different non His-tRNA^{His} substrates. Considering some residue changes were based on the pocket of *Snou*-CDPS which synthesises cyclo(L-Phe-L-Leu), it was a possibility that mutants containing the same residues would also accept phenylalanine or leucine as substrates. Indeed, this hypothesis was tested as before using purified tRNA and aaRSs with each individual mutant and the resultant peak area was quantified using a calibration curve with a commercial standard. This method however would only be

successful if the variant enzyme could tolerate both the change in the incoming amino acid and the tRNA body to which it would be attached initially. Unfortunately, this proved to be challenging for the CDPS and poor product formation was observed for both cFP and cLP. The highest yield of cFP when using Phe-tRNA^{Phe} was approximately 2 µM from E174L+Y189L and the highest yield of cLP with Leu-tRNA^{Leu} was around 5 µM for Y55V+E174L (Figure 5.12).

These results suggested that using aminoacylated tRNA was not the optimal substrate for the variants to produce new CDPs, which consequently led to the use of aa-DBE substrates instead. These minimal units bypassed the requirement for charged tRNA and aided the CDPS in accepting the new amino acid substrate. Using the same *in vitro* experimental setup highlighted that the *Parcu*CDPS mutants were capable of using aa-DBE to yield higher amounts of cFP and cLP compared to aa-tRNA substrates. For cFP, E174L+Y189L produced the largest quantity of product (6 µM) whilst Y55V+E174L was the most successful mutant for cLP (7 µM). It should be highlighted that Y55V was the most productive single mutant out of the three and produced comparable amounts of products as some of the double mutants. This would imply that this residue in particular is pivotal for amino acid recognition as proven by the lack of cHP production observed when Y55 is removed. Indeed, not all mutations were productive as is the case for Y189L which did not yield measurable quantities of either cFP or cLP, indicating this particular mutant was unable to improve recognition of any tested substrate. Furthermore, the triple mutant Y55V+E174L+Y189L could produce less than 1 µM of either new peptide as well as failing to yield the natural cHP. This suggested that the pocket had been entirely destroyed and the CDPS was unable to recognise substrates of any nature - this could, of course, be confirmed by structural investigations. Furthermore, the boundaries of this new substrate selectivity were tested by introducing amino acids of a similar structure like valine and isoleucine coupled to the DBE subunit. *In vitro* activity investigations did not detect the presence of any cyclo(L-Val-L-Pro) or cyclo(L-Ile-L-Pro) indicating that the *Parcu*CDPS mutants were still actively selecting their substrates and that whilst the overall specificity had been expanded, it remained restricted to a select few amino acids.

Verification of altered substrate binding to P1 of *Parcu*CDPS mutants was performed using the previously described trapped acyl enzyme intermediate experiment (Section 2.5.20). The double mutants, Y55V+Y189L, Y55V+E174L and E174L+Y189L - which were capable of producing cFP and cLP - were tested with His-DBE, Phe-DBE and Leu-DBE (Figure 5.13). All three variants exhibited reduced capacity to accept histidine in P1 with E174L+Y189L completely unable to form a covalent attachment to His-DBE. This would explain the reduced yield of cHP (Figure 5.11) from these mutants which clearly struggle to perform the first acyl-enzyme binding event that is integral to the overall reaction mechanism. The evidence of altered substrate specificity is strengthened by the presence of both Phe-

and Leu-enzyme intermediates when the mutants are incubated with the DBE substrates. Only Y55V+Y189L did not display binding to leucine, however production of cLP was minimal with this mutant therefore it is possible that such a low concentration of acyl-enzyme intermediate would be unobservable here. In general, this experiment reinforced the hypothesis that select mutations in P1 have triggered a switch in substrate specificity of *Parcu*CDPS.

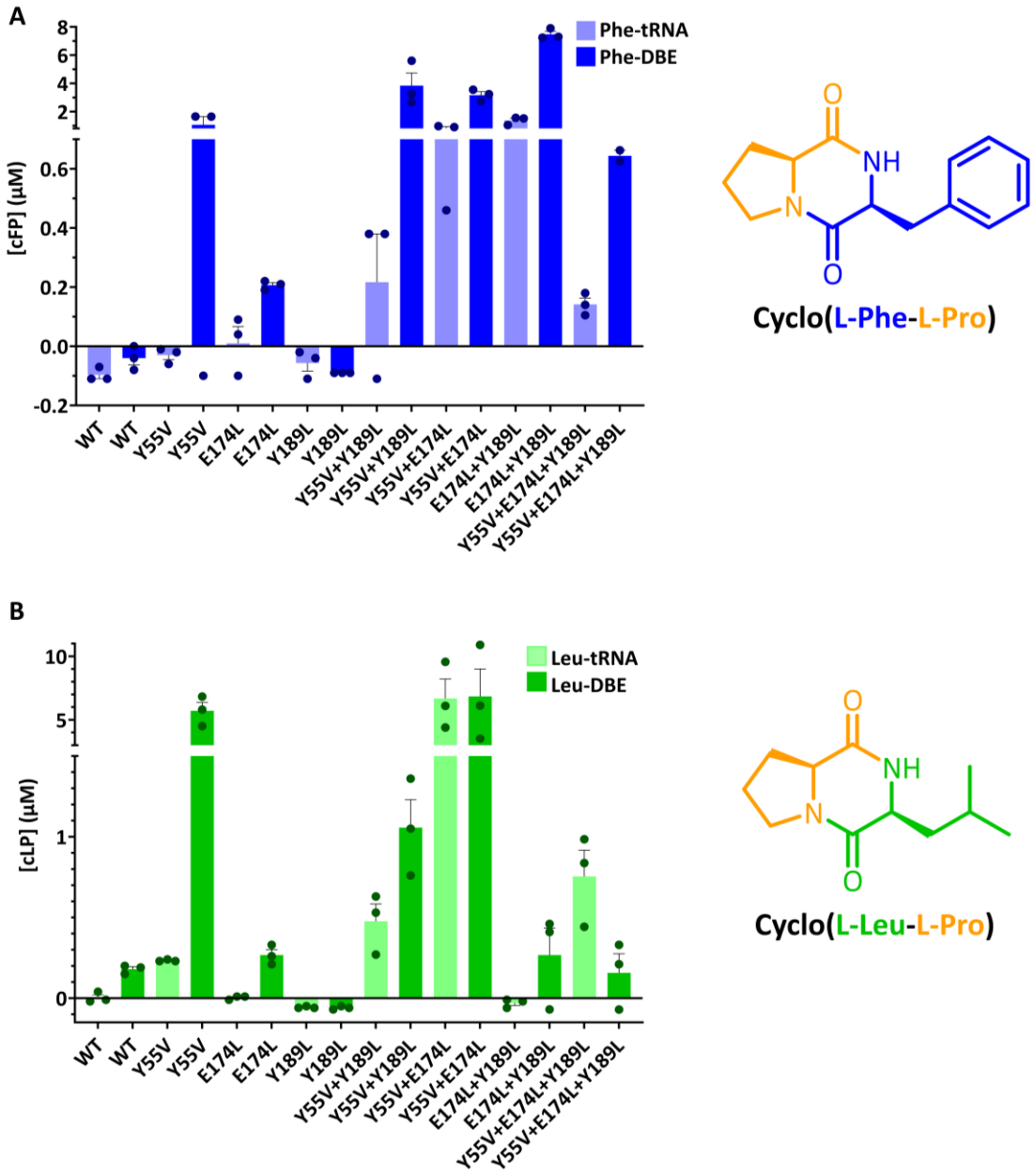


Figure 5.12 Alternative cyclic dipeptide production using mutant *Parcu*CDPS.
A – Quantification of cFP produced by a series of *Parcu*CDPS mutants. Both types of substrates are shown on the graph: Phe-tRNA^{Phe} in light blue and Phe-DBE in dark blue. The structure of the product is shown to the right of the graph with the proline segment in orange and phenylalanine in blue. **B** – Quantification of cLP as shown for cFP - both leucine substrates are coloured green. All experiments were performed in triplicate with each point plotted and the standard error of the mean shown as a line.

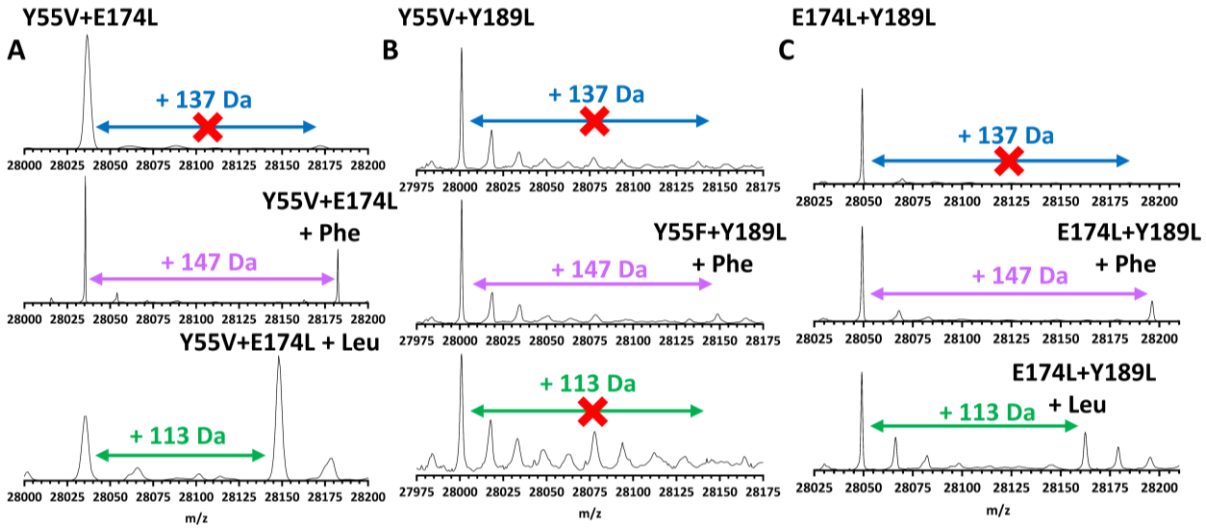


Figure 5.13 Investigation of *Parcu*CDPS mutant P1 binding.
The three double mutants - **A** – Y55V+E174L; **B** - Y55V+Y189L and **C** – E174L+Y189L - which displayed a shift in substrate specificity were subjected to the trapped acyl-enzyme intermediate experiment with the three potential substrates: histidine (blue), phenylalanine (purple) and leucine (green).

5.5 Biophysical properties of *Parcu*CDPS variants

5.5.1 Structural characterisation of select *Parcu*CDPS mutants

Structural investigations of the mutants aimed to confirm the integrity of the pocket structure remained intact throughout the rounds of mutagenesis. All variants were initially subjected to crystallisation trials, from which six mutants gave productive crystalline material for the determination of said structures using the wildtype structure as the molecular replacement model: Y55F, D58N, E171Q, E174A, E174L, Y189F (Figure 5.14 and Table 5.3). All solved structures appear similar to the wildtype *Parcu*CDPS with RMSD deviations of less than 1 Å when superimposed using WinCoot 0.9.4.1. The Rossman fold remained intact and the two pocket features were still recognisable regardless of mutation. Both tyrosine to phenylalanine mutants (Y55F and Y189F) still had the benzyl ring pointing in the same direction as the wildtype. As stated previously, the hydroxyl hydrogen bonding capability is not the only interaction tyrosine participates in and the benzyl π interactions are also important for CDPSs which these mutants satisfy with the phenylalanine group. On the other hand, both E174 variants highlight a change in the position of the residue which would open up the pocket site for other substrates. Both mutants cannot contribute to the hydrogen bonding network found in the wildtype which may explain their lack of activity concerning cHP production. Additionally, the E171Q highlights a change in the direction of the residue which now points out of the active site and fails to maintain the hydrogen bond with Y55. Thus, this structural evidence likely explains the poor yield of cHP when using E171Q. Lastly, the D58N mutant highlights that the position is not the only important factor as the asparagine residue does point in the same direction, however the amide replacing the carboxylate group removes the necessary interactions with the catalytic serine for enzymatic activity.

However, it should be noted that any conclusions drawn here are limited due to only six structures being successfully solved out of the 23 purified mutants.

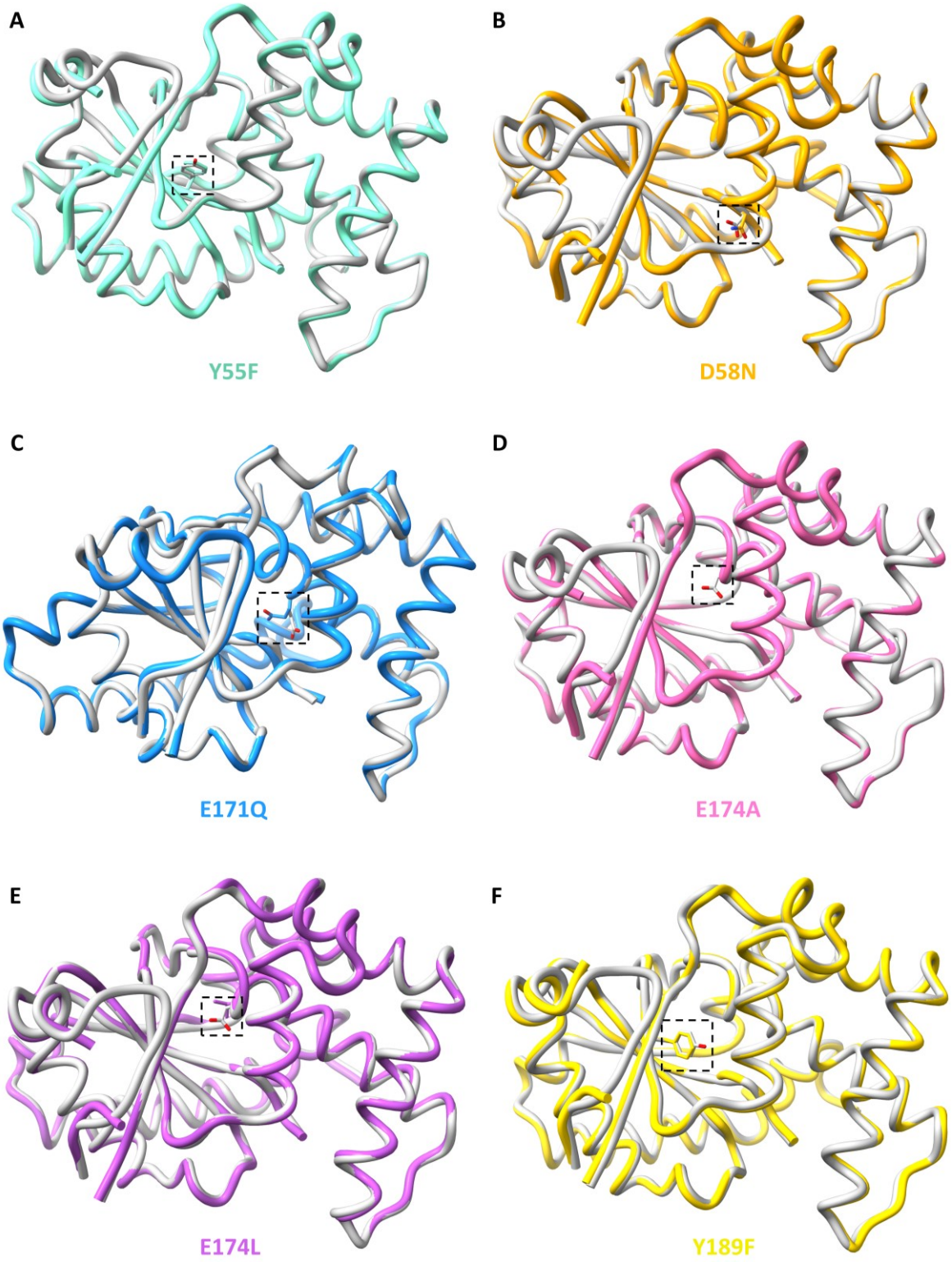


Figure 5.14 Structures of ParcuCDPS mutants.
Structures were superimposed using SSM in WinCoot 0.9.4.1 and visualised in UCSF ChimeraX version 1.2. *Parcu*CDPS mutants are coloured to match previous graphs whilst wildtype is white throughout. The residues of interest are highlighted by a black dashed line box. These particular variants superimposed well onto the original wildtype structure, indicating that the mutation does not interrupt the native protein folding.

Table 5.3 Crystallography data processing and refinement statistics of *Parcu*CDPS and related mutants.

	***Parcu*CDPS WT**	***Parcu*CDPS Y55F**	***Parcu*CDPS D58N**	***Parcu*CDPS E171Q**	***Parcu*CDPS E174A**	***Parcu*CDPS E174L**	***Parcu*CDPS Y189F**
PDB accession code	7QB8	7QAY	7QAU	7QAX	7QAQ	7QAT	7QAW
Data collection							
Resolution (Å)*	51.31 - 1.90 (1.97 - 1.90)	45.28 - 2.09 (2.12 - 2.09)	46.26 - 2.30 (2.38 - 2.30)	46.05 - 2.09 (2.16 - 2.09)	44.85 - 2.40 (2.49 - 2.40)	73.29 – 2.40 (2.44 – 2.4)	45.26 - 2.29 (2.37 - 2.29)
Space group	P 42 21 2	P 42 21 2	P 42 21 2	P 42 21 2	P 42 21 2	P 42 21 2	P 42 21 2
Unit cell parameters – a, b, c (Å)	102.61, 102.61, 49.76	101.26, 101.26, 50.07	103.44, 103.44, 49.95	102.97, 102.97, 49.81	102.60, 102.60, 49.86	103.63, 103.63, 49.68	101.20, 101.20, 50.32
Unit cell parameters – α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Multiplicity*	52.5 (48.2)	12.9 (7.9)	25.1 (25.4)	14.5 (5.5)	23.9 (24.3)	25.5 (22)	15.3 (14.5)
Completeness (%)*	99.86 (99.86)	94.1 (100.0)	98.92 (97.62)	92.57 (96.84)	99.94 (99.91)	100 (100)	99.12 (99.92)
Mean I/σ(I)*	31.45 (1.95)	23.74 (1.44)	27.09 (4.21)	15.80 (0.90)	26.17 (7.24)	26.00 (0.70)	18.80 (1.30)
R_{merge}*	0.073 (1.452)	0.069 (1.452)	0.088 (1.435)	0.090 (1.637)	0.155 (0.9867)	0.055 (0.884)	0.072 (2.230)
R_{pim}*	0.010 (0.207)	0.027 (0.804)	0.018 (0.287)	0.023 (0.807)	0.033 (0.202)	0.011 (0.189)	0.019 (0.604)
$CC_{1/2}$*†	0.999 (0.926)	1.000 (0.490)	1.000 (0.892)	0.999 (0.331)	0.999 (0.963)	1.000 (0.947)	1.000 (0.511)
Refinement							
Total reflections*	1128270 (101638)	192351 (6030)	311233 (30217)	222211 (4134)	260149 (25838)	282452 (11766)	187130 (8625)
Unique reflections*	21493 (2107)	14927 (1536)	12408 (1189)	15221 (1565)	10887 (1064)	11063 (534)	12240 (1215)
R_{work}	0.2248	0.2108	0.2096	0.2260	0.1955	0.2139	0.2187
R_{free}	0.2419	0.2625	0.2473	0.2647	0.2451	0.2606	0.2475
Number of atoms							
Protein	1870	1813	1825	1870	1810	1809	1820
Ligand	0	0	0	0	0	0	0
Water	40	15	11	15	24	8	9
***B*-factors (Å²)**							

Protein	63.53	71.38	69.86	68.31	55.44	90.13	85.10
Water	57.63	64.44	58.02	48.41	50.47	84.61	76.27
RMS deviations‡							
Bond lengths (Å)	0.008	0.008	0.007	0.008	0.009	0.008	0.008
Bond angles (°)	1.15	0.93	1.10	1.15	1.21	0.90	0.88
Ramachandran parameters							
Favoured (%)	98.62	96.65	97.61	97.24	97.61	96.65	97.62
Allowed (%)	1.38	2.87	1.91	2.30	1.91	2.87	1.90
Outliers (%)	0.0	0.48	0.48	0.46	0.48	0.48	0.48

*Values in brackets refer to the highest resolution shell. † $CC_{1/2}$ correlation coefficient ‡ RMS – root mean square deviation

5.5.2 Investigation of mutant thermal stability

Further studies to confirm the overall stability and folding of the mutants - especially those with unsolved structures - used a thermal shift assay to investigate the trends in melting temperatures (T_M). All mutants were subjected to differential scanning fluorimetry in the storage buffer and the results were analysed in GraphPad Prism 9.4.1 (Figure 5.15 and Table 5.4). It should be highlighted that a thermal profile could not be generated for Y55F+Y189F and this, alongside the lack of structure, indicated that Y55F+Y189F may be unfolded or highly unstable. This would also aid our understanding as to why no cHP was formed from this particular mutant. It should be highlighted that this hypothesis could be confirmed by performing circular dichroism where the secondary structural features could be identified by measuring the absorption different between left and right circularly polarised light[175]. On the other hand, the remaining mutants exhibited similar melting profiles to the WT with some improved melting temperatures, especially for the double mutants. Mutants with expanded substrate specificity such as Y55V+E174L or E174L+Y189L also displayed enhanced thermal stability with melting temperatures at least 10 degrees higher than the wildtype. This finding was not expected and has not been mentioned in the literature where other CDPSs have been subjected to mutagenesis investigations. It appears that there may be potential to alter both the substrate specificity and the overall thermal stability of this family of proteins by targeting a select few residues within the P1 region. Only three mutants presented a T_M lower than the wild-type - S26A, S26C and Y55V – all of which had reduced thermal stability by 1 °C. However, the standard deviation from these sets of triplicates measurements highlights that this difference could be attributed simply to error in the measurement. Additionally, this T_M difference does not appear to affect the activity of Y55V which remarkably showed a change in substrate tolerance, generating both cFP and cLP (Figure 5.12).

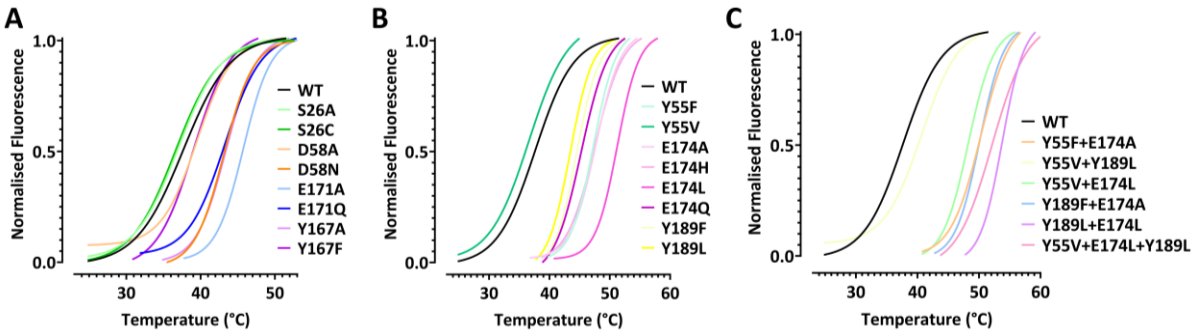


Figure 5.15 Differential scanning fluorimetry of *Parcu*CDPS mutants.
All mutants were subjected to the same thermal stability test in triplicate. Each graph contains a select group of mutants: **A** - active site; **B** - P1 and **C** - double/triple mutants. Using GraphPad Prism 9.4.1, a sigmoidal fit of the normalised mean from the 3 replicates was performed for each protein.

Table 5.4 Melting temperatures (T_M) of each *Parcu*CDPS variant.
Melting temperatures were calculated as described in as calculated from the thermal shift assay (Section 2.7.4).

Enzyme variant	T_M (°C)	Standard deviation of T_M (°C)	Δ T_M (°C)
WT	37.65	0.98	
S26A	36.98	1.54	-0.67
S26C	36.58	1.25	-1.07
Y55F	47.31	1.08	9.66
Y55V	36.59	0.77	-1.06
D58A	39.50	0.54	1.85
D58N	43.17	0.42	5.52
E171A	45.75	0.34	8.10
E171Q	43.25	1.17	5.60
E174A	47.70	0.40	10.05
E174H	47.53	0.38	9.88
E174L	51.25	0.31	13.6
E174Q	45.25	1.39	7.60
Y167A	43.47	0.30	5.82
Y167F	39.04	1.50	1.39
Y189F	44.54	1.01	6.89
Y189L	43.44	0.74	5.79
Y55F+E174A	50.27	0.60	12.62
Y55V+Y189L	40.68	0.44	3.03
Y55V+ E174L	48.27	0.38	10.62
Y189F+E174A	50.12	0.71	12.47
Y189L+E174L	53.79	0.14	16.14
Y55V+E174L+Y189L	52.14	2.41	14.49

5.6 Concluding observations

The structures of six cyclodipeptide synthases have previously been published, all of which reflect the high structural similarity present within the family. However, attempts at using these as molecular replacement models to solve *Parcu*CDPS were unsuccessful. This was the first indication that *Parcu*CDPS was unique in more than one aspect. Whilst the wildtype structure did contain the archetypical CDPS Rossman fold and two pocket structures, divergent features were highlighted including a large flexible alpha helix and a distinctive change in direction of an integral beta sheet. This investigation confirmed the enzyme active site did, in fact, remain the same suggesting that *Parcu*CDPS would participate in the same mechanism described for *Snou*-CDPS. The two pocket features contained the same residues as predicted by Gondry *et al.* but the particular composition of

amino acids that constitute these pockets remain uncommon amongst the CDPS family, further demonstrating the unique nature of *Parcu*CDPS.

Previous research confirming the substrate binding order allowed for the manipulation of P1 which initially destroyed the ability of *Parcu*CDPS variants to recognise histidine. An intriguing hypothesis then developed focussing on the identity of the substrate which suggested that perhaps the enzyme itself was not completely inactive, but simply inactive towards histidine. Therefore, alternative substrates were tested consequently yielding both cFP and cLP from a series of single and double mutant *Parcu*CDPS. Substrate selectivity itself was not completely disrupted as structurally similar amino acids such as isoleucine and valine were tested but not accepted by the variants thus clearly demonstrating that the enzymes were still able to consciously select for their substrates. Overall, the research presented here is the first record of successful engineering of a CDPS enzyme to yield different peptide products.

Chapter 6 - Functional characterisation of *Ndas*CDO

6.1 Introduction

Cyclodipeptide biosynthetic pathways have received considerable interest as it has been proposed they could expand the enzymatic toolkit to synthesise novel cyclodipeptides with attractive properties. CDPS gene clusters are classed as one of the smallest known natural product pathway with the largest reported cluster being 10 kB in total[90]. To date, CDPSs have mainly been identified from bacterial phyla with at least one predicted tailoring enzyme found in all bacterial genomes which contain a CDPS-encoding gene[176]. Whilst the most commonly found tailoring enzymes were N-acetyltransferases and sulfotransferases, global genomic analysis of 721 unique CDP gene clusters highlighted 42 of these had cyclodipeptide oxidase (CDO) genes associated with them[89,94]. The incorporation of a double bond has been shown to be quite effective at improving the therapeutic properties of certain compounds. For example, Kanzaki *et al.* reported a remarkable enhancement in inhibitor activity by 1000x for the tetra-dehydrogenated phenylahistin when compared to the naturally produced di-dehydrogenated natural product[177]. This discovery subsequently led to the development of plinabulin (NPI-2358), an effective anti-tumour therapeutic currently in stage III clinical trials for non-small cell lung cancer[38].

The earliest report of an enzyme installing an *exo*-double bond on a CDP came from a cell free extract of *Streptomyces albulus* KO-23 in 1999[178], however it was not until 2001 that this class of enzymes were categorised as cyclodipeptide oxidases (CDOs). The first isolation of this novel enzyme demonstrated the ability to convert cyclo(L-Phe-L-Leu) into the doubly oxidised version cyclo(L-Phe*-L-Leu*), where the asterisk represents dehydrogenation of the Cα-Cβ bond (Figure 6.1)[91,179]. Most recently, Jiang *et al.* described a CDO involved in the biosynthesis of drimentine B (DMT B) from *Streptomyces sp. NRRL F-5123*, bringing the total number of cyclodipeptide oxidases found in natural product biosynthetic pathways to six, as of January 2023 (Table 6.1)[180]. Interestingly, the majority of known CDOs have been discovered in the *Streptomyces* genus which is well known for producing molecules with unique activities[181].

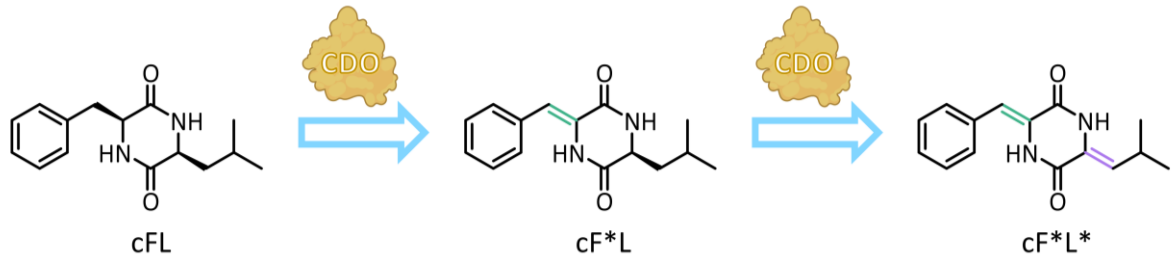


Figure 6.1 CDO-catalysed oxidation of a CDP.
As an example, the CDO isolated from *Streptomyces noursei* subsequently oxidises cyclo(L-Phe-L-Leu) to form two products: cyclo(L-Phe*-L-Leu) first and cyclo(L-Phe*-L-Leu*) where the asterisk represents oxidation of the C-C single bond. The first oxidised bond is shown in green and the second in purple.

Table 6.1 Summary of CDOs found in natural product biosynthetic pathways.
The peptide(s) in red denote the natural CDP produced during the biosynthetic pathway to the final product. In the case of Ndas_1146 and Ndas_1147, only the existence and not the exact location of the double bond was confirmed, therefore the asterisk highlights only the presence in this particular case.

Biosynthetic product pathway	Organism	CDO Name	Observed products	Publication Year
Albonoursin	*Streptomyces noursei*	AlbB/AlbC	cF*L/cF*L*, cF*F* cF*H, cWW*/ cW*W*, cL*A, cF*G, cL*G, cS*G, cE*G	2001[91]
Nocazine	*Nocardiopsis dassonvillei*	Ndas_1146/ Ndas_1147	cF*Y*, cFF*, cFL* cYL*, cYY*, cFM*	2013[182]
Guanitrypmycin	*Streptomyces monomycini* NRRL B-24309	$Gut(BC)_{24309}$	cWF*, cW*Y	2019[181]
Purincyclamide	*Streptomyces chrestomyceticus* NA4264	PcmB/PcmC	cWY*	2019[93]
Guatrypmethine C	*Streptomyces cinnamoneus* DSM 40646	GtmBC	cWM*	2022[183]
Drimentine B	*Streptomyces sp.* NRRL F-512	DmtD2_E2	cWP* c*WA cWV* cWY* cWL*	2023[180]

An established feature of a CDO is the presence of two subunits which form a multimeric complex with a covalently associated flavin mononucleotide (FMN) cofactor as proven by denaturation experiments[91]. Oxygen was shown to be the electron acceptor with the formation of hydrogen peroxide as a by-product as reactions under anaerobic conditions were ineffective at yielding oxidised CDPs[91]. However, the exact mechanism of action is yet to be uncovered with some hypothesizing that

oxidation could occur via one of three possible routes: direct dehydrogenation; α-hydroxylation with subsequent water loss or imine formation with enamine rearrangement (Figure 6.2)[184]. When CDOs turn over a cyclodipeptide with two oxidisable bonds, it has been postulated that the CDP will dissociate following one round of oxidation and then re-bind to the enzyme, however there is no direct evidence of this currently, therefore Figure 6.2 only demonstrates potential mechanisms for one oxidative cycle.

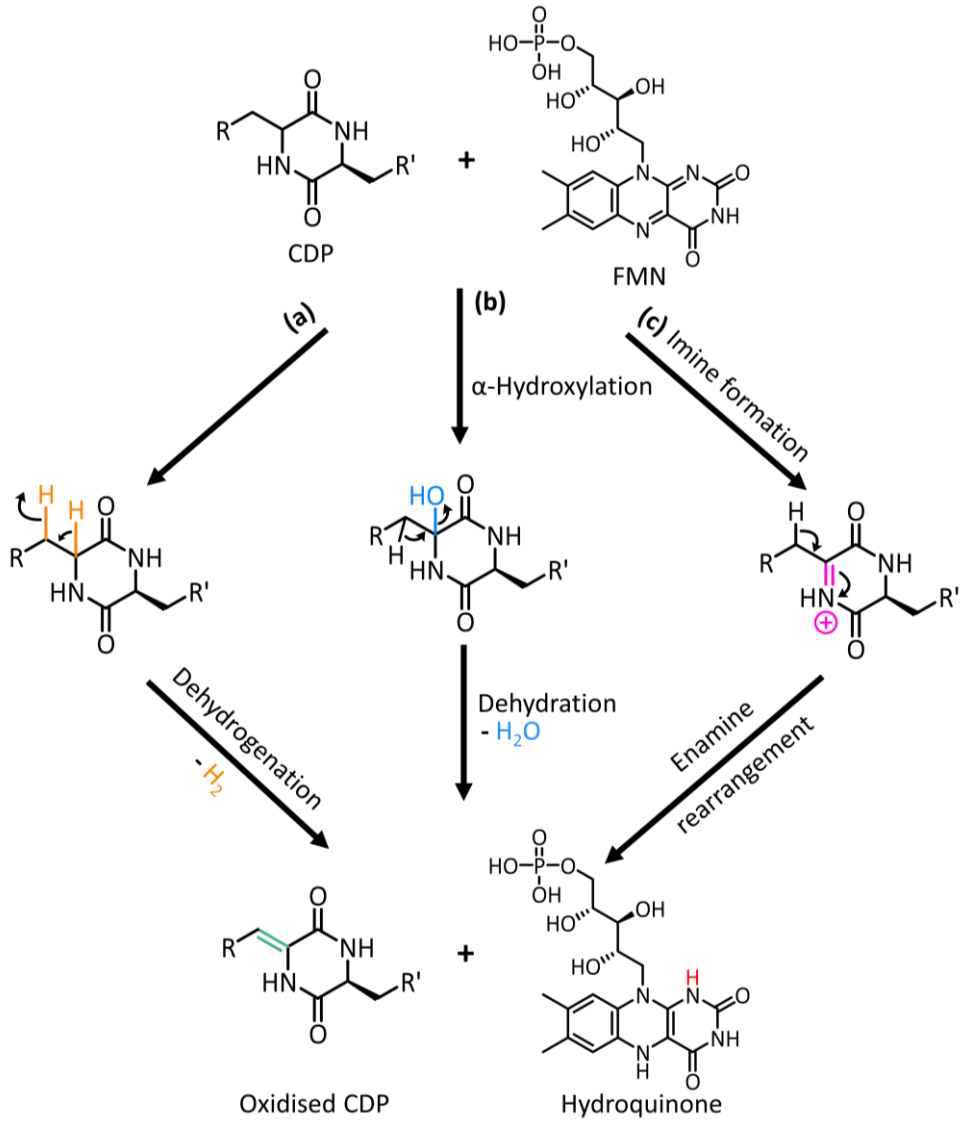


Figure 6.2 Proposed mechanistic pathways to cyclodipeptide oxidation.
Three potential options for oxidation by CDOs have been proposed in the literature but so far there are no experimental data to prove or disprove any of the mechanisms shown. Option (a) uses dehydrogenation of the α and β hydrogens of the side chain; (b) requires first hydroxylation at the α carbon followed by the loss of a water molecule and (c) involves an imine intermediate which is then rearranged to the corresponding enamine. The role of the flavin cofactor in these reactions is yet to be determined but is thought to cycle between oxidised and reduced (hydroquinone) throughout.

Initially, it was assumed that oxidation occurred using only the larger of the two gene products from the biosynthetic gene cluster which displayed close resemblance to the nitroreductase family.

Yet, dehydrogenation attempts with only this subunit were unsuccessful leading to the inclusion of the second smaller subunit which has no similarity to proteins of known structure or function[179]. Interestingly, the CDO genes have overlapping DNA sequences resulting in the start codon for the second subunit to be located far from the promoter sequence. This has been hypothesised to regulate gene expression during translation and thus may be essential for ensuring protein-protein interactions between the two subunits to allow the multimeric complex to successfully form[182]. With no crystal structure solved or biophysical characterisation of the CDO complex, the only information gathered on the oligomeric state of the complex was from size exclusion chromatography. The first chromatographic experiment in 2001 highlighted that the CDO from *Streptomyces noursei* (AlbA/AlbB) was larger than 2 mega Daltons as the sample eluted in the void of a Superose 6 column. Successive experiments indicated that the complex from *Nocardiopsis dassonvillei* (Ndas_1146/47) was in fact 4 mega Daltons, thus corroborating that this family of enzymes appears to exist as large multimeric species despite each subunit only being around 22 kDa and 11 kDa with no data known of how these two interact to form the overall active complex[182].

Factors underlying the substrate specificity exhibited by this enzyme class remain undetermined, but early investigations highlighted the ability of CDOs to oxidise alternative peptides not found in their natural biosynthetic pathway. Using a cell free extract from *Streptomyces albulus* KO-23, several phenylalanine containing CDPs were dehydrogenated, even including a derivatised version containing an isoprenyl group on the histidine moiety[177,178]. A more recent extensive analysis by Mikulski *et al.* compared the substrate specificity of three CDOs - AlbA/B, Ndas_1146/7 and CDO-Np from *Nocardiopsis prasina* which had never before been studied[185]. These experiments using 34 different cyclic dipeptides proved that CDOs could tolerate a wide array of substrates yet displayed distinct preference for phenylalanyl- or tyrosyl-containing CDPs. Furthermore, an intriguing study unveiled that CDOs could work in tandem with CDPSs from differing organisms if both were expressed in a suitable host. Using eighteen CDPSs with eight CDOs, conversion of cyclodipeptides into their oxidised derivatives was detected for 14 of the 27 different peptides[94]. Again, these CDOs demonstrated a clear bias towards aromatic or hydrophobic chains including tryptophan and leucine this time. From this, the idea of classifying CDOs based on their substrate tolerance was formulated using the terms 'broad-spectrum' and 'narrow-spectrum'[94]. Not only did broad-spectrum CDOs accept a greater variance of CDPs, they were also capable of performing double oxidation on CDPs to yield the final tetra-dehydrogenated form.

Our overall understanding of cyclodipeptide oxidases is therefore limited, especially given that most publications to date used *in vivo* expression systems or crude cell extracts to catalyse the oxidation of CDPs. Whilst it is accepted that this family exists as a large complex, information about

the nature of the subunits or the exact role of the flavin cofactor within the mechanism is lacking. The homology of subunit A to the nitroreductase family was initially used for comparison as it was discovered that AlbA contained the same residues responsible for FMN binding as other flavin-binding nitroreductases[179]. However, these nitroreductases are non-covalently associated with flavin and require NADPH as electron donors, whereas CDOs are understood to covalently bind FMN and not require NADPH for activity. Generally, studies to date failed to unveil the intricacies of how these enzymes are actively recognising substrates and discriminating against structurally similar peptides. Ergo, this elusive enzyme family would benefit from a global investigation of both the functional and structural characteristics to allow us to fully realise the potential expansion to the chemical realm of achievable CDP structures via enzymatic pathways.

Initially, tailoring enzymes were sought from the organisms of the two histidine accepting CDPSs characterised in Chapter 3. PRISM, developed by Skinnider *et al.*[186–189], is an algorithm to identify potential CDPS biosynthetic gene clusters, however this did not highlight the existence of any putative tailoring enzymes in either of the genomes containing *Para*CDPS or *Parcu*CDPS. AntiSMASH (Version 7.0.0) analysis also supported this finding, hence an alternative CDO was sought which would be likely to accept histidine as a substrate[190–195]. The comparative study of three CDOs by Mikulski and colleagues used four histidine containing CDPs including the two products from the CDPSs investigated in this project. Thus, we focussed on the cyclodipeptide oxidase from *Nocardiopsis dasonvillei* DSM 43111 (NC_014210.1), henceforth referred to as *Ndas*CDO, whose natural substrate is cyclo(L-Trp-L-Phe) (cWF).

Purification of *Ndas*CDO was key to in-depth characterisation of this enzyme therefore preliminary research focussed on isolating this protein from recombinant expression in *E. coli*. The substrate tolerance was investigated using LC-MS *in vitro* activity assays which was complemented further with kinetic analysis to reveal details about the chemical steps shown in Figure 6.3. Only the steps represented by the kinetic parameters specifically for the first oxidative cycle are shown for simplicity, however CDPs with two oxidisable bonds are likely to exhibit two sets of parameters relating to the first and second rounds of oxidation respectively. Lastly, the existence of a flavin radical intermediate was explored using electron paramagnetic resonance spectroscopy in collaboration with Dr Bela Bode (University of St Andrews). These preliminary results emphasised the lack of key facts available about cyclodipeptide oxidases and strengthen the argument that this family needs further exploration to fully harness their potential.

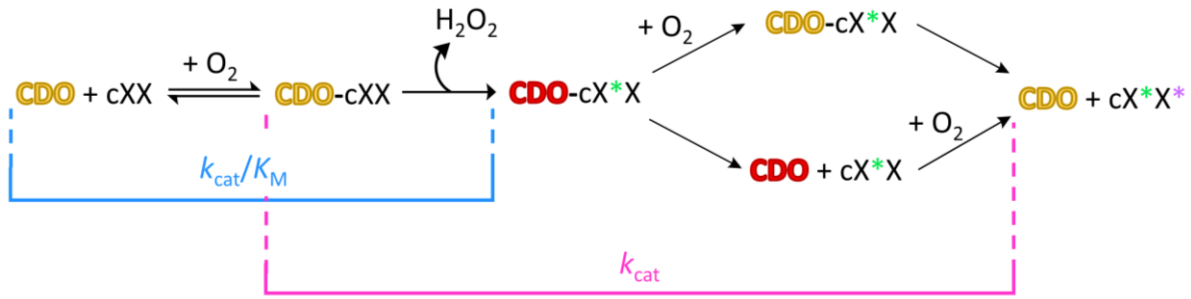


Figure 6.3 Hypothetical reaction scheme followed by cyclodipeptide oxidases.
The colour of the CDO relates to the state of the flavin cofactor: orange for oxidised FMN and red for reduced FMN. The presence of a double bond on the cyclodipeptide, cXX, is implied by an asterisk (green or purple). The macroscopic steps that k_{cat} and k_{cat}/K_M report on are depicted by the pink and blue brackets respectively.

6.2 Functional characterisation of *Ndas*CDO

6.2.1 Expression and purification of *Ndas*CDO

Initially, a gBlock™ featuring the DNA sequence encoding both subunits of *Ndas*CDO – *Ndas*1146 and *Ndas*1147 – was purchased alongside the relevant primers to perform Gibson Assembly using a pDUET vector (a gift from Dr Haunting Liu, University of St Andrews). Amplification of both the genes and the plasmid via a PCR was successful, however Sanger sequencing highlighted that genes were not properly cloned into the pDUET vector despite multiple attempts of Gibson Assembly. This failure was unexpected but could be attributed to the primer design for the cloning of the gBlock as the two genes contain some overlap where a primer could non-specifically bind to. Due to time constraints, a plasmid – pRSF-DUET *Ndas*CDO – was purchased from Genscript which contained both subunits of NdasCDO in the vector where only subunit A had a C-terminus 6xHis tag facilitating the co-expression of both subunits and subsequent purification if they formed a complex as previously thought (Appendix VI for plasmid map).

The small scale expression trial (method details in Section 2.8.2) highlighted the presence of the larger CDO subunit, *Ndas*1146 (22 kDa), in the soluble protein fraction which was then not observed post-elution from the nickel affinity beads using the Biosprint (Figure 6.4A). Expression was further investigated by purifying protein from 3 L of cell culture using nickel-affinity chromatography whereupon *Ndas*CDO was successfully purified from *E. coli*, producing a yellow protein (Figure 6.4B). SDS-PAGE analysis of *Ndas*CDO post-nickel revealed a band from the elution profile relating to the molecular weight expected for the larger subunit (22 kDa) however there was no indication of elution of the second smaller subunit (11 kDa) (Figure 6.4C). Afterward, it was realised the smaller subunit of *Ndas*CDO contained no tryptophan residues, explaining why there was no band in the stain-free gel used for analysis. Size exclusion chromatography, however, was unsuccessful using a HiLoad® 16/600

Superdex® 200 pg whereupon *Ndas*CDO did not elute from the column, indicative of possibly instability or aggregation in the current buffer conditions (Figure 6.4D). To confirm the presence of both subunits, mass spectrometric analysis of the protein was conducted. A trypsin digest confirmed that the series of observed peptide fragments displayed the highest similarity to sequences of both subunits from *Ndas*CDO (Appendix V). This technique coupled with SDS-PAGE analysis did not confirm the final oligomeric state of the *Ndas*CDO complex unfortunately, therefore alternative techniques should be investigated such as Native PAGE analysis. Theoretically, Native PAGE does not denature the protein therefore the complex should remain intact as it runs on the gel, thus alluding to the potential size of this protein.

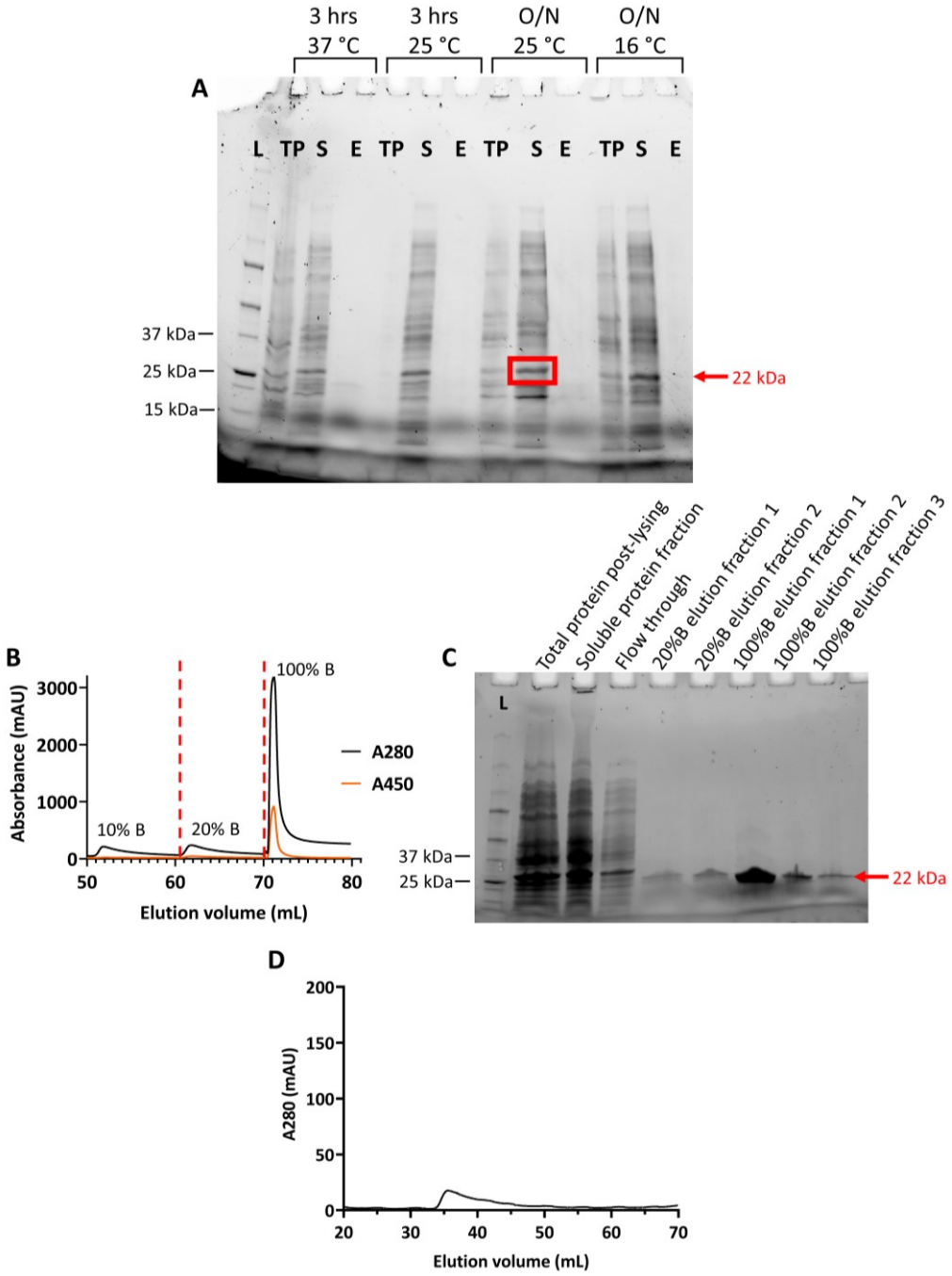
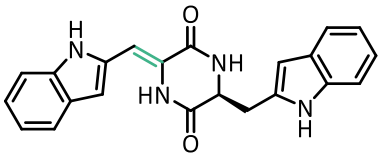
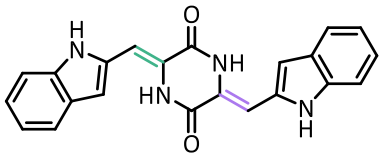
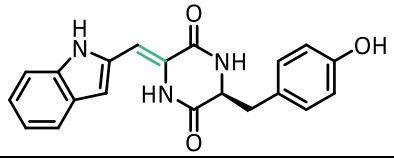
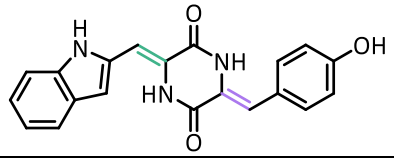
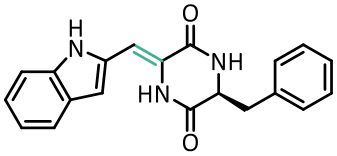
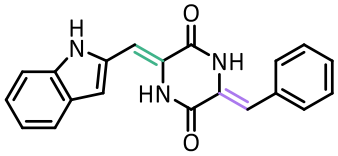
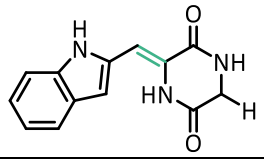
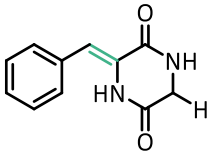
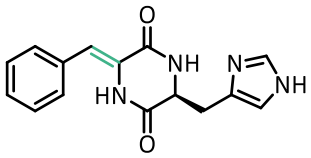
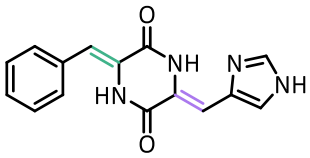
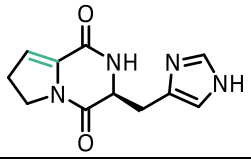
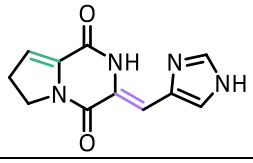
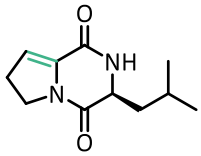
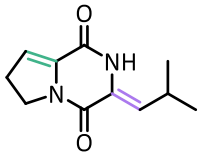
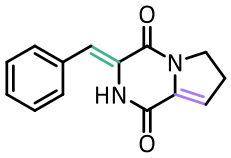


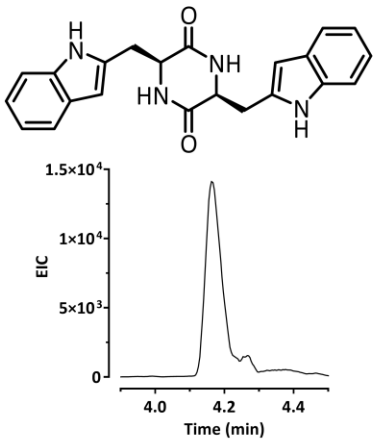
Figure 6.4 Recombinant expression of *Ndas*CDO.
A – SDS-PAGE analysis of small scale expression trials of *Ndas*CDO where growth conditions were varied as noted above the respective gel lanes. The labelling of gel lanes are as follows: L – protein standards ladder; TP – total protein; S – soluble fraction after cell lysis and E – elution from nickel affinity beads. **B** – UV absorbance trace of nickel-affinity chromatography of *Ndas*CDO where the imidazole concentration was increased to elute the protein from the column. The red dotted lines highlight the regions where the imidazole concentration changed, starting at 10% B (30 mM imidazole) and increasing to 20% (60 mM,) and finally 100% (300 mM). *Ndas*CDO is known to have an FMN cofactor, thus the wavelength at 450 nm was also monitored to highlight the presence of this molecule. **C** – SDS-PAGE analysis of fractions from the nickel affinity chromatography. The expected molecular weights of *Ndas*1146 and *Ndas*1147 are 22 kDa and 11 kDa respectively. **D** – UV absorbance trace of size exclusion chromatography of *Ndas*CDO using a Superdex 200 16/600 column. The protein sample did not elute from this column highlighting possible stability issues of the complex.

6.2.2 Confirmation of *Ndas*CDO activity using mass spectrometry

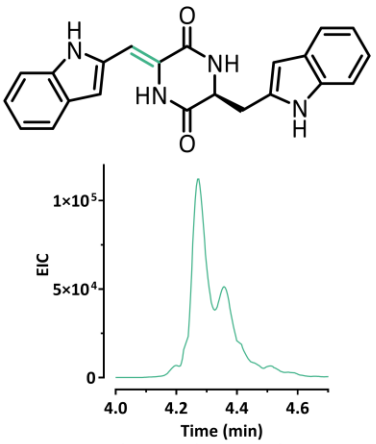
The enzymatic activity of *Ndas*CDO was investigated using a similar *in vitro* style assay developed for CDPSs, where the products were observed via LC-MS (Section 2.8.4) . A range of commercial cyclodipeptides were tested whereupon a partition of two possible oxidation states was possible. For cyclo(L-Phe-L-Pro), only the tetra-dehydrogenated form was detected whilst all other peptides contained a mixture of di- and tetra-dehydrogenated (Table 6.2). The extracted ion chromatograms from the MS analysis of each product are shown in Figure 6.5. Work by other groups had previously reported the acceptance of cFY, cFF, cFL, cYL, cYY and cFM by *Ndas*CDO which were therefore not tested here[182]. It was evident that *Ndas*CDO was particularly active with tryptophan containing CDPs which was to be expected as this enzyme is naturally part of the nocazine biosynthetic pathway[182]. This particular assay, however, did have some limitations - without a calibration curve, quantification of the oxidised product yields was not possible and given that each product would not ionise to the same extent, peak areas could not be directly compared. Moreover, for CDPs with two oxidisable bonds, mass spectrometric analysis did not indicate the position of the double bond after the first oxidative step, but rather simply the presence of one. With reference to Table 6.2, the first double bond (drawn in green) has been placed on the left amino acid consistently for simplicity, but it is entirely possible that the double bond could be on the right amino acid if that bond is also a carbon-carbon single bond. HPLC analysis of these reactions would be useful in future to determine the overall conversion of substrate to product, however this is only applicable to cyclic dipeptides containing UV active side groups. The presence of two oxidisable bonds on the CDP ring could lead to a mixture of two di-dehydrogenated products (one double bond on either amino acid) as well as the tetra-dehydrogenated product which would further complicate the separation and analysis of these molecules. NMR analysis could be of use here to determine the location of the single double bond by investigating the environments experienced by the hydrogen nuclei within the CDP which would be perturbed by the presence of a double bond. Alternatively, previous research has shown that MS-induced fragmentation of the 2,5-diektopiperazine ring could lead to a fingerprint signature unique to the CDP substrate or oxidised product which could be interpreted to identify the location of the double bond[185]. Moreover, these experiments could elucidate the prerequisites required by the enzyme for the installation of the first double bond which could be related to the currently unknown reaction mechanism.

Table 6.2 Structures of oxidised CDPs observed after incubation with *Ndas*CDO *in vitro*.
The position of the first double bond is unknown therefore it has been drawn on the left amino acid for simplicity.

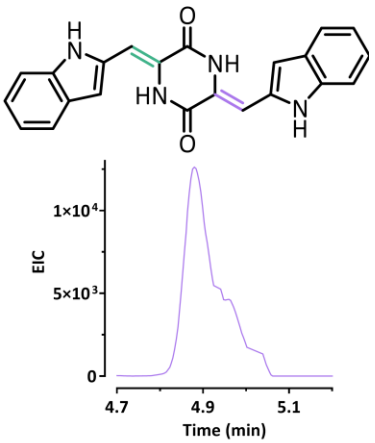
Cyclodipeptide	1st oxidation cycle	2nd oxidation cycle
Cyclo(L-Trp-L-Trp)		
Cyclo(L-Trp-L-Tyr)		
Cyclo(L-Trp-L-Phe)		
Cyclo(L-Trp-L-Gly)		**Not applicable**
Cyclo(L-Phe-L-Gly)		**Not applicable**
Cyclo(L-His-L-Phe)		
Cyclo(L-His-L-Pro)		
Cyclo(L-Leu-L-Pro)		
Cyclo(L-Phe-L-Pro)	**Not observed**	



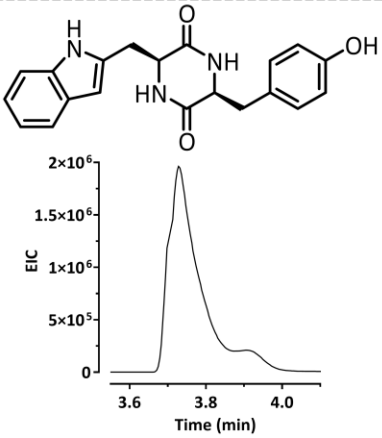
Cyclo(L-Trp-L-Trp)
Expected m/z: 373.1659
Observed m/z: 373.1656
Ppm deviation: -0.8039



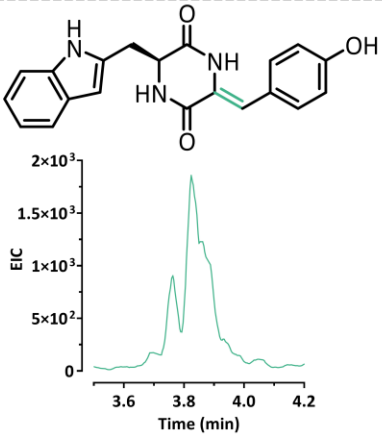
Cyclo(L-Trp*-L-Trp)
Expected m/z: 371.1503
Observed m/z: 371.1500
Ppm deviation: -0.8083



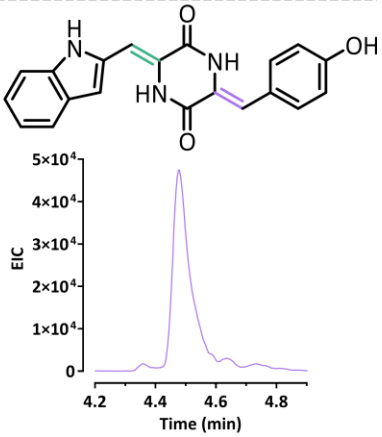
Cyclo(L-Trp*-L-Trp*)
Expected m/z: 369.1346
Observed m/z: 369.1346
Ppm deviation: 0.0



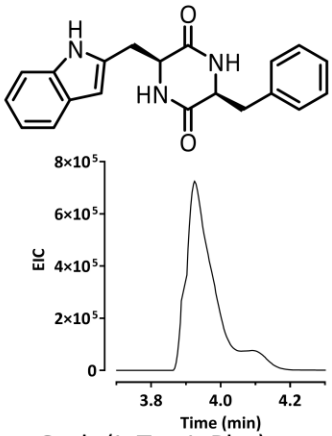
Cyclo(L-Trp-L-Tyr)
Expected m/z: 350.1499
Observed m/z: 350.1506
Ppm deviation: 1.9991



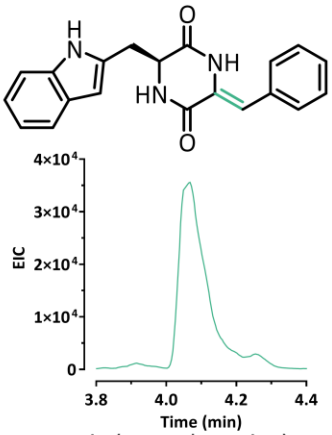
Cyclo(L-Trp*-L-Tyr)
Expected m/z: 348.1343
Observed m/z: 348.1351
Ppm deviation: 2.2979



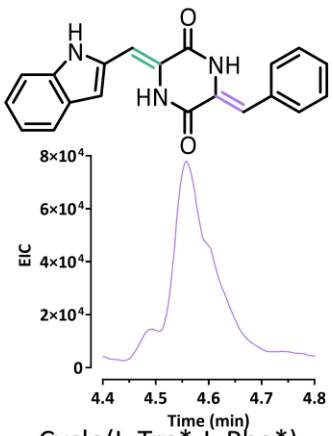
Cyclo(L-Trp*-L-Tyr*)
Expected m/z: 346.1186
Observed m/z: 346.1185
Ppm deviation: -0.2889



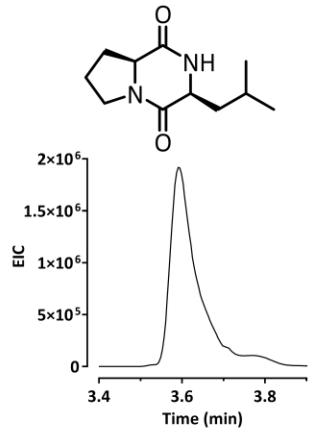
Cyclo(L-Trp-L-Phe)
Expected m/z: 334.1550
Observed m/z: 334.1554
Ppm deviation: 1.1971



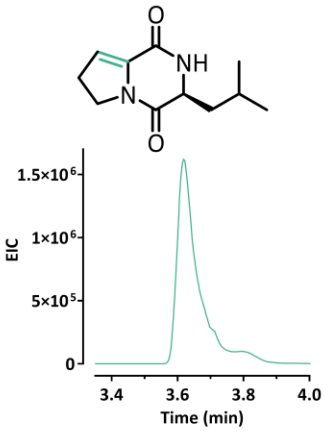
Cyclo(L-Trp*-L-Phe)
Expected m/z: 332.1394
Observed m/z: 332.1395
Ppm deviation: 0.3011



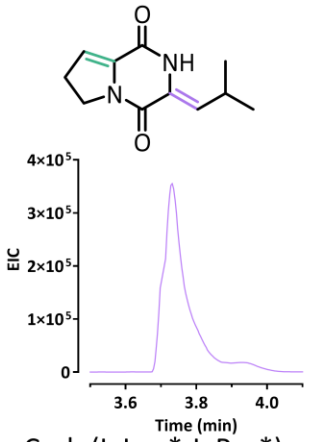
Cyclo(L-Trp*-L-Phe*)
Expected m/z: 330.1237
Observed m/z: 330.1235
Ppm deviation: -0.6058



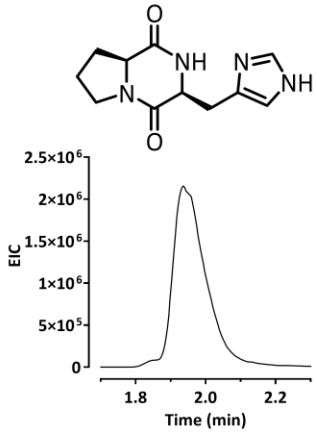
Cyclo(L-Leu-L-Pro)
Expected m/z: 211.1441
Observed m/z: 211.1442
Ppm deviation: 0.4736



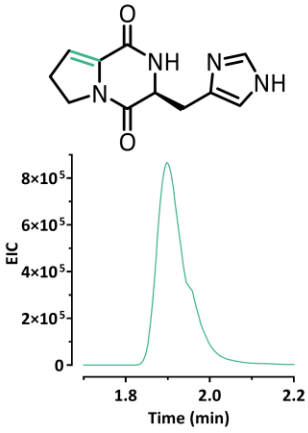
Cyclo(L-Leu-L-Pro*)
Expected m/z: 209.1285
Observed m/z: 209.1287
Ppm deviation: 0.9563



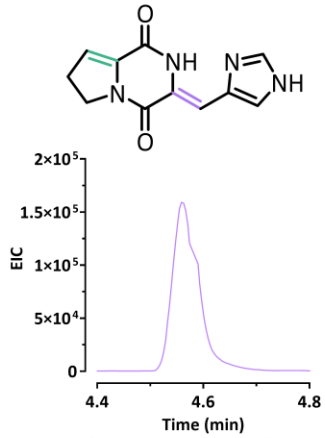
Cyclo(L-Leu*-L-Pro*)
Expected m/z: 207.1128
Observed m/z: 207.1131
Ppm deviation: 1.4484



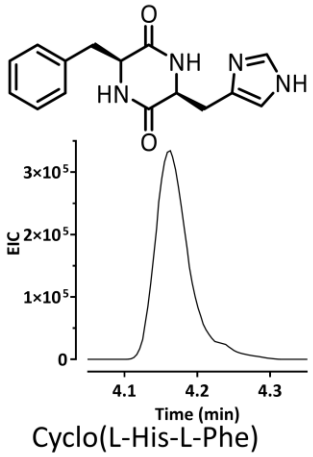
Cyclo(L-His-L-Pro)
Expected m/z: 235.1190
Observed m/z: 235.1195
Ppm deviation: 2.1265



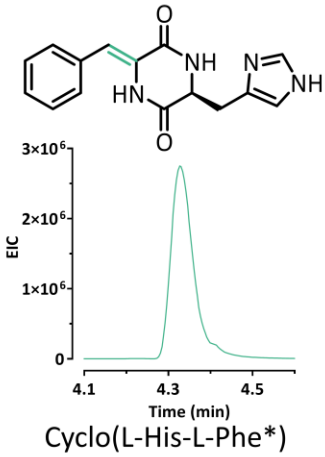
Cyclo(L-His-L-Pro*)
Expected m/z: 233.1033
Observed m/z: 233.1032
Ppm deviation: -0.4289



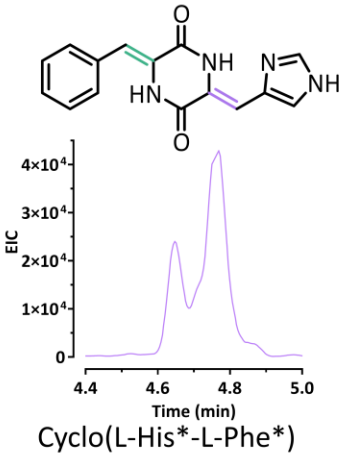
Cyclo(L-His*-L-Pro*)
Expected m/z: 231.0877
Observed m/z: 231.0876
Ppm deviation: -0.4327



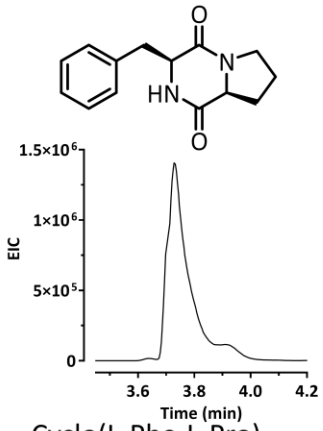
Cyclo(L-His-L-Phe)
Expected m/z: 285.1346
Observed m/z: 285.1347
Ppm deviation: 0.3507



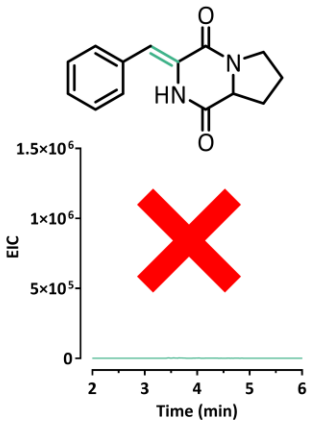
Cyclo(L-His-L-Phe*)
Expected m/z: 283.1190
Observed m/z: 283.1194
Ppm deviation: 1.4128

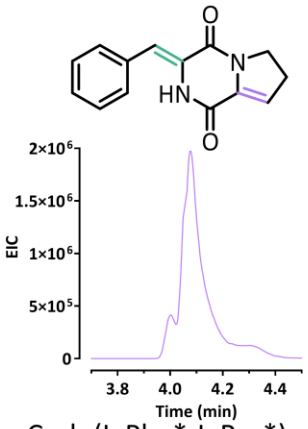


Cyclo(L-His*-L-Phe*)
Expected m/z: 281.1033
Observed m/z: 281.1033
Ppm deviation: 0.0

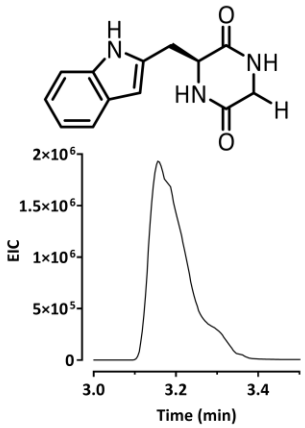


Cyclo(L-Phe-L-Pro)
Expected m/z: 245.1285
Observed m/z: 245.1288
Ppm deviation: 1.2238

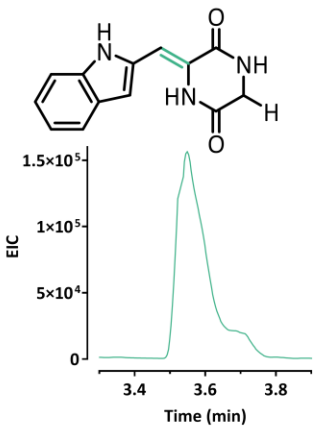




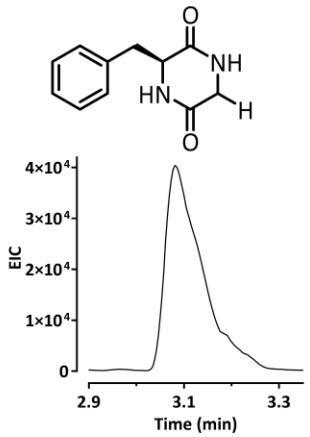
Cyclo(L-Phe*-L-Pro*)
Expected m/z: 241.0972
Observed m/z: 241.0975
Ppm deviation: 1.2443



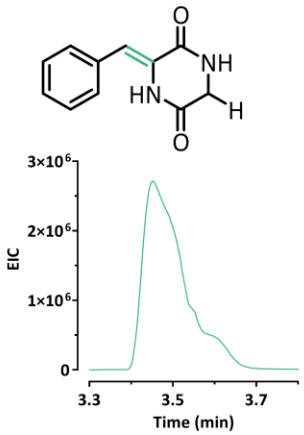
Cyclo(L-Trp-L-Gly)
Expected m/z: 244.1081
Observed m/z: 244.1085
Ppm deviation: 1.6386



Cyclo(L-Trp*-L-Gly)
Expected m/z: 242.0924
Observed m/z: 242.0923
Ppm deviation: -0.4131



Cyclo(L-Phe-L-Gly)
Expected m/z: 205.0972
Observed m/z: 205.0973
Ppm deviation: 0.4875



Cyclo(L-Phe*-L-Gly)
Expected m/z: 203.0815
Observed m/z: 203.0823
Ppm deviation: 3.9393

Figure 6.5 LC-MS analysis of oxidised CDPs.
Extracted ion chromatograms are shown for each oxidised CDP along with the observed m/z value and ppm deviation. All observed products have m/z values within the allowed ± 5 ppm deviation.

6.2.3 Investigation of *Ndas*CDO activity using UV absorbance

With the aim of investigating key kinetic details about CDOs, it was beneficial to utilise a direct, UV-dependent assay allowing for quick read-out measurements with reliable signal amongst replicates. Previous literature investigating the catalytic efficiency of a CDO from *Streptomyces noursei* measured the increase in absorbance at 297 nm when catalysing the oxidation of cHF[91]. To verify this wavelength was suitable to monitor the reaction for the available CDPs, a UV scan was performed at the beginning of the reaction and then again after 24 hours of incubation (Figure 6.6). As expected, CDPs featuring UV-active side chains displayed the greatest change in absorbance in the 200-400 nm range, with detector saturation observed in both cFP and cHF. The overall absorbance was consistently greater in scans of reaction after 24 hours indicating the presence of a new species which can absorb UV light, however detector saturation was prevalent throughout this experiment. Additionally, for cHP and cLP, an increase in absorbance was observed at shorter wavelengths (around 260 nm) despite there being no obvious UV active group.

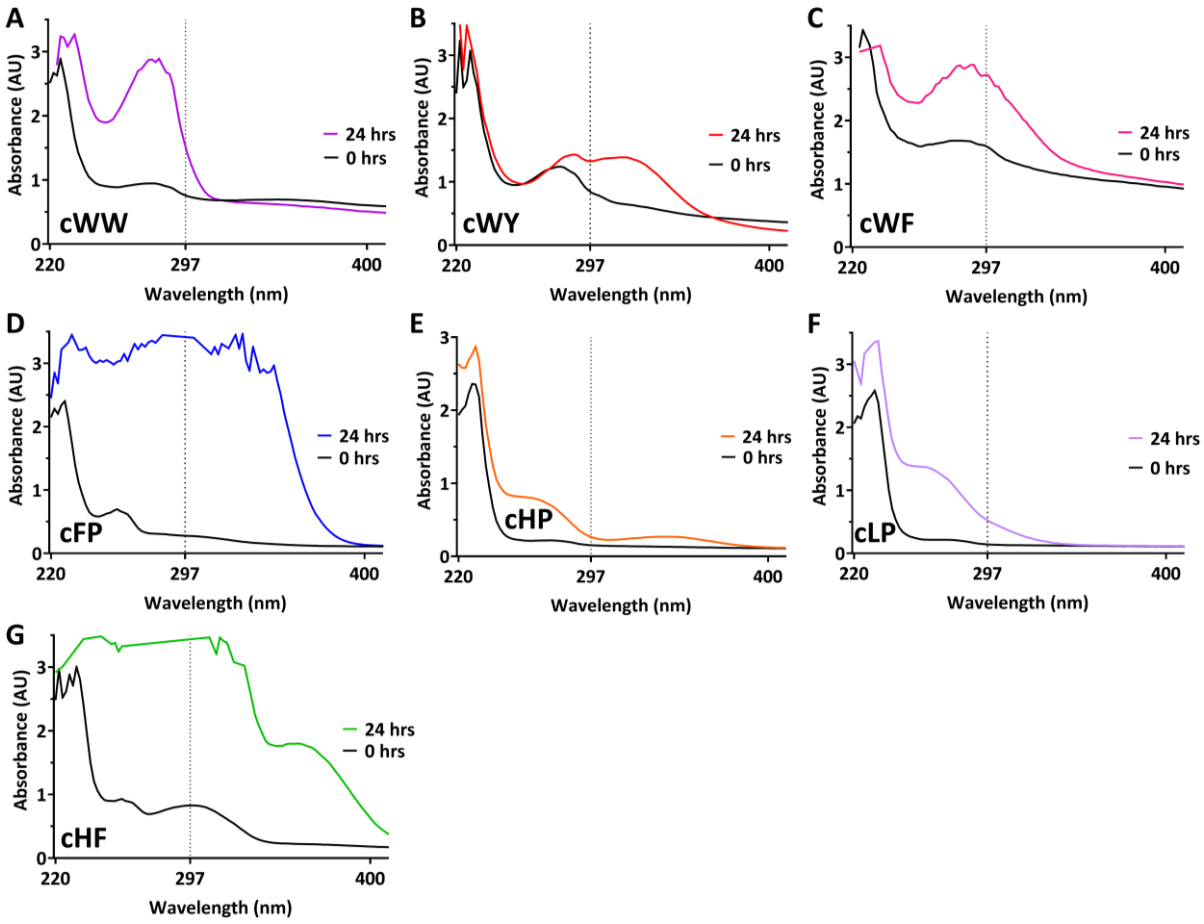


Figure 6.6 UV absorbance profiles of oxidised cyclodipeptides.
Two UV scans were taken – 0 and 24 hours – to compare the UV absorbance of the oxidised product generated by *Ndas*CDO. All available cyclodipeptides at the time were tested: **A** – cyclo(L-Trp-L-Trp) in purple; **B** – cyclo(L-Trp-L-Tyr) in red; **C** – cyclo(L-Trp-L-Phe) in pink; **D** – cyclo(L-Phe-L-Pro) in blue; **E** – cyclo(L-His-L-Pro) in orange; **F** – cyclo(L-Leu-L-Pro) in lavender, and **G** – cyclo(L-His-L-Phe) in green.

The lack of information gathered about the oxidation of these peptides at 297 nm prompted a second experiment using less CDP in an attempt to avoid detector saturation which may be masking important wavelengths for observation. A lower concentration of CDP was used this time – 200 µM for all tryptophan containing peptides and 1 mM for all other CDPs – whereas the previous experiment used 500 µM and 5 mM respectively. By reducing the substrate concentration, trends in the UV absorbance spectrum of the reaction mixture after 24 hours of incubation with *Ndas*CDO were realised (Figure 6.7). cWW, cWY, cFP and cHF all displayed a shifted absorbance maxima possibly relating to the presence of the newly form oxidised bonds which would introduce further conjugation to the ring, thus affecting the overnight absorbance wavelength and extinction coefficient. Both CDPs without a UV-active group, cHP and cLP, did not conform to this trend which is to be expected. However, an increase in absorbance was observed at 260 nm for cLP after 24 hours of reaction whilst cHP displayed little change in absorbance between the incubation times.

The extended incubation time of these reactions allows *Ndas*CDO to catalyse the formation of di and/or tetra-dehydrogenated products which would both contribute to the absorbance spectra in Figure 6.7. Therefore, the newly observed wavelengths (highlighted in Figure 6.7) cannot be definitively labelled to a single oxidised product but rather a mixture of both. Further analysis of the resultant reaction after 24 hours incubation could use LC-MS to confirm the partition of products present, however this would not stop both products from contributing the absorbance profiles. A time course experiment measuring the absorbance spectra for shorter time periods could possibly detect the presence of the first di-dehydrogenated product if it displayed a change in absorbance wavelength as well as the subsequent tetra-dehydrogenated product. LC-MS analysis of each time point should corroborate the presence of the oxidised products and allow us tot infer which are contributing to the altered UV absorbance profiles.

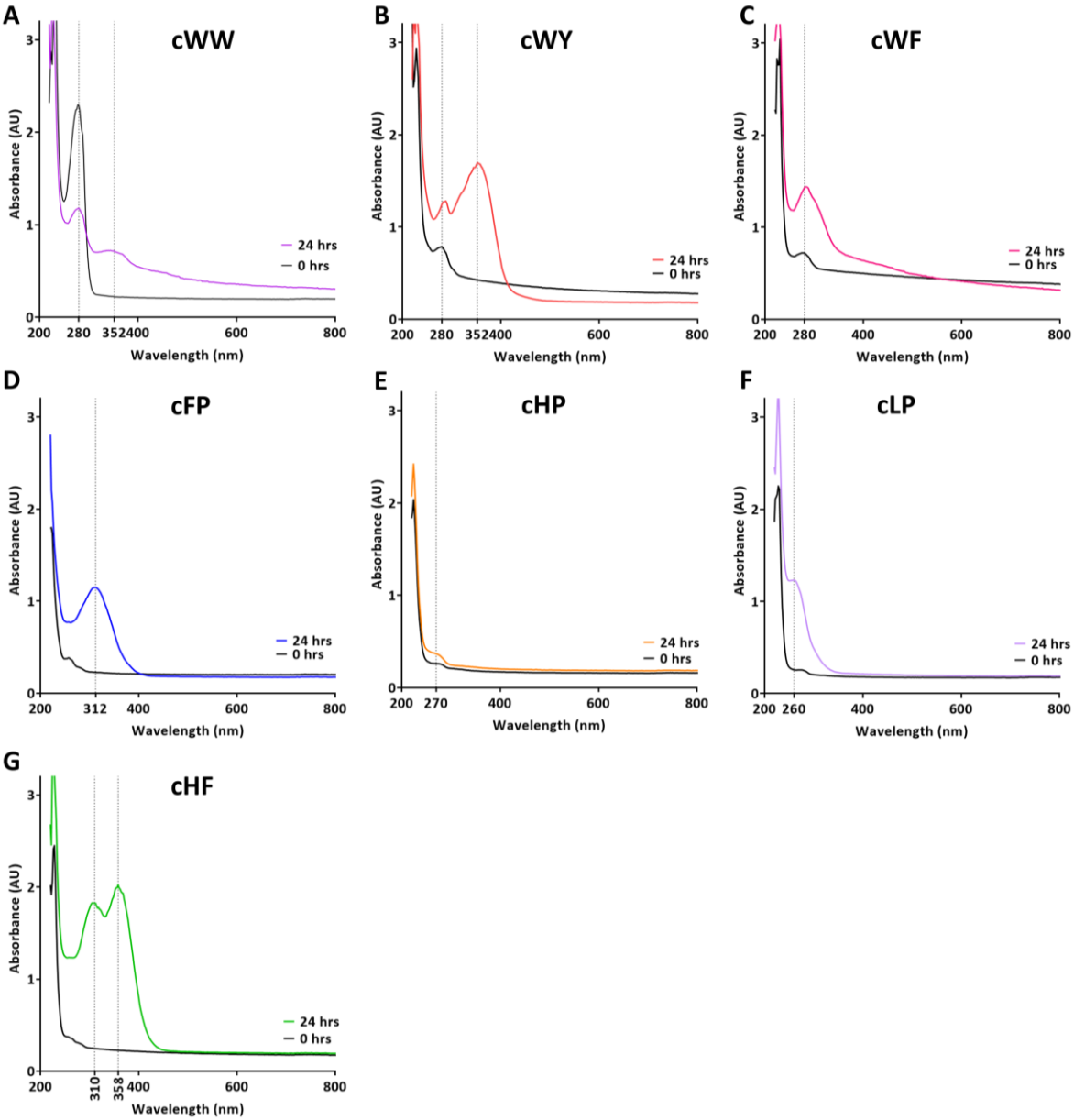


Figure 6.7 Repeat of UV absorbance profiles of oxidised cyclodipeptides.
A lower concentration CDP was used to avoid detector saturation observed in Figure 6.6. Two UV scans were taken – 0 and 24 hours – to compare the UV absorbance of the oxidised product generated by *Ndas*CDO. All available cyclodipeptides at the time were tested: **A** – cyclo(L-Trp-L-Trp) in purple; **B** – cyclo(L-Trp-L-Tyr) in red; **C** – cyclo(L-Trp-L-Phe) in pink; **D** – cyclo(L-Phe-L-Pro) in blue; **E** – cyclo(L-His-L-Pro) in orange; **F** – cyclo(L-Leu-L-Pro) in lavender, and **G** – cyclo(L-His-L-Phe) in green.

6.2.4 Influence of pH on *Ndas*CDO activity

Prior to performing a pH-rate profile, the activity of *Ndas*CDO was tested at extreme pH conditions to gather information on which values would be detrimental to the overall protein activity. Product formation was severely impacted when the pH was lower than 8 or greater than 10.5, therefore a pH-rate profile was assembled within this range by measuring the reaction rate of cFP oxidation at 297 nm under pseudo first order and steady state conditions ([S]>[E] therefore $^{d[ES]}/_{dt} = 0$). At the time

of performing this experiment, only the first attempt at obtaining the UV absorbance spectra of oxidised CDPs had been performed, therefore the monitoring wavelength was selected to be 297 nm. In hindsight it is clear that this experiment should be repeated with cFP monitoring the absorbance at 312 nm which has more accurate correlation to the oxidised product over 297 nm. Based on the LC-MS screen and the short time interval for which this assay was performed, it was assumed that carrying out the reaction under steady state conditions would lead to only one oxidised form, presumably cF*P as *Ndas*CDO is predicted to prefer aromatic side chains.

Investigations into the effect of pH on enzyme activity can allude to which groups could be participating in the acid-base catalysis of the reaction. When the plot of log(kinetic parameter) against pH displays a curvature trend, equations can be fitted to this data to describe one or more ionisable groups and obtain apparent pK_a values for each. Plotting the data for *Ndas*CDO revealed a bell shaped curve with a maxima at pH 9 for *Ndas*CDO, indicating the involvement of at least one ionisable group on each side of the acid-base equilibrium i.e. one group acting as an acid whilst the other functioning as a conjugate base (Figure 6.8). Using the model for 1 acid, 1 base catalysis, two apparent pK_a values were determined for two ionizable groups: the group with an apparent pK_a of 8.0-8.6 must be unprotonated whilst the group with a pK_a of 9.5-10.0 must be protonated. For comparison, models relating to two ionizable groups were also fitted to the data however the slope of the lines on either side of the curve did not correlate to a model containing two acids and two bases. It should be noted that typically the measured pK_a values do not describe a single residue but more likely reflect the cumulative effect from a subset of residue interactions, hence the apparent nomenclature[196].

A sequence alignment with other known flavoenzymes confirmed the conservation of residues in AlbA/B and *Ndas*CDO believed to be essential for flavin association[179]. Intriguingly, most of these residues were titratable groups such as arginine, histidine, lysine, and glutamate which could be contributing to the pH effects observed here within reason. For example, the pK_a values for histidine and glutamate are perhaps far from the range observed for *Ndas*CDO, hence it is more probable that it is contributions from arginine or lysine residues we are witnessing. Overall, the pK_a values for both ionisabe groups appear rather high and possibly unachievable in the natural environment of this organism. *Nocardiopsis dassonvillei* belongs to the group of actinobacteria and has been found primarily in coastal wetland such as the Yellow River of the Shandong peninsula[197]. Whilst most actinobacteria are known to grow at neutral pH, there have been reports of alkalophilic strains of *Nocardiopsis dassonvillei* which would allow the CDO to experience optimal pH for catalysis[198].

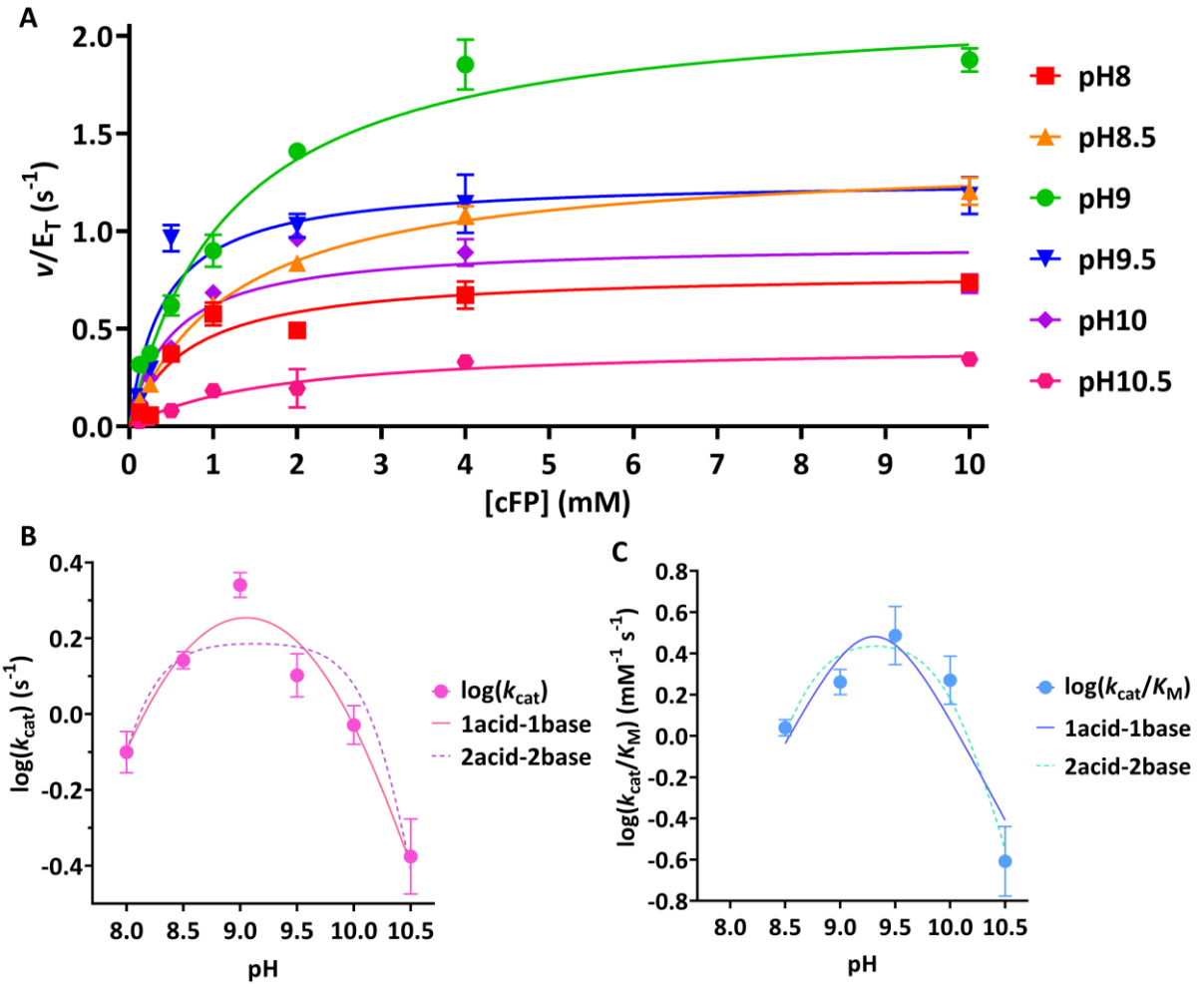


Figure 6.8 pH-rate profile of *Ndas*CDO with cFP at a constant temperature.
A – Substrate saturation curves for *Ndas*CDO in the range of pH 8 to 10.5. Data were fitted to Equation 2.3 **B** – pH-rate profile of *Ndas*CDO k_{cat} fit to two different models. Data were fitted to two models: 1 acid, 1 base and 2 acid, 2 base (Equation 2.4 or 2.5). **C** – pH-rate profile for k_{cat}/K_M of *Ndas*CDO fit to two different models. All experiments were performed in triplicate and the standard error of the mean was shown as error bars (as calculated by GraphPad Prism 9.5.1).

6.2.5 Kinetic analysis of *Ndas*CDO

In order to characterise substrate preference and enzyme catalytic proficiency, the kinetic parameters – k_{cat} and k_{cat}/K_M - for different CDP substrates were determined using the direct spectrometric plate reader assay described above. Consequently, this relies on the enzyme acting under Henri-Michaelis-Menten conditions which assumes that product formation and release is slower than the formation of the ES complex which is in equilibrium with both reactants (CDO and CDP in this case)[196].

These investigations aimed to uncover kinetic parameters for *Ndas*CDO when turning over different substates. The K_M parameter highlights the substrate concentration that evokes an overall reaction velocity that is half of the expected maximum velocity (V_{max}). Alternatively, this can be viewed as the substrate concentration where half of the active sites are filled by substrate molecules. The

turnover number k_{cat} can be calculated by dividing V_{max} by the enzyme concentration, however V_{max} must be converted into productive units (concentration per second), therefore requiring the extinction coefficient of the product. Under steady state conditions and a short timeframe, it was assumed that the CDP substrate would be converted into the di-dehydrogenated product only. This oxidation could, however, alter the extinction coefficient therefore it was imperative to first identify the extinction coefficient of the product.

One method to estimate the change in extinction coefficient of the oxidised product involves the use of a progress curve where a wavelength relating to the product is monitored until the reaction reaches completion (i.e. a plateau). Initial attempts at this experiment monitored only the wavelength at 297 nm (Figure 6.9), relating to the absorbance spectra collected in Figure 6.6. The experiment was performed at a range of concentrations of substrate whilst keeping the enzyme constant under steady state conditions. After 24 hours of observation, it was assumed that all substrate had been converted into product and so the last twenty points from the curve were averaged and plotted against the substrate concentration which is effectively the product concentration now to give a straight line where the slope was equal to the change in absorbance per unit concentration (ΔAbs μM^{-1}, Table 6.3). This variable could be used as a substitute for the unknown extinction coefficient to transform V_{max} from units of ΔAbs s^{-1} to μM s^{-1} which would be useful for determining k_{cat}.

Upon closer inspection of the progress curves, it appears that there was no trend for cWW when monitored at 297 nm which could now be explained by the repeat of the UV absorbance spectra (Figure 6.7) highlighting a more appropriate wavelength to monitor would be 352 nm. Furthermore, both progress curves for cFP and cHF did not reach plateau after measuring for 24 hours indicating that not all substrate has been oxidised to product. Therefore, the straight line and equivalent extinction coefficient calculated for these two CDPs is not wholly reflective of the actual extinction coefficient of the product and the reaction should be monitored for an extended period of time to ensure a plateau is reached. Additionally, the UV absorbance spectra for these two peptides displays alternative wavelengths that may be more indicative of the oxidised product rather than monitoring at 297 nm – cFP at 312 nm, and cHF at either 310 nm or 358 nm. The relatively strange profile of the curves for cWF indicates that 297 nm is not the most optimal wavelength to monitor for the generation of oxidised cWF. Overall, this preliminary experiment confirmed that a progress curve monitoring product formation could be used to calculate the altered extinction coefficient of a said product, however, repeats of this should incorporate the newly discovered wavelengths from Figure 6.7 which appear optimal for these oxidised CDP products.

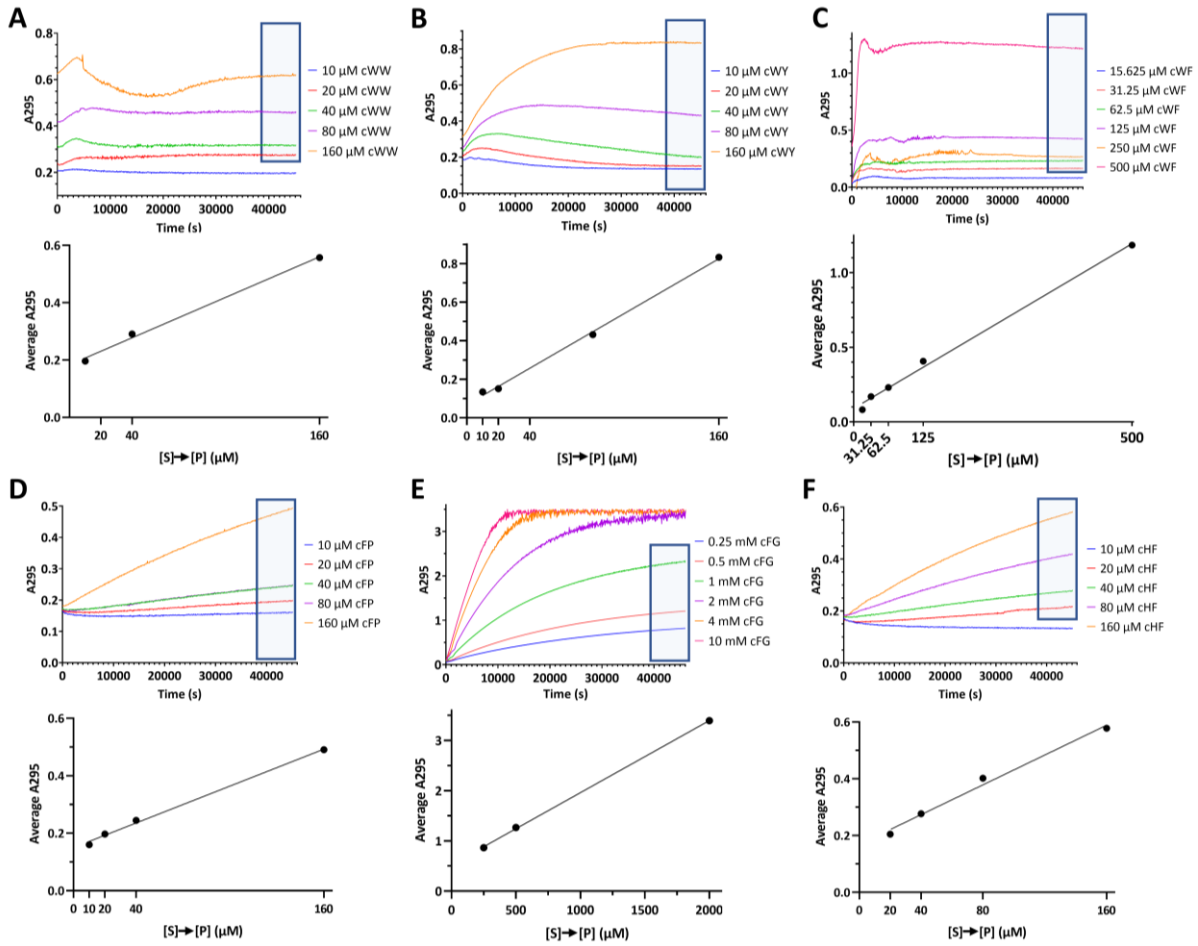


Figure 6.9 Overnight progress curve monitoring formation of oxidised CDP.
Progress curves for cWW (**A**), cWY (**B**), cWF (**C**), cFP (**D**), cFG (**E**) and cHF (**F**) at different substrate concentrations, dependent on the solubility of the peptide. The top graph of each set depicts the increase in absorbance at 295nm over time as the reactant is transformed into the oxidised product. The blue box depicts the region of points averaged to give the absorbance values which were plotted against substrate concentration. The bottom graph plots the substrate concentration which is effectively the product concentration against the absorbance at 295 nm averaged by taking the last 20 points of the overnight progress curve. A linear trend can be fitted to this data where the slope is the change in absorbance (ΔAbs) per μM units.

Table 6.3 Slope of the line as determined by a linear fit to the average data from a progress curve against substrate/product concentration.
A linear fit was applied to the data and the slope with the goodness of fit (R^2) shown to 4 significant figures.

Peptide	Slope (ΔAbs μM^{-1})	R^2
cWW	0.002355	0.9958
cWY	0.004731	0.9975
cWF	0.002207	0.9947
cFP	0.002141	0.9965
cFG	0.001435	0.9997
cHF	0.002625	0.9883

At the time of performing this experiment, only absorbance data collected at 297 nm was available, therefore the following kinetic parameters were calculated using the extinction coefficients shown in Table 6.3. It is now obvious that these extinction coefficients are not accurate and future experiments should first focus on obtaining better representatives of this parameter to use in the calculation of k_{cat} and K_M for *Ndas*CDO. Initial velocity plots were obtained in triplicate for each CDP where the rate was corrected with the enzyme concentration to plot v/[E] in per second units (s^{-1}) against the substrate concentration in molar (typically millimolar in this case, mM) (Figure 6.10). The elucidation of these steady state parameters allowed us to determine k_{cat}/K_M which is a direct measurement of the catalytic efficiency of an enzyme (Table 6.4). From this analysis, *Ndas*CDO had the highest k_{cat}/K_M - parameter when oxidising cFP followed by cWW. Whilst the K_M for cWW is the lowest of all the CDPs tested, the time it takes for *Ndas*CDO to turnover is rather slow in comparison to other substrates, especially for cHF with a k_{cat} of 5.682 s^{-1}. It should be highlighted that the natural substrate, cFL, was not investigated here and future work should therefore aim to identify key kinetic parameters using this CDP for complete characterisation of *Ndas*CDO. Again, these kinetic parameters rely on the inaccurate extinction coefficients calculated previously therefore any information inferred from these is not an accurate representation of the kinetic capabilities of *Ndas*CDO.

Table 6.4 Comparison of NdasCDO steady-state kinetic parameters with different cyclodipeptide substrates.
All measurements were performed in triplicate and the mean presented alongside the standard error of the mean (SEM) to 2 significant figures.

Cyclodipeptide	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (mM^{-1} s^{-1})
cWW	0.61 ± 0.058	0.19 ± 0.042	3200 ± 780
cWF	0.29 ± 0.029	0.29 ± 0.062	990 ± 230
cFP	1.2 ± 0.041	0.31 ± 0.048	3900 ± 620
cFG	1.9 ± 0.13	3.1 ± 0.55	620 ± 120
cHF	5.7 ± 0.49	6.1 ± 0.96	930 ± 170

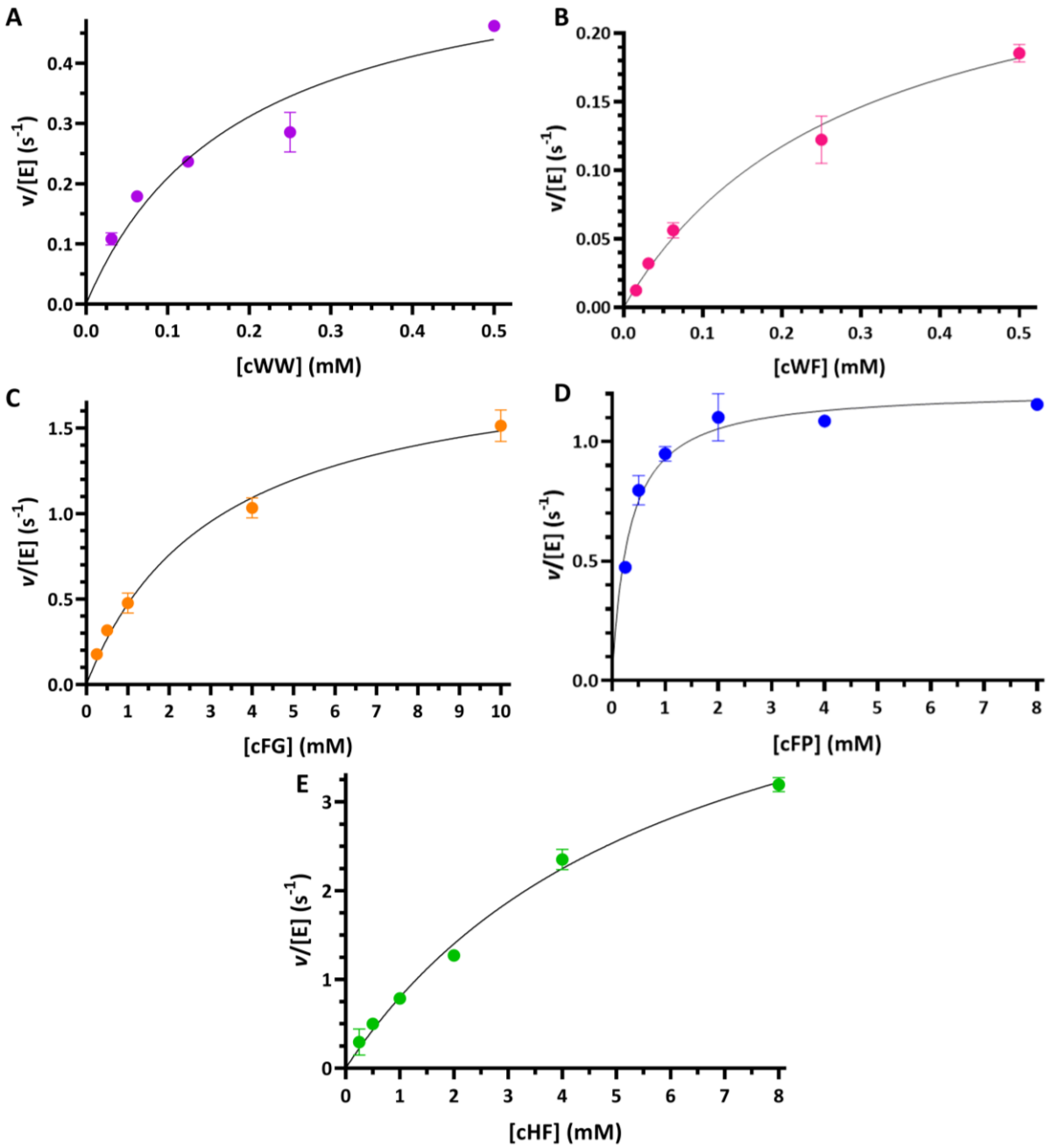


Figure 6.10 Substrate saturation curves for *Ndas*CDO with varying substrates.
Trends were fitted using equation 2.3 to yield the respective K_M and k_{cat} parameters for each CDP: **A** – cWW (purple); **B** – cWF (pink); **C** – cFG (orange); **D** – cFP (blue) and **E** – cHF (green). All data points were performed in triplicate and the standard error of the mean is represented by error bars.

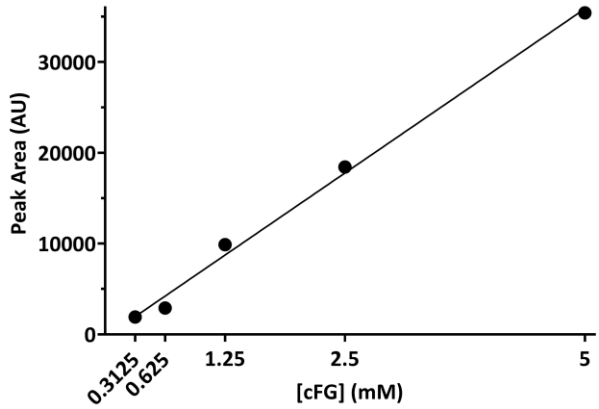
6.2.6 Investigation of oxidised cyclodipeptide extinction coefficient

An extinction coefficient is a conventional characteristic used often with regards to peptides and proteins as they feature amino acids with aromatic side groups that absorb mostly at 280 nm[199]. To further this point, CDOs introduce additional double bonds which extend the conjugation within the cyclodipeptide structure, likely increasing both the extinction coefficient and the absorption wavelength. Hence, it can be assumed that the oxidised products from CDO reactions would exhibit different extinction coefficients to the starting substrate. To ensure accurate characterisation of CDO

kinetic parameters, it would be imperative to determine the extinction coefficient of each individual oxidised product and confirm this hypothesis.

Isolation of oxidised cyclodipeptides was proposed to be possible using HPLC analysis followed by fraction collection of peaks pertaining to product(s). For simplicity, cFG was chosen for proof of concept where only one product was expected, and phenylalanine had a known extinction coefficient at 257 nm[200]. It was hypothesised that the starting material would absorb at 254 nm but not at 297 nm, thereby monitoring both wavelengths would distinguish peaks relating to substrate and resultant product for purification. Ultimately, oxidised cyclo(L-Phe*-L-Gly) (cF*G) was successfully isolated from a reaction mixture using HPLC following the method stated in Section 2.8.7. The peak relating to this product had a shift in retention time and a greater absorbance at 297 nm compared to the unoxidized counterpart as expected (Figure 6.12). A calibration curve was created from the commercial cFG standard (Figure 6.11) and this was used to calculate the new extinction coefficient for cF*G at 297 nmwhich was estimated to be 13.28x greater than the extinction coefficient for cFG at 297 nm (2589.6 M^{-1} cm^{-1}). This parameter was used to calculate the turnover number of *Ndas*CDO when using cFG as a substrate which was found to be 1.080 ± 0.074 s^{-1} – a reduction in k_{cat} by nearly a half in comparison to using the slope calculated from the overnight progress curves.

When using the overnight progress curves, the measurement reveals the change in absorbance over the measured time period. At this particular wavelength, several molecules also absorb light including residual starting material as well as the CDO protein itself present within the reaction. Thus, the extinction coefficient gathered from HPLC is a more realistic representation of the actual change in absorbance caused by the formation of oxidised product. On the other hand, the overnight progress curves of some peptides did display detector saturation or fail to reach completion which would prevent an accurate insight into the product yield. As a result, the previously calculated k_{cat} values in Section 6.2.5 should be interpreted cautiously until the experiment can be repeated without saturation effects.



[cFG] (mM)	Peak Area (AU)
0.3125	1912.72
0.625	2899.96
1.25	9879.96
2.5	18438.56
5	35412.44

Figure 6.11 Calibration curve of cyclo(L-Phe-L-Gly).
HPLC analysis of cFG standards highlighted a linear correlation between substrate concentration and peak area. This was fitted to a straight line trend with an R^2 of 0.9952 using Prism (GraphPad).

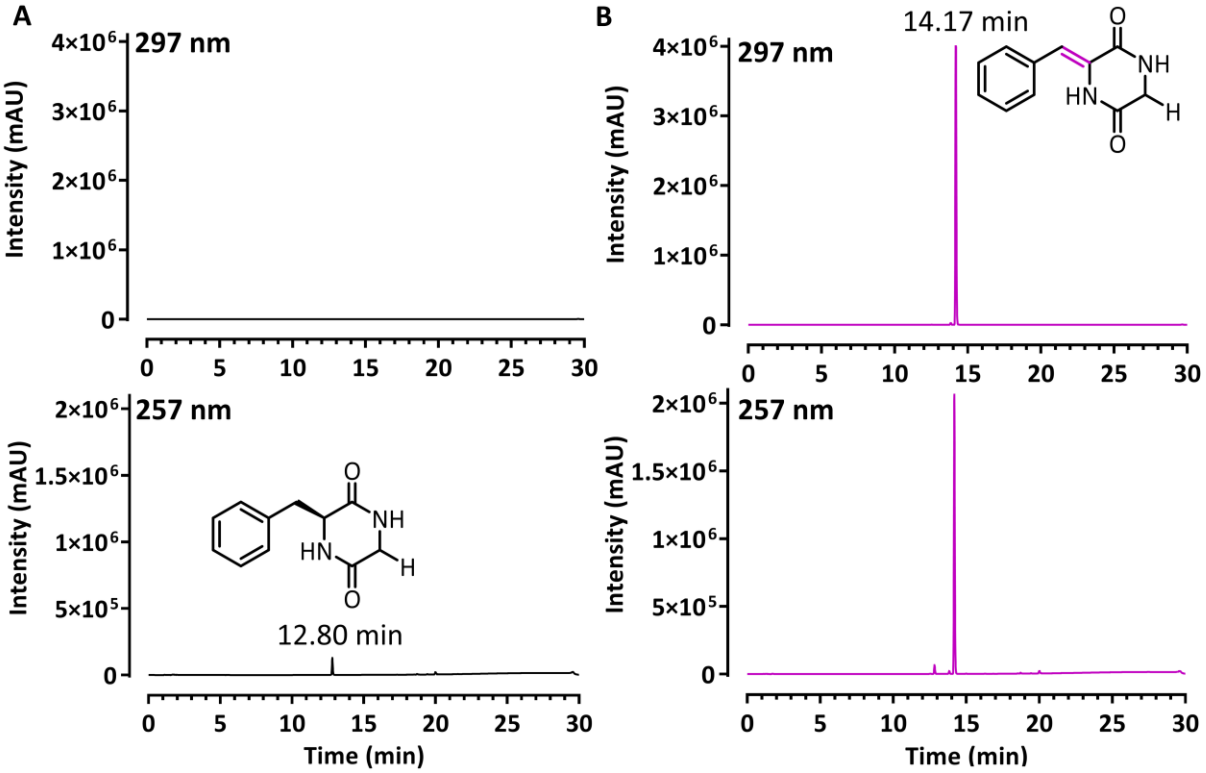


Figure 6.12 HPLC analysis of *Ndas*CDO reaction with cFG.
A – HPLC trace of control reaction containing peptide with no enzyme displayed a peak at 12.80 minutes believed to be substrate which only absorbed at the wavelength expected for phenylalanine (257 nm). **B** – HPLC trace of reaction (in purple) with *Ndas*CDO highlighted the presence of a new peak at 14.17 minutes which absorbed at both 257 and 297 nm, indicative of an oxidised product.

6.2.7 Electron paramagnetic resonance to probe existence of flavin radical

The mechanism of cyclodipeptide oxidases remains unknown to date with very little research addressing this family of enzymes. Understanding the mechanism of action would be highly beneficial to the general characterisation of these enzymes which would not only contribute to basic flavoenzyme research but also allow them to be of better use for protein engineering efforts. Given

that there is no published structure and all previous crystallography attempted with *Ndas*CDO for this project was unsuccessful, alternative methods were sought to probe the unknown path to oxidation. The flavin cofactor, essential for activity, was envisioned to give a potential insight into the overall workings of CDOs as the flavin redox states are well documented in the literature. Oxidised flavin is able to accept electron(s) to create two different states: semiquinone (SQ) and hydroquinone (HQ), both of which can exist as either anionic (ASQ/AHQ) or neutral (NSQ/NHQ) species (Figure 6.13)[201]. The semiquinone structure features a radical which can be difficult to identify from an analytical point of view as radical formation is typically fast and transient in nature i.e. short lived in low concentrations[202]. Therefore, specialised techniques such as electron paramagnetic resonance (EPR) spectroscopy have been employed to directly measure such elusive intermediates.

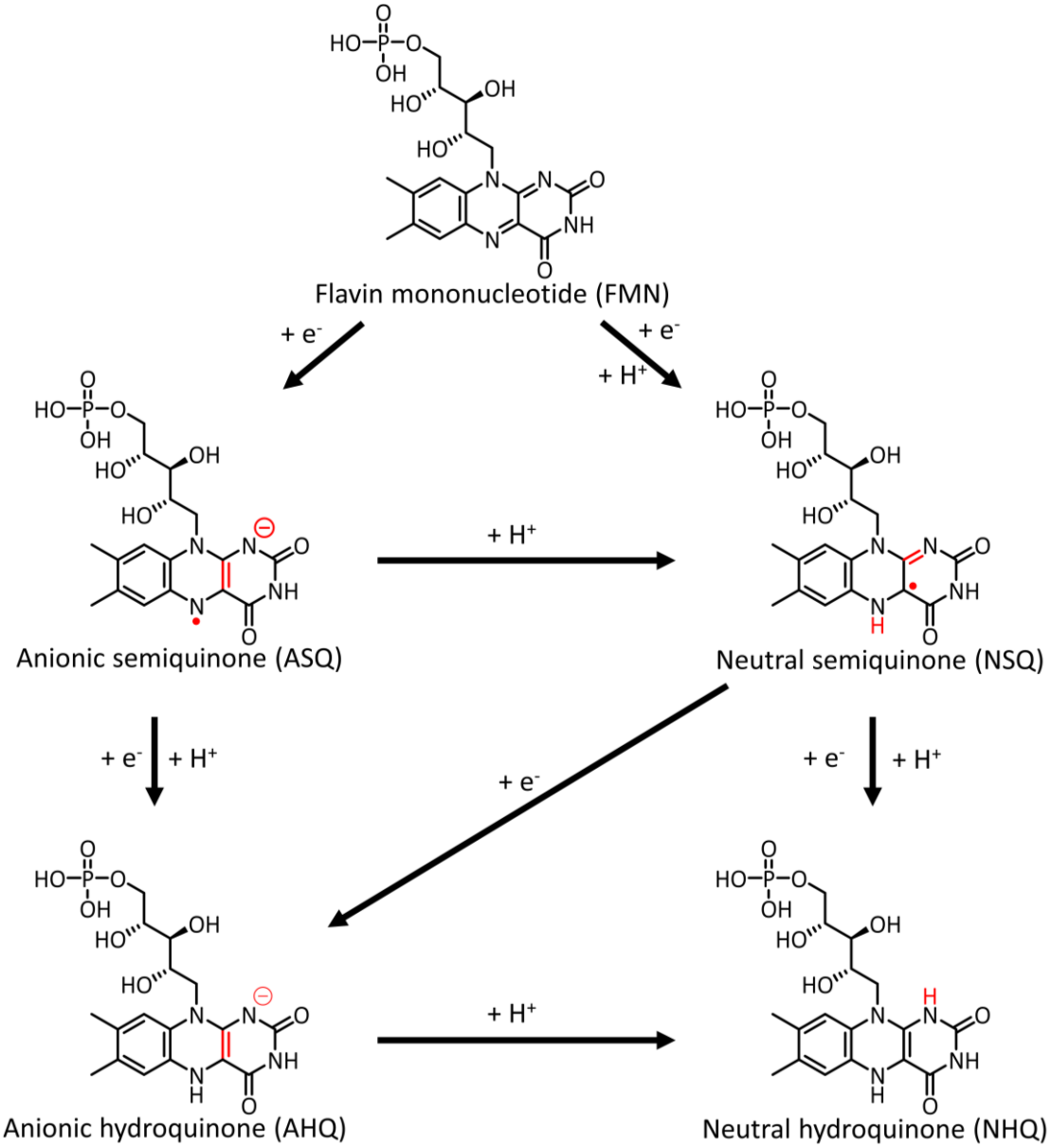


Figure 6.13 Possible redox states of flavin mononucleotide[201].

EPR measures the absorption of electromagnetic radiation of paramagnetic species experiencing a magnetic field[202]. Paramagnetism arises from the presence of more than one unpaired electron therefore EPR is highly specific for radicals and transition metal ions. By observing a specific flavin redox state, either SQ or HQ, details of the enzymatic mechanism could potentially be deduced. EPR measurements were therefore performed in collaboration with Dr Bela Bode from the School of Chemistry at University of St Andrews. Using cFG in excess of the calculated K_M, spectra were obtained for both the reaction and two controls containing either solely cFG or *Ndas*CDO separately. These data highlighted the presence of an organic radical species participating in the oxidation of cFG (Figure 6.14). However, this specific analysis could not divulge further information on the nature of the radical thus it could be either an anionic or a neutral semiquinone flavin radical. Low enzyme concentration limited the signal-to-noise of this assay and further experiments should aim to improve on this in order to elucidate mechanistic properties of CDOs.

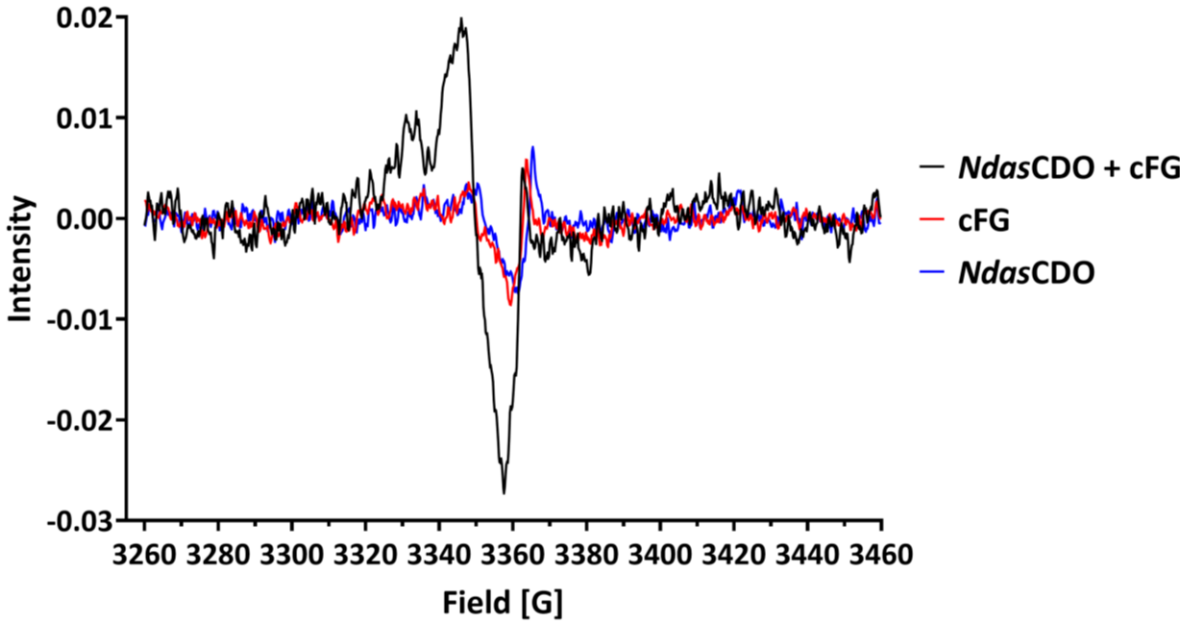


Figure 6.14 EPR spectroscopy of *Ndas*CDO incubated with cyclo(L-Phe-L-Gly).
The reaction, *Ndas*CDO and cFG, displayed a pronounced signal in comparison to either control indicating the presence of an organic radical species originating from the reaction itself.

6.3 Chapter summary

This chapter explored previously unknown characteristics of cyclodipeptide oxidases, specifically the catalytic and chemical mechanism of *Ndas*CDO. A plasmid containing two genes encoding both subunits of *Ndas*CDO was purchased from Genscript allowing for the purification of this protein complex from *E. coli* using a hexahistidine tag with nickel affinity chromatography. The substrate specificity of *Ndas*CDO was explored using a range of commercially purchased cyclodipeptides which were available in the lab at the time. LC-MS analysis revealed that *Ndas*CDO was able to oxidise every

tested CDP substrate to both the di- and tetra-dehydrogenated versions, indicating that this family of enzymes display advanced promiscuity, akin to their counterparts, CDPSs. Kinetic parameters for *Ndas*CDO were determined using the data collected at the time of thesis submission which has since been realised to be incorrect. The latest experiments with *Ndas*CDO revealed that there are specific UV absorbance wavelengths which are unique to each oxidised CDP and these should be used when determining the k_{cat} and K_M parameters to ensure these are reflective of the oxidised product. Additionally, the natural substrates of *Ndas*CDO - cFY, cFF and cFL - were not tested here and should be included in the kinetic characterisation of this enzyme to give a fair comparison of catalytic efficiency between different substrates.

With no structural information to date, it is difficult to postulate specific residues or interaction networks participating in the seemingly intricate reaction mechanism of the CDO family. It is accepted that an FMN cofactor is essential for enzymatic activity and sequence alignment of the CDO subunit A with other flavin-binding nitroreductase homologs highlighted that some key residues which interact with flavin are also conserved in CDOs[179]. When given the sequence of *Ndas*1146, the structural prediction server, I-TASSER (Iterative Threading ASSEMbly Refinement) was able to identify the flavin binding region of *Ndas*1146 based purely on structural homology[203]. The flavin oxidation-reduction cycle must be participating in the overall mechanism of the CDOs and here we attempted EPR spectroscopy to investigate this. Preliminary results did indicate the presence of a radical, however this could not be confirmed to be originating from a reactive semiquinone flavin species purely from this experiment. Indeed, other studies with flavoenzymes have used EPR to probe the redox states of flavin and have successfully observed a flavin-semiquinone species[204]. Our current data, whilst noisy, does bear similarities to published EPR data on flavin-semiquinones, thus strengthening the argument that the flavin cofactor must be cycling between oxidised and reduced forms during the CDO reaction. Regardless of the information presented here, the overall reaction mechanism of cyclodipeptide oxidases remains unknown and is clearly a complex topic that requires significantly more data to elucidate even the smallest of details about the catalytic mechanism at work here.

Chapter 7 - Conclusions and future work

The research described herein details the potential enzymatic routes to novel cyclodipeptides featuring histidine. Whilst there are established synthetic routes to the 2,5-diketopiperazine scaffold, these lack the capability to incorporate a myriad of complex derivatisations often found in natural products. Thus, it would be valuable to exploit the same biosynthetic pathways found in nature to generate novel molecules harbouring key properties for clinical applications. Chapter 3 introduced a family of enzymes known as cyclodipeptide synthase enzymes where only two are currently known to accept histidine as a substrate, *Para*CDPS and *Parcu*CDPS. *Para*CDPS, isolated from *Parabacteroides sp. 20_3*, synthesises cyclo(L-His-L-Phe) and *Parcu*CDPS from *Parcubacteria bacterium RAAC4_OD1_1* is known to generate two products: cyclo(L-His-L-Glu) and cyclo(L-His-L-Pro).

The initial aim to optimise product formation by CDPSs targeted their natural substrates, aminoacylated tRNA. Two tRNA isolates were prepared – tRNA pool and S30 extract – to rival the conventional *in vitro* transcriptional process to obtain the tRNA necessary for the CDPS reaction. Our original hypothesis proposed that using the S30 extract would simplify product formation thus allowing for a fast high throughput method to CDPs. However, the tRNA pool was discovered to be the better route to higher yields from CDPSs thereby removing the need to perform *in vitro* transcription. Whilst the S30 extract was not useful for its original purposes, it could be repurposed as a straightforward method to confirm CDPS activity *in vitro* as it removes the need to individually purify each required tRNA and it's respective aminoacyl-tRNA synthetase. Furthermore, investigations into large scale productions of CDPs using CDPSs are yet to be performed but are imperative to the argument that CDPSs are useful biosynthetic tools that rival conventional organic synthesis to these types of molecules.

An alternative method to bypass the requirement for tRNA all together was also considered, where an activated minimal substrate unit, referred to as aa-DBE, was used instead. This did not prove to be a more efficient substitute for aminoacylated tRNA, however these experiments revealed an opposing trend where more product was formed using different combinations of aminoacylated tRNA with aa-DBE. By trapping the enzyme in an intermediate state using the DBE substrates, the exact binding order of each CDPS was deciphered, whereupon it was realised that *Para*CDPS accepts phenylalanine in the second binding pocket whilst *Parcu*CDPS accepts histidine first. The use of DBE substrates in place of tRNA has been developed for other enzymes, including a family of RNA based flexizymes[146]. It was previously reported that flexizymes can aa-DBE substrates to charge shortened versions of tRNA with the related amino acid, thus providing a facile route to aminoacylated tRNA. Given that no kinetic analysis was performed on *Para*CDPS or *Parcu*CDPS, this route of using aa-DBE with flexizymes could

allow us to interrogate the kinetic properties of these enzymes and possibly reveal additional distinct differences between the binding of histidine in either pocket.

The establishment of a simplified route to tRNA production was also beneficial to *in vitro* studies of CDPS promiscuity with non-canonical amino acids. Referring to the literature highlighted an abundance of unnatural amino acids that could be charged to tRNA bodies by their respective aminoacyl-tRNA synthetases. Therefore, using CDPSs and commercially available unnatural amino, a small library of novel cyclodipeptides, with no known organic synthetic method to date, was created. Chemically interesting groups were able to be incorporated onto the ring side chains such as fluorine or terminal alkynes and nitriles. Fluorination of molecules has been shown to an asset which can be exploited for NMR or vibrational spectroscopic probes whilst alkynes are valuable click chemistry handles for use in copper free selective labelling[86]. This work lays the foundation for using CDPSs to generate diverse molecules which can be applied in several directions, however future work should confirm the purification of these from the CDPS reaction mixture. Separation using HPLC is conceivable for this task yet the yield of product obtained at the end of the purification may pose quite the issue for future applications of these new CDPs. Additionally, only commercially available amino acids were tested thus providing a biased view of the expansive promiscuity of CDPSs. A future avenue to explore should involve the incorporation of amino acid mimics only achievable via synthetic methods which give a wider view of the potential products available from CDPSs.

A direct mass spectrometric assay was applied provisionally to observe tRNA charged with unnatural amino acids – unfortunately this was found to be unsuccessful for such purposes. It should be stated that further development of an MS assay capable of detecting aminoacylated tRNA could be a benefit for CDPSs and tRNA research in general. For example, it would allow for a greater scope of unnatural amino acids to be tested as substrates in a high throughput manner, as current mass spectrometric efforts require advanced derivatisation before analysis which may not always be compatible with the amino acid under investigation. The development of a direct MS assay detailed herein to detect methylation of tRNA proves that this concept, whilst difficult, could be achievable.

Determining the structure of *Parcu*CDPS provided valuable insight into the unique properties of this enzyme by highlighting divergent regions from the archetypical CDPS structure. The first indication of these differences originated from the failure of molecular replacement using other CDPSs to solve the structure of *Parcu*CDPS which ultimately required iodide-SAD phasing. Structural comparison, described in Chapter 5, highlighted changes in direction of certain secondary structural features including one β strand which cut through the active site and second pocket feature in *Parcu*CDPS. Conserved residues near the active site likely play a similar role in the catalytic mechanism of

*Parcu*CDPS as site directed mutagenesis of residues in this area resulted in inactive variants. Consequently, it was assumed *Parcu*CDPS participated in the same mechanistic pathway like all other known CDPSs and any observed structural differences in *Parcu*CDPS had no effect on this aspect.

Considering *Parcu*CDPS could accept histidine in the first binding pocket, attention was directed to understanding the recognition pathway imposed by this enzyme on incoming substrates. By comparison with the binding network of histidyl-tRNA synthetase (HisRS), the role of three residues of interest was investigated: Tyr55, Glu174, Tyr189. Extensive mutagenesis of these amino acids highlighted the undoubtedly vital role they play in recognising histidine by *Parcu*CDPS. Successive double and triple mutants were most affected when examining the product yield, therefore it was evident that it was not a single residue but instead an intricate balance of these three pocket residues which were responsible for interacting with the histidine substrate. Moreover, supposedly inactive mutants were tested for cyclodipeptide formation using alternative amino acids readily available in the lab at this time such as phenylalanine and leucine. Initial attempts using aminoacylated tRNA proved unremarkable with limited product yield, yet this was improved when using aa-DBE with Pro-tRNA^{Pro} suggesting that mutant *Parcu*CDPS retained the ability to discriminate between the accepted histidyl-tRNA body and the alternative incoming tRNA. Thus, activated amino acids circumvented this 'screening' process by removing the potential for unfavourable interactions with the tRNA body which would cause the substrate to be rejected. This rational engineering is the first instance of successful modification to the selectivity of a CDPS and alludes to the sophisticated recognition networks applied by this family of enzymes. Previous published studies concentrated on altering large regions of a CDPS which failed to yield positive results; therefore our more tailored and structure-guided approach demonstrated the subtleties of a few key pocket residues that were previously overlooked. Our strategy and findings could be applied to other CDPS members to first pinpoint pocket residues responsible for substrate recognition and then alter these to influence the overall selectivity, thus generating multiple products from a single catalyst. An alternative approach could involve altering pocket residues of additional CDPS members to evoke a similar switch in the substrate specificity. An obvious starting point for this would be the incorporation of the three key pocket residues of *Parcu*CDPS – Y55, E174, Y189 - into another CDPS to investigate if this would allow the enzyme to now accept histidine as a substrate.

Tailoring enzymes complement the diversity observed in natural products created via biosynthetic pathways and are often found next to CDPSs within the genome. In Chapter 6, a cyclodipeptide oxidase, *Ndas*CDO, was investigated as a potential modifier to the cyclic products produced by *Para*CDPS and *Parcu*CDPS. However, biochemical characterisation of *Ndas*CDO was first necessary, considering very little had been published on this class of enzymes. The promiscuity of the CDO was

verified using mass spectrometry which highlighted that all CDPs tested here were oxidised to either di- or tetra-dehyrogenated products. This promiscuity is not unlike CDPSs and is a promising finding that the CDOs are also not limited by substrate specificity. Preliminary kinetic analysis of *Ndas*CDO with a range of CDPs was limited due to the lack of known extinction coefficients for the oxidised products, especially for peptides with no conjugated side groups for UV absorbance. Research performed since the submission of this thesis has highlighted that unique UV absorbance wavelengths related to the individual oxidised CDPs. Therefore, a complete repeat of all kinetic experiments detailed in Chapter 6 should involve the newly discovered wavelengths to provide accurate interpretations of the parameters for each substrate.

One of the key motivations for investigating tailoring enzymes originated in the hypothesis that an enzymatic cascade could be designed from CDPSs with the various types of tailoring enzymes. This style of experiment was envisioned to introduce a wider range of diversity to the 2,5-diketopiperazine ring by combining different enzymes to incorporate the equivalent groups. This has yet to be tested but previous research has shown it did work with CDPSs and CDOs from different organisms *in vivo*[94]. This is encouraging evidence that creating a combinatorial system *in vitro* would also work. As shown in Chapter 4 and in previous literature, CDPSs are capable of accepting unnatural amino acids, however these have not been tested with CDOs or any tailoring enzymes to date. The establishment of successive enzymatic processes would therefore allow the investigation of CDO-catalysed oxidation of unnatural amino acid groups.

Furthermore, the unique and complicated oligomeric state of CDOs is yet to be addressed here or in the literature at present. Alphafold predictions of the flavin binding subunit, *Ndas*1146, indicated that this protein was similar to other nitroreductases which also bind flavin. There was an elongated unstructured region at the N-terminus of Ndas1146 which did not superimpose with the homologs. Therefore, current research in our lab is dedicated to determining if this section is responsible for the structural organisation of the subunits to form the complex oligomeric state by creating truncated versions and testing their respective enzymatic activity. Given that the only structural information available is from predicted homologs which do not catalyse the same reaction as CDOs, intensive structural studies should be pursued. Our efforts have shifted to the potential to use cryo-EM to visualise a structure of potentially mega Daltons in size but for now this remains limited as a pure homogenous sample is required for this technique and previous attempts at SEC with *Ndas*CDO have failed. A tailored buffer screen with *Ndas*CDO to explore the stability of the complex in different systems should unveil an optimal condition that can then be applied for cryo-EM purposes.

In summary, the work described in this thesis acts as a spotlight into histidine accepting CDPSs which have been overlooked in current literature. Exploitation of their distinct substrate promiscuity could pave the way to potentially novel cyclodipeptides with untapped capabilities, especially in a therapeutic setting. As the first report of successful rational tailoring with a CDPS, the future impact of this finding is anticipated to be of particular use for creating an advanced enzyme with unmatched substrate selectivity for efficient production of a vast cyclodipeptide library. Furthermore, the introduction of tailoring enzymes to form a cascade of product diversity would only enhance the types of molecules available from biosynthetic pathways. This evidence complements current work on research of routes to natural products and highlights a potentially alternative facile method to such interesting molecules.

8 - References

1. Ji, H.-F., Li, X.-J. & Zhang, H.-Y. Natural products and drug discovery. *EMBO Rep* **10**, (2009).

2. Cragg, G. M. & Newman, D. J. Biodiversity: A continuing source of novel drug leads. *Pure and Applied Chemistry* **77**, 7–24 (2005).

3. Dias, D. A., Urban, S. & Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2**, 303–336 (2012).

4. Maplestone, R. A., Stone, M. J. & Williams, D. H. The evolutionary role of secondary metabolites--a review. *Gene* **115**, 151–157 (1992).

5. Hutchings, M. I., Truman, A. W. & Wilkinson, B. Antibiotics: past, present and future. *Curr Opin Microbiol* **51**, 72–80 (2019).

6. Zhang, X., Hindra & Elliot, M. A. Unlocking the trove of metabolic treasures: activating silent biosynthetic gene clusters in bacteria and fungi. *Curr Opin Microbiol* **51**, 9–15 (2019).

7. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol* **43**, 155–176 (2016).

8. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147 (2002).

9. Ikeda, H. *et al.* Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**, 526–531 (2003).

10. Bauman, K. D., Butler, K. S., Moore, B. S. & Chekan, J. R. Genome mining methods to discover bioactive natural products. *Nat Prod Rep* **38**, 2100–2129 (2021).

11. Demain, A. L. Importance of microbial natural products and the need to revitalize their discovery. *J Ind Microbiol Biotechnol* **41**, 185–201 (2014).

12. Lautru, S., Gondry, M., Genet, R. & Pernodet, J.-L. The Albonoursin Gene Cluster of S. noursei: Biosynthesis of Diketopiperazine Metabolites Independent of Nonribosomal Peptide Synthetases. *Chem. Biol.* **9**, 1355–1364 (2002).

13. Jia, J. *et al.* 2,5-Diketopiperazines: A Review of Source, Synthesis, Bioactivity, Structure, and MS Fragmentation. *Curr Med Chem* **29**, 1–26 (2022).

14. Borthwick, A. D. 2,5-Diketopiperazines: Synthesis, Reactions, Medicinal Chemistry, and Bioactive Natural Products. *Chem. Rev.* **112**, 3641–3716 (2012).

15. Horton, D. A., Bourne, G. T. & Smythe, M. L. Exploring privileged structures: the combinatorial synthesis of cyclic peptides. *J Comput Aided Mol Des* **16**, 415–431 (2002).

16. González, J. F., Ortín, I., de la Cuesta, E. & Menéndez, J. C. Privileged scaffolds in synthesis: 2,5-piperazinediones as templates for the preparation of structurally diverse heterocycles. *Chem. Soc. Rev.* **41**, 6902–6915 (2012).

17. Martins, M. B. & Carvalho, I. Dikeopiperazines: biological activity and synthesis. *Tetrahedron* **63**, 9923–9932 (2007).

18. E. Evans, B. *et al.* Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J Med Chem* **31**, 2235–2246 (1988).

19. RW, D., KS, C., SA, M., JW, D. & DA, P. Privileged structures: applications in drug discovery. *Comb Chem High Throughput Screen* **7**, 473–493 (2004).

20. Corey, R. B. The Crystal Structure of Diketopiperazine. *J Am Chem Soc* **60**, 1598–1604 (1938).

21. Mendham, A. P. *et al.* Vibrational spectroscopy and crystal structure analysis of two polymorphs of the di-amino acid peptide cyclo(L-Glu-L-Glu). *Journal of Raman Spectroscopy* **41**, 288–302 (2010).

22. Borthwick, A. D. & da Costa, N. C. 2,5-diketopiperazines in food and beverages: Taste and bioactivity. *Crit Rev Food Sci Nutr* **57**, 718–742 (2017).

23. Holden, M. T. G. *et al.* Quorum-sensing cross talk: isolation and chemical characterization of cyclic dipeptides from Pseudomonas aeruginosa and other Gram-negative bacteria. *Mol Microbiol* **33**, 1254–1266 (1999).

24. Elkahoui, S. *et al.* Cyclo-(His,Leu): A new microbial diketopiperazine from a terrestrial Bacillus subtilis strain B38. *Nat Prod Res* **27**, 108–116 (2013).

25. Zhao, P. *et al.* Non-lipopeptide fungi-derived peptide antibiotics developed since 2000. *Biotechnol Lett* **41**, 651–673 (2019).

26. Song, Z., Hou, Y., Yang, Q., Li, X. & Wu, S. Structures and Biological Activities of Diketopiperazines from Marine Organisms: A Review. *Mar Drugs* **19**, 403 (2021).

27. Bojarska, J. *et al.* Cyclic Dipeptides: The Biological and Structural Landscape with Special Focus on the Anti-Cancer Proline-Based Scaffold. *Biomolecules* **11**, 1515 (2021).

28. Balachandra, C., Padhi, D., Govindaraju, T. & Reviews, C. Cyclic Dipeptide: A Privileged Molecular Scaffold to Derive Structural Diversity and Functional Utility. *ChemMedChem* **16**, 2558–2587 (2021).

29. Bellezza, I., Peirce, M. J. & Minelli, A. Cyclic dipeptides: from bugs to brain. *Trends Mol. Med.* **20**, 551–558 (2014).

30. Ortiz-Castro, R. *et al.* Transkingdom signaling based on bacterial cyclodipeptides with auxin activity in plants. *PNAS* **108**, (2011).

31. Li, J., Wang, W., Xu, S. X., Magarvey, N. A. & McCormick, J. K. Lactobacillus reuteri-produced cyclic dipeptides quench agr-mediated expression of toxic shock syndrome toxin-1 in staphylococci. *Proc Natl Acad Sci U S A* **108**, 3360–3365 (2011).

32. Bofinger, M. R., de Sousa, L. S., Fontes, J. N. & Marsaioli, A. J. Diketopiperazines as Cross-Communication Quorum-Sensing Signals between Cronobacter sakazakii and Bacillus cereus. (2017) doi:10.1021/acsomega.6b00513.

33. Prasad, C. & Peterkofsky, A. Demonstration of Pyroglutamylpeptidase and Amidase Activities toward Thyrotropin-releasing Hormone in Hamster Hypothalamus Extracts. *J Biol Chem* **251**, 3229–3234 (1976).

34. Guan, J. & Gluckman, P. D. IGF-1 derived small neuropeptides and analogues: a novel strategy for the development of pharmaceuticals for neurological conditions. *Br J Pharmacol* **157**, 881–891 (2009).

35. Zhao, K., Xing, R. & Yan, X. Cyclic dipeptides: Biological activities and self-assembled materials. *Peptide Science* **113**, e24202 (2021).

36. Kanoh, K. *et al.* (−)-Phenylahistin: A new mammalian cell cycle inhibitor produced by aspergillus ustus. *Bioorg. Med. Chem. Lett.* **7**, 2847–2852 (1997).

37. Kanoh, K. *et al.* Antitumor Activity of Phenylahistin in Vitro and in Vivo. *Biosci Biotechnol Biochem* **63**, 1130–1133 (1999).

38. Nicholson, B. *et al.* NPI-2358 is a tubulin-depolymerizing agent: in-vitro evidence for activity as a tumor vascular-disrupting agent. *Anticancer Drugs* **17**, 25–31 (2006).

39. Hayashi, Y., Yamazaki-Nakamura, Y. & Yakushiji, F. Medicinal Chemistry and Chemical Biology of Diketopiperazine-Type Antimicrotubule and Vascular-Disrupting Agents. *Chem Pharm Bull (Tokyo)* **61**, 889–901 (2013).

40. Fdhila, F., Vázquez, V., Sánchez, J. L. & Riguera, R. Diketopiperazines: Antibiotics Active against Vibrioaguillarum Isolated from Marine Bacteria Associated with Cultures of Pectenmaximus. *J Nat Prod* **66**, 1299–1301 (2003).

41. Bellezza, I., Peirce, M. J. & Minelli, A. Chapter 10 - Cyclic Peptides in Neurological Disorders: The Case of Cyclo(His-Pro)a. in *Quorum Sensing: Molecular mechanism and biotechnological application* (ed. Tommonaro, G.) 257–286 (Academic Press, 2019).

42. Minelli, A., Bellezza, I., Grottelli, S. & Galli, F. Focus on cyclo(His-Pro): history and perspectives as antioxidant peptide. *Amino Acids* **35**, 283–289 (2008).

43. Grottelli I.; Pietrini G.; Peirce M.J.; Minelli A.; Bellezza I., S. ; F. The Role of Cyclo(His-Pro) in Neurodegeneration. *Int. J. Mol. Sci.* **17**, 1332 (2016).

44. Grottelli, S. *et al.* Cyclo(His-Pro) inhibits NLRP3 inflammasome cascade in ALS microglial cells. *Mol. Cell. Neurosci.* **94**, 23–31 (2019).

45. Hwang, I. K. *et al.* Effects of cyclo (his-pro) plus zinc on glucose metabolism in genetically diabetic obese mice. *Diabetes Obes Metab* **5**, 317–324 (2003).

46. Tullberg, M., Grøtli, M. & Luthman, K. Efficient synthesis of 2,5-diketopiperazines using microwave assisted heating. *Tetrahedron* **62**, 7484–7491 (2006).

47. Fridkin, M., Patchornik, A. & Katchalski, E. A Synthesis of Cyclic Peptides Utilizing High Molecular Weight Carriers. *J Am Chem Soc* **87**, 4646–4648 (1965).

48. Martínez-Núñez, M. A. & López, V. E. L. y. Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes 2016 4:1* **4**, 1–8 (2016).

49. Strieker, M., Tanović, A. & Marahiel, M. A. Nonribosomal peptide synthetases: structures and dynamics. *Curr Opin Struct Biol* **20**, 234–240 (2010).

50. Gondry, M. *et al.* Cyclodipeptide synthases are a family of tRNA-dependent peptide bond–forming enzymes. *Nat Chem Biol* **5**, 414–420 (2009).

51. Shi, H. & Moore, P. B. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **6**, 1091–1105 (2000).

52. Mishra, A. K., Choi, J., Choi, S.-J. & Baek, K.-H. Cyclodipeptides: An Overview of Their Biosynthesis and Biological Activity. *Molecules* **22**, 1796–1809 (2017).

53. Aravind, L., de Souza, R. F. & Iyer, L. M. Predicted class-I aminoacyl tRNA synthetase-like proteins in non-ribosomal peptide synthesis. *Biol Direct* **5**, 48 (2010).

54. Vetting, M. W., Hegde, S. S. & Blanchard, J. S. The structure and mechanism of the Mycobacterium tuberculosis cyclodityrosine synthetase. *Nat Chem Biol* **6**, 797–799 (2010).

55. Sauguet, L. *et al.* Cyclodipeptide synthases, a family of class-I aminoacyl-tRNA synthetase-like enzymes involved in non-ribosomal peptide synthesis. *Nucleic Acids Res.* **39**, 4475–4489 (2011).

56. Moutiez, M. *et al.* Unravelling the mechanism of non-ribosomal peptide synthesis by cyclodipeptide synthases. *Nat Commun* **5**, (2014).

57. Bonnefond, L. *et al.* Structural basis for nonribosomal peptide synthesis by an aminoacyl-tRNA synthetase paralog. *PNAS* **108**, 3912–3917 (2011).

58. Schmitt, E., Bourgeois, G., Gondry, M. & Aleksandrov, A. Cyclization Reaction Catalyzed by Cyclodipeptide Synthases Relies on a Conserved Tyrosine Residue. *Sci. Rep.* **8**, 7031 (2018).

59. Canu, N., Moutiez, M., Belin, P. & Gondry, M. Cyclodipeptide synthases: a promising biotechnological tool for the synthesis of diverse 2,5-diketopiperazines. *Nat Prod Rep* **37**, 312–321 (2020).

60. Canu, N. New insights into the recognition of the substrates of cyclodipeptide synthases. (2019).

61. Sutherland, E., Harding, C. J. & Czekster, C. M. Active site remodelling of a cyclodipeptide synthase redefines substrate scope. *Commun Chem* **5**, (2022).

62. Gondry, M. *et al.* A comprehensive overview of the cyclodipeptide synthase family enriched with the characterization of 32 new enzymes. *Front Microbiol* **9**, 46 (2018).

63. Seguin, J. *et al.* Nonribosomal peptide synthesis in animals: the cyclodipeptide synthase of Nematostella. *Chem Biol* **18**, 1362–1368 (2011).

64. Jacques, I. B. *et al.* Analysis of 51 cyclodipeptide synthases reveals the basis for substrate specificity. *Nat Chem Biol* (2015) doi:10.1038/nCHeMBIO.1868.

65. Bourgeois, G. *et al.* Structural basis for partition of the cyclodipeptide synthases into two subfamilies. *J Struct Biol* **203**, 17–26 (2018).

66. Bourgeois, G. *et al.* Structural basis of the interaction between cyclodipeptide synthases and aminoacylated tRNA substrates. *RNA* **26**, 1589–1602 (2020).

67. Harding, C. J., Sutherland, E., Hanna, J. G., Houston, D. R. & Czekster, C. M. Bypassing the requirement for aminoacyl-tRNA by a cyclodipeptide synthase enzyme. *RSC Chem. Bio.* **2**, 230–240 (2021).

68. Canu, N., Moutiez, M., Belin, P. & Gondry, M. Cyclodipeptide synthases: a promising biotechnological tool for the synthesis of diverse 2,5-diketopiperazines †. *Nat Prod Rep* **37**, (2020).

69. Gomez, M. A. R. & Ibba, M. Aminoacyl-tRNA synthetases. *RNA* **26**, 910–936 (2020).

70. Hoon Kwon, N., Fox, P. L. & Kim, S. Aminoacyl-tRNA synthetases as therapeutic targets. *Nat Rev Drug Discov* **18**, 629–650 (2019).

71. Kaiser, F. *et al.* Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases. *PLoS Comput Biol* **14**, e1006101 (2018).

72. Dutta Banik, S. & Nandi, N. Mechanism of the activation step of the aminoacylation reaction: a significant difference between class I and class II synthetases Mechanism of the activation step of the aminoacylation reaction: a significant difference between class I and class II synthetases. *J Biomol Struct Dyn* **30**, 701–715 (2012).

73. Kaiser, F. *et al.* The structural basis of the genetic code: amino acid recognition by aminoacyl-tRnA synthetases. *Sci Rep* **10**, (2020).

74. Cowie, D. B. & Cohen, G. N. Biosynthesis by Escherichia coli of active altered proteins containing selenium instead of sulfur. *Biochim Biophys Acta* **26**, 252–261 (1957).

75. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **244**, 182–188 (1989).

76. Bain, J. D., Glabe, C. G., Dix, T. A., Chamberlin, A. R. & Diala, E. S. Biosynthetic Site-Specific Incorporation of a Non-Natural Amino Acid into a Polypeptide. *J Am Chem Soc* **111**, 8013–8014 (1989).

77. Robertson, S. A., Ellman, J. A. & Schultz, P. G. A General and Efficient Route for Chemical Aminoacylation of Transfer RNAs. *J Am Chem Soc* **113**, 2722–2729 (1991).

78. Santoro, S. W., Wang, L., Herberich, B., King, D. S. & Schultz, P. G. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. (2002) doi:10.1038/nbt742.

79. Hendrickson, T. L., de Crécy-Lagard, V. & Schimmel, P. Incorporation of nonnatural amino acids into proteins. *Annu Rev Biochem* **73**, 147–176 (2004).

80. Wang, L., Xie, J. & Schultz, P. G. Expanding the genetic code. *Annu Rev Biophys Biomol Struct* **35**, 225–249 (2006).

81. Bessho, Y., Hodgson, D. R. W. & Suga, H. A tRNA aminoacylation system for non-natural amino acids based on a programmable ribozyme. (2002).

82. Kourouklis, D., Murakami, H. & Suga, H. Programmable ribozymes for mischarging tRNA with nonnatural amino acids and their applications to translation. *Methods* **36**, 239–244 (2005).

83. Ohuchi, M., Murakami, H. & Suga, H. The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus. *Curr Opin Chem Biol* **11**, 537–542 (2007).

84. Ngo, J. T. & Tirrell, D. A. Non-Canonical Amino Acids in the Interrogation of Cellular Protein Synthesis. *Acc Chem Res* **44**, 677–685 (2011).

85. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proc Natl Acad Sci U S A* **103**, 9482–9487 (2006).

86. Hartman, M. C. T. Non-canonical Amino Acid Substrates of E. coli Aminoacyl-tRNA Synthetases. *ChemBioChem* **23**, e202100299 (2022).

87. Canu, N. *et al.* Incorporation of Non-canonical Amino Acids into 2,5-Diketopiperazines by Cyclodipeptide Synthases. *Angewandte Chemie International Edition* **57**, 3118–3122 (2018).

88. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res* (2022) doi:10.1093/NAR/GKAC993.

89. Skinnider, M. A., Johnston, C. W., Merwin, N. J., Dejong, C. A. & Magarvey, N. A. Global analysis of prokaryotic tRNA-derived cyclodipeptide biosynthesis. *BMC Genomics* **19**, 1–11 (2018).

90. Borgman, P., Lopez, R. D. & Lane, A. L. The expanding spectrum of diketopiperazine natural product biosynthetic pathways containing cyclodipeptide synthases. *Org Biomol Chem* **17**, 2305 (2019).

91. Gondry, M. *et al.* Cyclic dipeptide oxidase from Streptomyces noursei. *Eur J Biochem* **268**, 1712–1721 (2001).

92. Giessen, T. W., von Tesmar, A. M. & Marahiel, M. A. Insights into the Generation of Structural Diversity in a tRNA-Dependent Pathway for Highly Modified Bioactive Cyclic Dipeptides. *Chem Biol* **20**, 828–838 (2013).

93. Shi, J. *et al.* Genome Mining and Enzymatic Total Biosynthesis of Purincyclamide. *Org Lett* **21**, 6825–6829 (2019).

94. Le Chevalier, F. *et al.* In vivo characterization of the activities of novel cyclodipeptide oxidases: New tools for increasing chemical diversity of bioproduced 2,5-diketopiperazines in Escherichia coli. *Microb Cell Fact* **19**, 1–16 (2020).

95. Meng, S. *et al.* A Six-Oxidase Cascade for Tandem C−H Bond Activation Revealed by Reconstitution of Bicyclomycin Biosynthesis. *Angewandte Chemie - International Edition* **57**, 719–723 (2018).

96. Patteson, J. B., Cai, W., Johnson, R. A., Santa Maria, K. C. & Li, B. Identification of the Biosynthetic Pathway for the Antibiotic Bicyclomycin. *Biochemistry* **57**, 61 (2018).

97. Giessen, T. W., von Tesmar, A. M. & Marahiel, M. A. A tRNA-Dependent Two-Enzyme Pathway for the Generation of Singly and Doubly Methylated Ditryptophan 2,5-Diketopiperazines. *Biochemistry* **52**, 4274–4283 (2013).

98. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571.

99. van den Ent, F. & Löwe, J. RF cloning: A restriction-free method for inserting target genes into plasmids. *J Biochem Biophys Methods* **67**, 67–74 (2006).

100. New England BioLabs (NEB). NEB Tm Calculator. https://tmcalculator.neb.com/#!/main.

101. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).

102. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom Rev* **39**, 229–244 (2020).

103. Adams, K. J. *et al.* Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. *Cite This: J. Proteome Res* **19**, 1447–1458 (2020).

104. Bertand Beckert, B. M. Synthesis of RNA by In Vitro Transcription. *RNA: Methods in molecular biology* **703**, 29–41 (2011).

105. Genomic tRNA Database for E. coli. http://gtrnadb.ucsc.edu/genomes/bacteria/Esch_coli_K_12_MG1655/Esch_coli_K_12_MG1655-gene-list.html.

106. Parcubacteria bacterium RAAC4_OD1_1 O210_OD1C00001, whole genome . https://www.ncbi.nlm.nih.gov/nuccore/AWSN01000001.1.

107. Tian, S. & Das, R. Primerize-2D: automated primer design for RNA multidimensional chemical mapping. *Bioinformatics* **33**, 1405–1406 (2017).

108. Tian, S., Yesselman, J. D., Cordero, P. & Das, R. Primerize: automated primer assembly for transcribing non-coding RNA domains. *Nucleic Acids Res* **43**, W522–W526 (2015).

109. Mechulam, Y., Guillon, L., Yatime, L., Blanquet, S. & Schmitt, E. Protection-Based Assays to Measure Aminoacyl-tRNA Binding to Translation Initiation Factors. in *Methods in Enzymology* (ed. Lorsch, J.) vol. 430 265–281 (Academic Press, 2007).

110. Krinsky, N. *et al.* A Simple and Rapid Method for Preparing a Cell-Free Bacterial Lysate for Protein Synthesis. *PLoS One* **11**, e0165137 (2016).

111. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19**, 751–755 (2001).

112. Yuet, K. P. *et al.* Cell-specific proteomic analysis in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* **112**, 2705–2710 (2015).

113. Hernandez, V. *et al.* Discovery of a Novel Class of Boron-Based Antibacterials with Activity against Gram-Negative Bacteria. *Antimicrob Agents Chemother* **57**, 1394–1403 (2013).

114. Peacock, J. R. *et al.* Amino acid-dependent stability of the acyl linkage in aminoacyl-tRNA. *RNA* **20**, 758–764 (2014).

115. Sweeney, P. *et al.* Structure, dynamics, and molecular inhibition of the Staphylococcus aureus m1A22-tRNA methyltransferase TrmK. *Journal of Biological Chemistry* **298**, (2022).

116. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallographica Section D* **69**, 1204–1214 (2013).

117. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D* **66**, 133–144 (2010).

118. Gildea, R. J. *et al.* New methods for indexing multi-lattice diffraction data. *Acta Crystallographica Section D* **70**, 2652–2666 (2014).

119. Dauter, Z., Dauter, M. & Rajashankar, K. R. Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Crystallographica Section D* **56**, 232–237 (2000).

120. Dauter, M. & Dauter, Z. Phase Determination Using Halide Ions. in *Macromolecular Crystallography Protocols: Volume 2: Structure Determination* (ed. Doublié, S.) 149–158 (Humana Press, 2007). doi:10.1385/1-59745-266-1:149.

121. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* **67**, 271–281 (2011).

122. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235 (2011).

123. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221 (2010).

124. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *urn:issn:0907-4449* **64**, 61–69 (2007).

125. Afonine, P. v *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D* **68**, 352–367 (2012).

126. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

127. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674 (2007).

128. Kantardjieff, K. A. & Rupp, B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci* **12**, 1865–1871 (2003).

129. Zwart, P., Grosse-Kunstleve, R. W. & Adams, P. Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsletter* **43**, (2005).

130. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).

131. New England Biolabs (NEB). Q5® Site-Directed Mutagenesis (E0554) V.2. (2022) doi:dx.doi.org/10.17504/protocols.io.bddfi23n.

132. Liu, J., Yu, H. & Li, S. M. Expanding tryptophan-containing cyclodipeptide synthase spectrum by identification of nine members from Streptomyces strains. *Appl Microbiol Biotechnol* **102**, 4435–4444 (2018).

133. Liao, Y.-D., Jeng, J.-C., Wang, C.-F., Wang, S.-C. & Chang, S.-T. Removal of N-terminal methionine from recombinant proteins by engineered E. coli methionine aminopeptidase . *Protein Science* **13**, 1802–1810 (2004).

134. Maclean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

135. Moutiez, M., Belin, P. & Gondry, M. Aminoacyl-tRNA-Utilizing Enzymes in Natural Product Biosynthesis. (2017) doi:10.1021/acs.chemrev.6b00523.

136. Murakami, H., Ohta, A., Goto, Y., Sako, Y. & Suga, H. Flexizyme as a versatile tRNA acylation catalyst and the application for translation. *Nucleic Acids Symp Ser (Oxf)* 35–36 (2006) doi:10.1093/NASS/NRL018.

137. Sharif, S., Kim, S. J., Labischinski, H. & Schaefer, J. Characterization of Peptidoglycan in Fem-deletion Mutants of Methicillin-resistant Staphylococcus aureus by Solid-State NMR. *Biochemistry* **48**, 3100 (2009).

138. Fonvielle, M. *et al.* Aminoacyl-tRNA recognition by the FemX Wv transferase for bacterial cell wall synthesis. *Nucleic Acids Res* **37**, 1589–1601 (2009).

139. Mainardi, J.-L., Egis Villet, R. ´, Bugg, T. D., Mayer, C. & Arthur, M. Evolution of peptidoglycan biosynthesis under the selective pressure of antibiotics in Gram-positive bacteria. (2008) doi:10.1111/j.1574-6976.2007.00097.x.

140. Banik, S. D. & Nandi, N. Chirality and Protein Biosynthesis. *Top Curr Chem* **333**, 255–305 (2012).

141. Ling, Y., Pang, J., Poruri, K. & Martinis, S. A. tRNA synthetase: tRNA Aminoacylation and beyond. *Wiley Interdiscip Rev RNA* **5**, 461–480 (2014).

142. Hoben, P. & Söll, D. [8] Glutaminyl-tRNA synthetase of Escherichia coli. *Methods Enzymol* **113**, 55–59 (1985).

143. Ledoux, S. & Uhlenbeck, O. C. [3'-32P]-labeling tRNA with nucleotidyltransferase for assaying aminoacylation and peptide bond formation. *Methods* **44**, 74–80 (2008).

144. Hartman, M. C. T., Josephson, K. & Szostak, J. W. Enzymatic aminoacylation of tRNA with unnatural amino acids. *PNAS* **103**, 4356–4361 (2006).

145. Gamper, H. & Hou, Y.-M. A Label-Free Assay for Aminoacylation of tRNA. doi:10.3390/genes11101173.

146. Morimoto, J., Hayashi, Y., Iwasaki, K. & Suga, H. Flexizymes: their evolutionary history and the origin of catalytic function. *Acc Chem Res* **44**, 1359–1368 (2011).

147. Eigner, E. A. & Loftfield, R. B. [48] Kinetic Techniques for the Investigation of Amino Acid: TRNA Ligases (Aminoacyl-tRNA Synthetases, Amino Acid Activating Enzymes). *Methods Enzymol* **29**, 601–619 (1974).

148. Wolfson, A. D. & Uhlenbeck, O. C. Modulation of tRNAAla identity by inorganic pyrophosphatase. *Proc Natl Acad Sci U S A* **99**, 5965–5970 (2002).

149. Ledoux, S. & Uhlenbeck, O. C. 3'-32 P Labeling tRNA with Nucleotidyltransferase for Assaying Aminoacylation and Peptide Bond Formation. *Methods* **44**, 74–80 (2008).

150. Kirshenbaum, K., Carrico, I. S. & Tirrell, D. A. Biosynthesis of Proteins Incorporating a Versatile Set of Phenylalanine Analogues. *ChemBioChem* **02**, 235–237 (2002).

151. Sang-woong, K., Hong, C., Young-gwan, K. & Kim, I. Method for preparing diketopiperazine derivatives using solid phase reaction. (1999).

152. Noe, C. R., Weigand, A. & Pirker, S. Studies on cyclic dipeptides, I: Aryl modifications ofcyclo-[Phe-His]. *Monatshefte für Chemie / Chemical Monthly 1996 127:10* **127**, 1081–1097 (1996).

153. Sheppard, K., Akochy, P. M. & Söll, D. Assays for transfer RNA-dependent amino acid biosynthesis. *Methods* **44**, 139–145 (2008).

154. Suzuki, T. The expanding world of tRNA modifications and their disease relevance. *Nature Reviews Molecular Cell Biology 2021 22:6* **22**, 375–392 (2021).

155. Lorenz, C., Lünse, C. E. & Mörl, M. tRNA Modifications: Impact on Structure and Thermal Adaptation. doi:10.3390/biom7020035.

156. Roovers, M. *et al.* The YqfN protein of Bacillus subtilis is the tRNA: m 1 A22 methyltransferase (TrmK). *Nucleic Acids Res* **36**, 3252–3262 (2008).

157. Sprinzl, M. & Vassilenko, K. S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res* **33**, D139–D140 (2004).

158. Menichi, B., Arnold, H. H., Heyman, T., Dirheimer, G. & Keith, G. Primary structure of Bacillus subtilis tRNAsTyr. *Biochem Biophys Res Commun* **95**, 461–467 (1980).

159. Andachi, Y., Yamao, F., Muto, A. & Osawa, S. Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in Mycoplasma capricolum: Resemblance to mitochondria. *J Mol Biol* **209**, 37–54 (1989).

160. Chaudhuri, R. R. *et al.* Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* **10**, 1–18 (2009).

161. Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A. & Pucci, M. J. Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. *Nucleic Acids Res* **30**, 3152–3162 (2002).

162. Hsiao, K., Zegzouti, H. & Goueli, S. A. Methyltransferase-Glo: A universal, bioluminescent and homogenous assay for monitoring all classes of methyltransferases. *Epigenomics* **8**, 321–339 (2016).

163. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

164. Ahdritz, G. *et al.* OpenFold Software v1.0.1. (2021) doi:10.5281/ZENODO.7352612.

165. Krissinel, E. *et al.* CCP4 Cloud for structure determination and project management in macromolecular crystallography. *Acta Crystallographica D* **78**, 1079–1089 (2022).

166. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73**, 765–783 (2008).

167. Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61**, 704–721 (2005).

168. Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res* **46**, W363–W367 (2018).

169. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **72**, e108 (2020).

170. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237–2243 (2018).

171. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**, 478–480 (1995).

172. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica D* **60**, 2256–2268 (2004).

173. Arnez, J. G. *et al.* Crystal structure of histidyl-tRNA synthetase from Escherichia coli complexed with histidyl-adenylate. *EMBO J* **14**, 4143–4155 (1995).

174. Arnez, J. G., Augustine, J. G., Moras, D. & Francklyn, C. S. The first step of aminoacylation at the atomic level in histidyl-tRNA synthetase. *Proc Natl Acad Sci U S A* **94**, 7144–7149 (1997).

175. Andrews, S. S. & Tretton, J. Physical Principles of Circular Dichroism. *J Chem Educ* **97**, 4370–4376 (2020).

176. Giessen, T. W. & Marahiel, M. A. The tRNA-dependent biosynthesis of modified cyclic dipeptides. *Int J Mol Sci* **15**, 14610–14631 (2014).

177. Kanzaki, H., Yanagisawa, S., Kanoh, K. & Nitoda, T. A novel potent cell cycle inhibitor dehydrophenylahistin-enzymatic synthesis and inhibitory activity toward sea urchin embryo. *J Antibiot (Tokyo)* **55**, 1042–1047 (2002).

178. Kanzaki, H., Imura, D., Nitoda, T. & Kawazu, K. Enzymatic conversion of cyclic dipeptides to dehydro derivatives that inhibit cell division. *J Biosci Bioeng* **90**, 86–89 (2000).

179. Lautru, S., Gondry, M., Genet, R. & Pernodet, J. L. The Albonoursin Gene Cluster of S. noursei: Biosynthesis of Diketopiperazine Metabolites Independent of Nonribosomal Peptide Synthetases. *Chem Biol* **9**, 1355–1364 (2002).

180. Jiang, Y., Yao, T., Geng, Z., Xiao, F. & Li, W. Characterization of the cyclodipeptide synthase gene cluster in Streptomyces sp. NRRL F-5123 by unraveling the biosynthesis of drimentine B. *Tetrahedron* **131**, 133213 (2023).

181. Liu, J., Xie, X. & Li, S. M. Guanitrypmycin Biosynthetic Pathways Imply Cytochrome P450 Mediated Regio- and Stereospecific Guaninyl-Transfer Reactions. *Angewandte Chemie - International Edition* **58**, 11534–11540 (2019).

182. Giessen, T. W., Von Tesmar, A. M. & Marahiel, M. A. Insights into the generation of structural diversity in a tRNA-dependent pathway for highly modified bioactive cyclic dipeptides. *Chem Biol* **20**, 828–838 (2013).

183. Harken, L., Liu, J., Kreuz, O., Berger, R. & Li, S. M. Biosynthesis of Guatrypmethine C Implies Two Different Oxidases for exo Double Bond Installation at the Diketopiperazine Ring. *ACS Catal* **12**, 648–654 (2022).

184. Giessen, T. W. & Marahiel, M. A. Rational and combinatorial tailoring of bioactive cyclic dipeptides. *Front. Microbiol.* **6**, 785 (2015).

185. Mikulski, L., Schäfer, J., Brockmeyer, K., Kraut, R. & Li, S.-M. Comparative studies on similarities and differences of cyclodipeptide oxidases for installation of C–C double bonds at the diketopiperazine ring. *Appl Microbiol Biotechnol* **104**, 2523–2536 (2020).

186. Skinnider, M. A. *et al.* Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res* **43**, 9645–9662 (2015).

187. Skinnider, M. A. *et al.* Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci U S A* **113**, E6343–E6351 (2016).

188. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res* **45**, W49–W54 (2017).

189. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature Communications 2020 11:1* **11**, 1–9 (2020).

190. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**, W339–W346 (2011).

191. Blin, K. *et al.* antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* **41**, W204–W212 (2013).

192. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237–W243 (2015).

193. Blin, K. *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36–W41 (2017).

194. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**, W81–W87 (2019).

195. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* **49**, W29–W35 (2021).

196. Copeland, R. A. *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*. (John Wiley & Sons, 2000).

197. Fu, P. *et al.* α-Pyrones and Diketopiperazine Derivatives from the Marine-Derived Actinomycete Nocardiopsis dassonvillei HR10-5. *J Nat Prod* **74**, 2219–2223 (2011).

198. Zenova, G. M., Manucharova, N. A. & Zvyagintsev, D. G. Extremophilic and extremotolerant actinomycetes in different soil types. *Eurasian Soil Science* **44**, 417–436 (2011).

199. Kuipers, B. J. H. & Gruppen, H. Prediction of Molar Extinction Coefficients of Proteins and Peptides Using UV Absorption of the Constituent Amino Acids at 214 nm To Enable Quantitative Reverse Phase High-Performance Liquid Chromatography−Mass Spectrometry Analysis. *J Agric Food Chem* **55**, 5445–5451 (2007).

200. Fasman, G. D. *Handbook of Biochemistry*. vol. I (CRC Press, 1976).

201. Kar, R. K., Miller, A. F. & Mroginski, M. A. Understanding flavin electronic structure and spectra. *Wiley Interdiscip Rev Comput Mol Sci* **12**, (2022).

202. Hawkins, C. L. & Davies, M. J. Detection and characterisation of radicals in biological materials using EPR methodology. *Biochim Biophys Acta Gen Subj* **1840**, 708–721 (2014).

203. Zhou, X. *et al.* I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nat Protoc* **17**, 2326–2353 (2022).

204. Francis, K., Russell, B. & Gadda, G. Involvement of a Flavosemiquinone in the Enzymatic Oxidation of Nitroalkanes Catalyzed by 2-Nitropropane Dioxygenase. *Journal of Biological Chemistry* **280**, 5195–5204 (2005).

9 - Appendices

Appendix I – aa-DBE ^{1}H NMR spectra

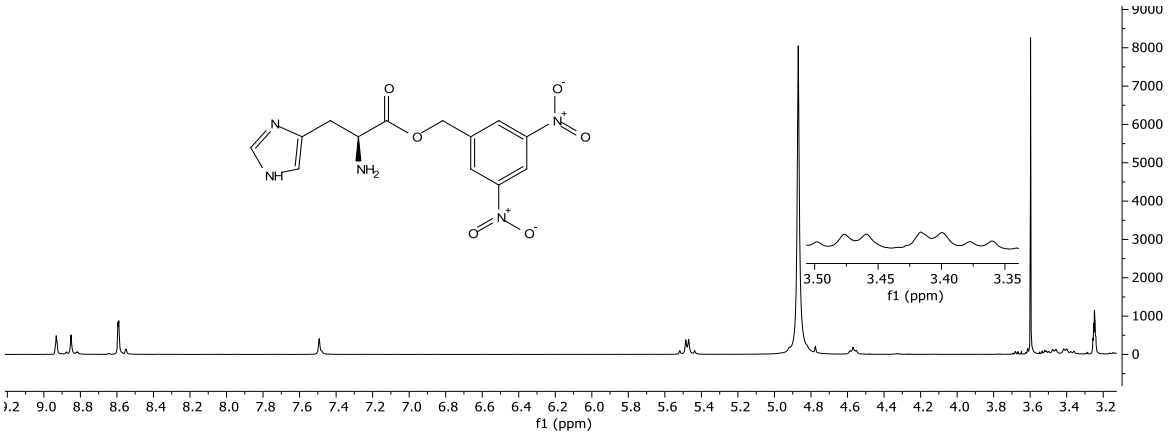


Figure 9.1 ^{1}H NMR (400 MHz, Methanol-d_4) spectrum of His-DBE.

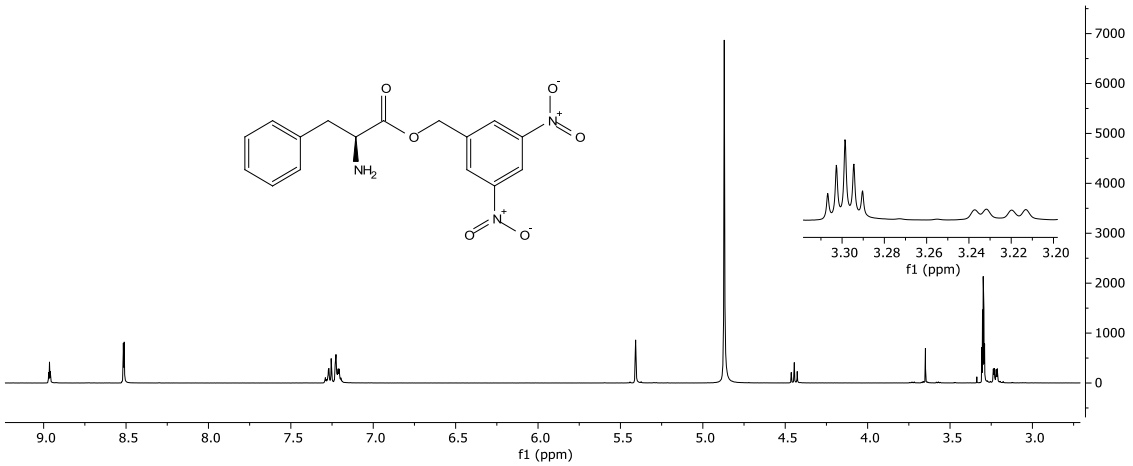


Figure 9.2 ^{1}H NMR (400 MHz, Methanol-d_4) spectrum of Phe-DBE.

Pro-DBE NMR spectrum can be found in the Supplementary Information from Harding *et al*[67].

Appendix II – LC-MS analysis of rejected non-canonical amino acids

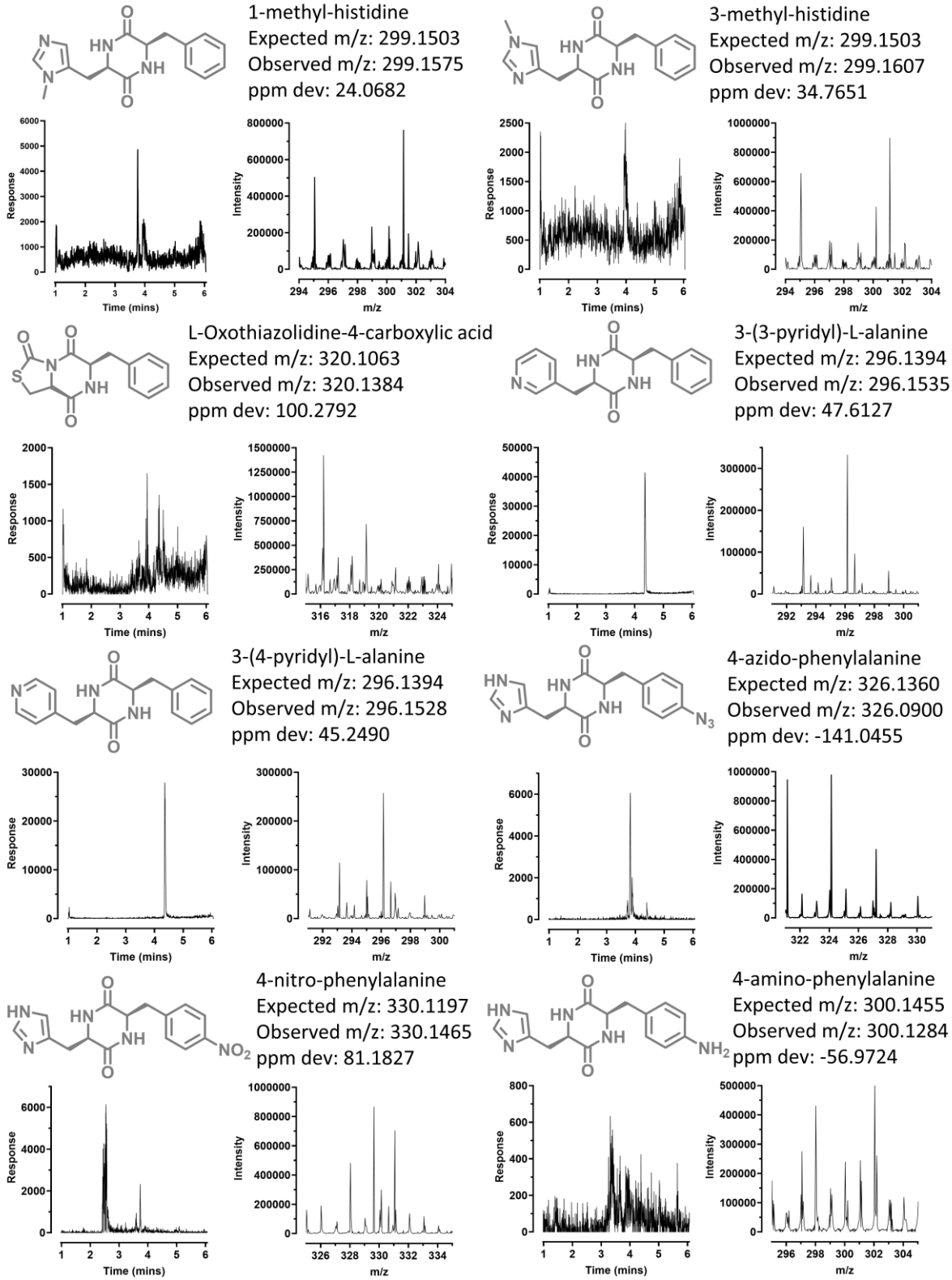
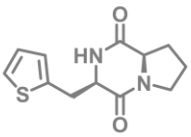
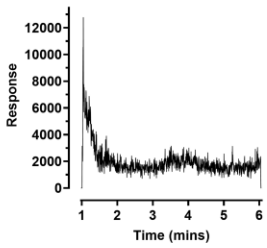
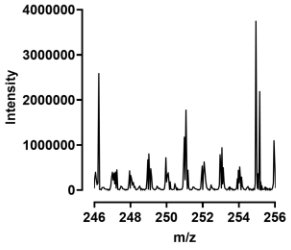


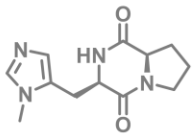
Figure 9.3 MS analysis of CDPs not produced by ParaCDPS.
Mass spectrometry analysis illustrating the cyclodipeptides (drawn in grey) that were not detected from the use of non-canonical amino acids with *Para*CDPS. Both the EIC and mass chromatogram are given with the observed m/z for each.



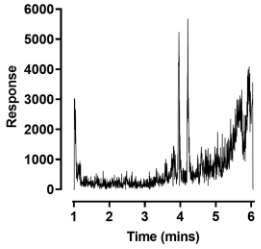
3,2-thienyl-alanine
Expected m/z: 251.0849
Observed m/z: 251.0720
ppm dev: -51.3770

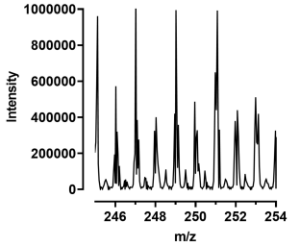


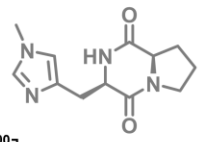




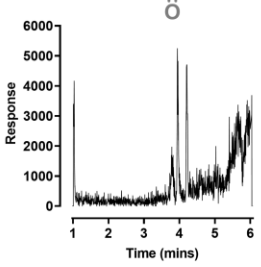
1-methyl-histidine
Expected m/z: 249.1346
Observed m/z: 249.0224
ppm dev: -450.3589

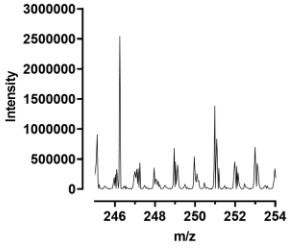


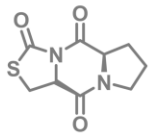




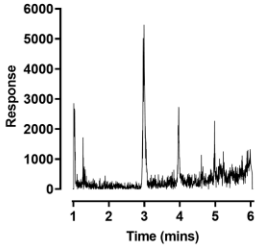
3-methyl-histidine
Expected m/z: 249.1346
Observed m/z: 249.0902
ppm dev: -178.2169

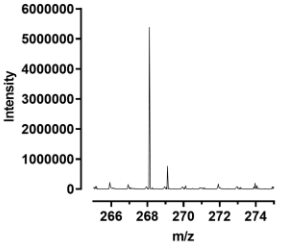


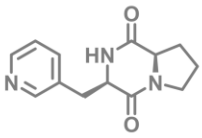




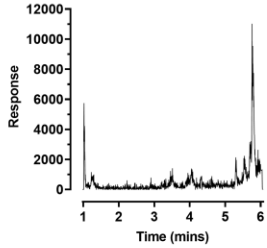
L-Oxothiazolidine-4-carboxylic acid
Expected m/z: 270.0907
Observed m/z: 270.1087
ppm dev: 66.6443

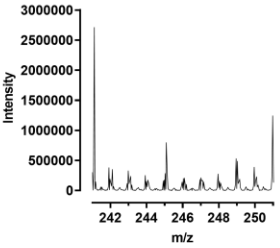


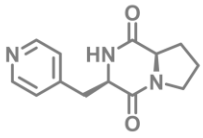




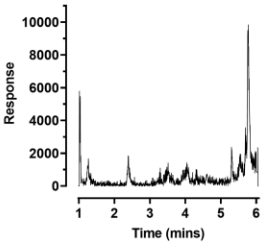
3-(3-pyridyl)-L-alanine
Expected m/z: 246.1237
Observed m/z: 246.0893
ppm dev: -139.7671

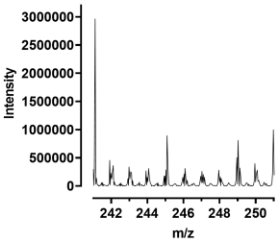


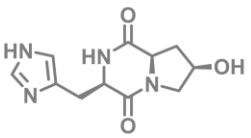




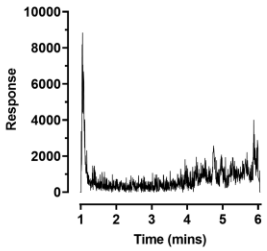
3-(4-pyridyl)-L-alanine
Expected m/z: 246.1237
Observed m/z: 246.0855
ppm dev: -155.2065

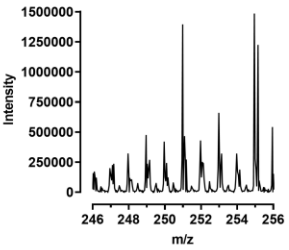


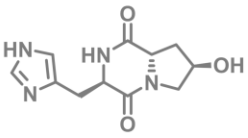




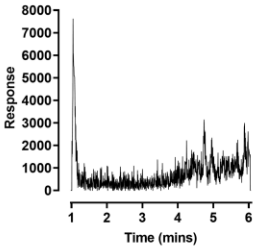
2S,4S-hydroxy-L-proline
Expected m/z: 251.1139
Observed m/z: 251.0856
ppm dev: -112.6979

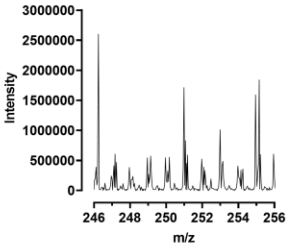






2R,4S-hydroxy-L-proline
Expected m/z: 251.1139
Observed m/z: 251.1245
ppm dev: 42.2119





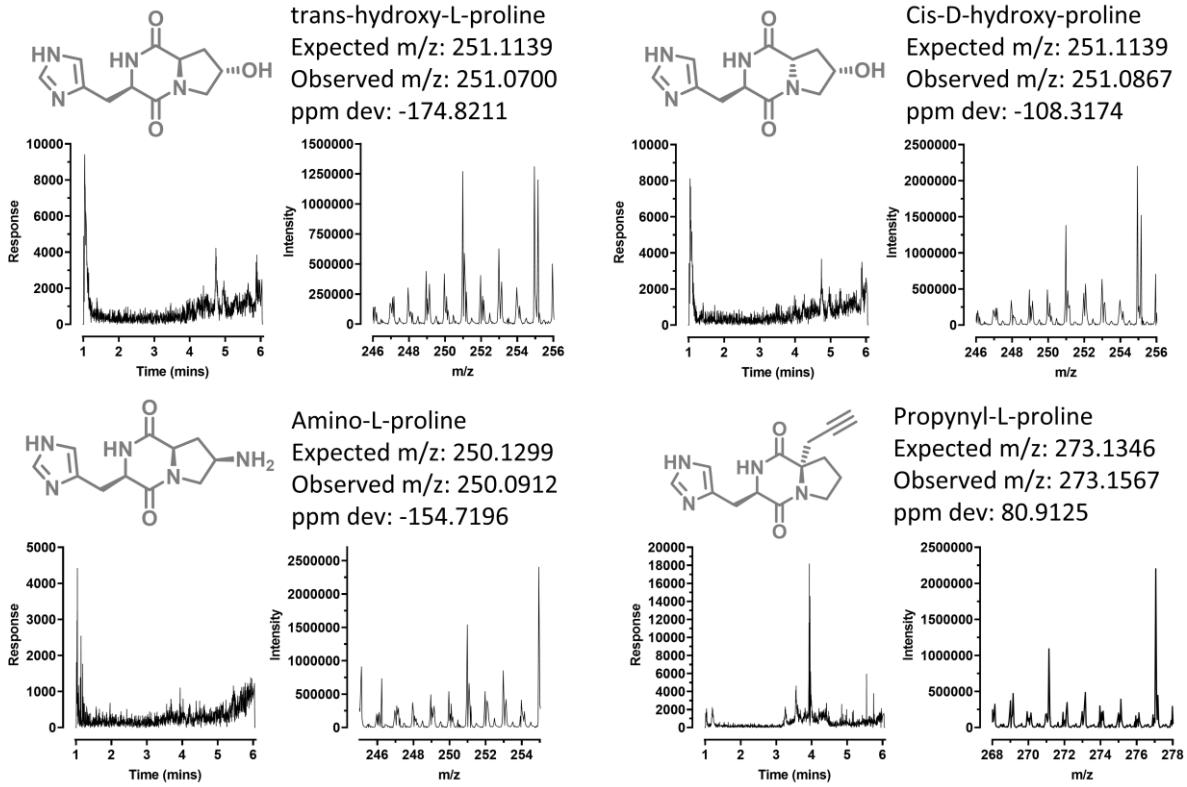


Figure 9.4 MS analysis of CDPs not produced by ParcuCDPS.
Mass spectrometry analysis illustrating the cyclodipeptides (drawn in grey) that were not detected from the use of non-canonical amino acids with *Parcu*CDPS. Both the EIC and mass chromatogram are given with the observed m/z for each.

Appendix III – Intact protein mass of *Parcu*CDPS mutants

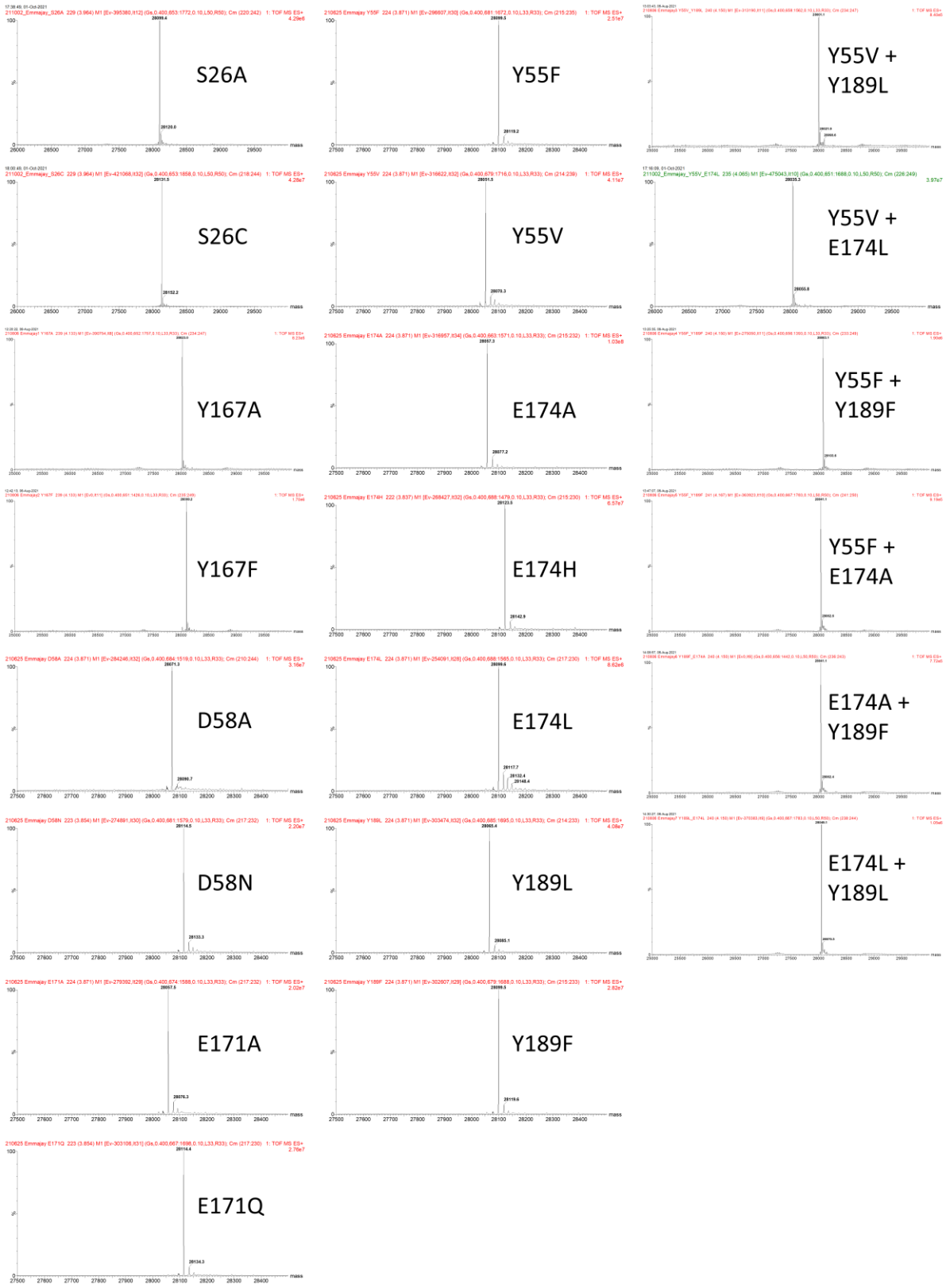


Figure 9.5 Intact protein MS of *Parcu*CDPS mutants

Appendix VI – *Ndas*CDO plasmid map

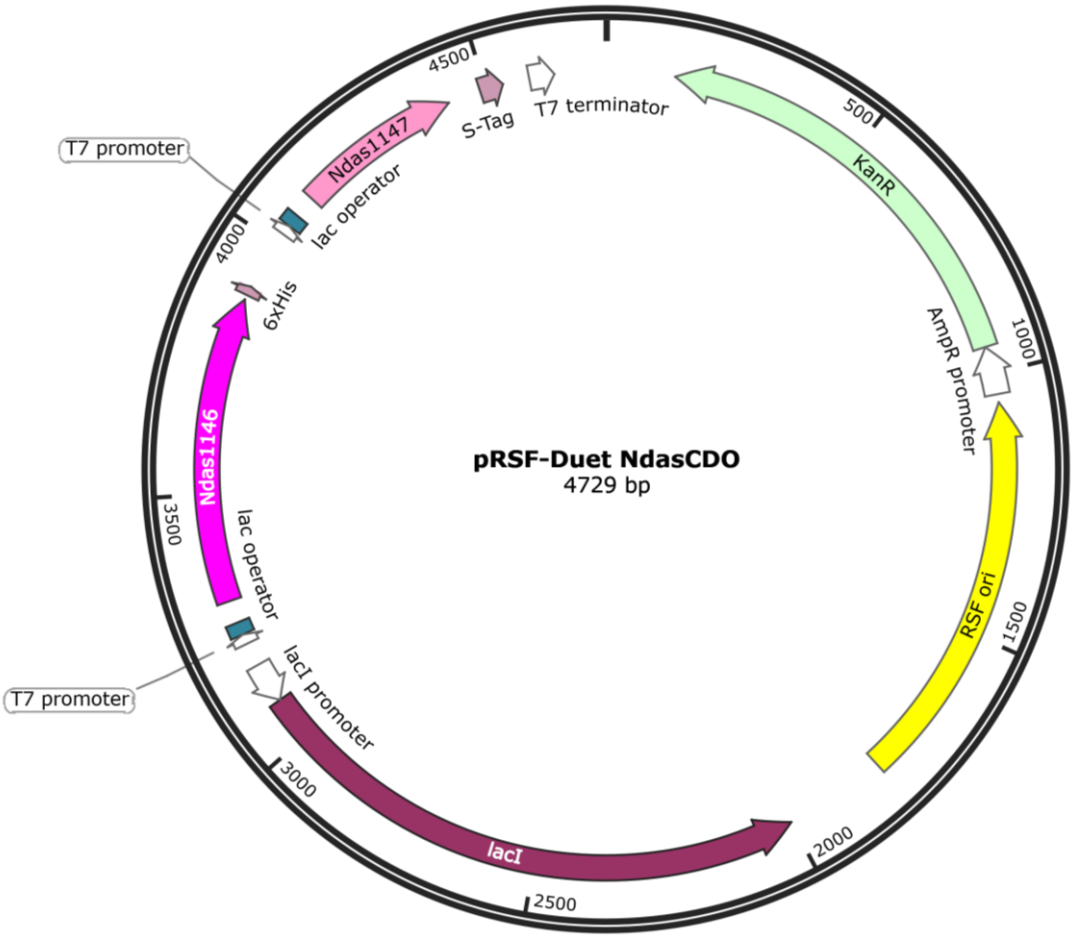


Figure 9.6 Plasmid map of *Ndas*CDO.
This plasmid was purchased from Genscript with the two subunits of *Ndas*CDO (*Ndas*1146 and *Ndas*1147) cloned into pRSF-Duet.

Appendix V – Peptide mapping of *Ndas*CDO

***Ndas*1146**

Protein sequence coverage: 49%

Matched peptides shown in ***bold red***.

```
  1 MDTGSSEPDA NRCPSQRSSH ALQTLTTRRA VRAFADRPVD DSLLDPMLDA
 51 MLAAPSASNK QAWAFVAVRE RRALRLLRAF SPGIIELPPL VVAACFDRSR
101 AVGGSGNSTD SGDSWDEGML CVAMAVENLL LAAHCLGLGG CPSGSFRRGP
151 VRRLLGLPDH LEPLLLVPIG HPARPLAPAP RRDRNEVVSH ERWGTEHHHH
201 HH
```

***Ndas*1147**

Protein sequence coverage: 70%

Matched peptides shown in ***bold red***.

```
  1 MSAGEPEVRQ VGEELLLLAA YLLSSGRGLL DEPRQYGTFR CLDAARRVLA
 51 LAAGTGPHHP ELDALRGRMD DVMCGPMGDH ELDTLLDQMC ERLATVLEDP
101 DVISD
```

Figure 9.7 MS/MS results of *Ndas*CDO following a trypsin digest.
Peptide fragments contain the highest similarity to sequences of both *Ndas* subunits (1146 and 1147).