Structures of linear gramicidin synthetase reveal its elegant synthetic cycle

Janice M. Reimer
Department of Biochemistry
McGill University
Montréal, Québec, Canada
July 2018

A thesis submitted to McGill University in partial fulfilment
of the requirements for the degree of Doctor of Philosophy.

As long as you believe what you're doing is meaningful, you can cut through fear and exhaustion and take the next step.
- Arlene Blum

Dedicated to my Oma, Maria Reimer, for her perseverance and dedication to her dream of having her children and her children's children go to school and learn.

Abstract

Nonribosomal peptide synthetases (NRPSs) are large enzymes that synthesize diverse secondary metabolites ranging from antibiotics to industrial solvents. They are arranged as an assembly line of modules where each module is responsible for incorporating one specific substrate, or "building block," into the final nonribosomal peptide (NRP), through the action of 3 domains. In canonical NRPSs, substrates are selected and activated by the adenylation (A) domain and then transferred onto the peptidyl carrier protein (PCP) domain. The PCP domain transports them to the condensation (C) domain for incorporation into the nascent peptide. NRPSs may also include specialized tailoring domains to further modify the NRP.

To determine how a tailoring domain is adopted into the architecture and synthetic logic of an NRPS, I determined 4 structures of the initiation module of the linear gramicidin synthetase subunit LgrA and performed accompanying small angle X-ray scattering and bioinformatics analysis. The module contains an A domain, a PCP domain and a tailoring formylation (F) domain. The structures reveal every major conformation required in the synthetic cycle of the initiation module with large conformational changes to transport substrate between active sites. This was the first time an initiation module has been solved, the first time a single NRPS has been visualized in so many of its conformational states, and the first time any tailoring domain has been determined within its NRPS.

Tailoring domains are acquired by NRPSs through horizontal gene transfer, and our bioinformatics identified the F domain to originate from sugar-nucleotide formyltransferases (FTs). To understand the adaptions that were needed to co-opt and evolve a FT into a functional and useful NRPS domain, I characterized PseFT, a homologous sugar FT found in *Anoxybacillus kamchatkensis*, which represents the pre-transfer FT prior to gene fusion with the NRPS. PseFT belongs to a novel biosynthetic pathway for CMP-7-formamidopseudaminic acid, a sugar-nucleotide used in glycosylation of flagellin. I solved 4 crystal structures of PseFT alone and in complex with substrates, which reveal substantial contrasts to other studied sugar FTs in substrate binding and architecture.

Lastly, I have solved the structure of dimodular LgrA, as well as 6 truncated constructs of LgrA that reveal the flexible and dynamic modular organization of NRPSs. This is the first dimodular structure of an NRPS and reveals the novel condensation state with both donor and acceptor PCP domains docked at the C domain. The structures show the initiation and elongation modules in multiple orientations relative to each other at several stages of the catalytic cycle. Together, I illustrate that NRPSs are unlikely to adopt any sort of super-modular architecture to complete their complex synthetic cycle.

Resumé

Les enzymes de synthèse de peptides non ribosomiques (NRPS) sont de larges protéines qui produisent divers métabolites secondaires allant des antibiotiques aux solvants industriels. Elles sont organisées en tant que chaînes de montage composées de modules, chacun d'entre eux incorporant un substrat spécifique – ou composante de base – dans le peptide non ribosomique (NRP). Chaque module est constitué de domaines; dans les NRPS classiques, les substrats sont sélectionnés et activés par le domaine d'adénylation (A), puis transférés au domaine porteur de peptide (PCP). Ce dernier les transporte vers le domaine de condensation (C), qui les incorpore alors dans le peptide naissant. Ces enzymes contiennent aussi parfois des domaines d'adaptation pour modifier les NRP qu'elles produisent.

En vue d'élucider la manière dont le domaine d'adaptation est intégré dans l'architecture et dans le procédé de synthèse des NRPS, j'ai déterminé quatre structures cristallines du module d'initiation de l'enzyme synthétisant la gramicidine linéaire (LgrA), et j'ai caractérisé cette protéine par la diffusion radiologique à petit angle et par des analyses bioinformatiques. Ce module est composé des domaines A et PCP ainsi que d'un domaine d'adaptation – le domaine de formylation (F). Ces structures ont révélé chaque conformation principale présente lors du cycle de synthèse du module d'initiation, incluant de larges changements de conformation nécessaires au transport du substrat entre les sites actifs. Ceci représente la première structure d'un module d'initiation élucidée, et la première NRPS visualisée dans autant d'états. De plus, aucune structure d'un domaine d'adaptation dans le contexte d'un module de NRPS intact n'a été déterminée auparavant.

Les NRPS acquièrent leurs domaines d'adaptation par le transfert horizontal de gènes, et notre analyse bioinformatique a établi que le domaine F est originaire des nucléotide-ose formyltransférases (FT). Afin de comprendre les changements nécessaires à l'incorporation et l'évolution d'une FT en tant que domaine fonctionnel dans une NRPS, j'ai caractérisé la protéine PseFT, une FT homologue provenant d'*Anoxybacillus kamchatkensis* et représentant le domaine F avant le transfert et la fusion avec le gène de la NRPS. PseFT est membre d'une nouvelle voie de biosynthèse de l'acide CMP-7-formamidopseudaminque, un nucléotide-ose employé dans la glycosylation de la flagelline. J'ai déterminé quatre structures cristallines de PseFT avec ou sans substrats, qui ont révélé des différences marquées dans son arrangement et sa liaison aux substrats, par rapport à d'autres FT.

Enfin, j'ai déterminé la structure de LgrA bimodulaire ainsi que de cinq constructions tronquées de cette enzyme, qui démontrent l'organisation flexible et dynamique des NRPS. Ceci représente la première structure de NRPS bimodulaire et révèle un nouvel état de condensation où les domaines PCP donateur et accepteur sont simultanément arrimés au domaine C. Ces structures illustrent les diverses orientations des modules d'initiation et d'élongation lors des différentes étapes du cycle catalytique. Dans leur ensemble, mes résultats suggèrent que les NRPS ne forment probablement pas d'arrangement supra-modulaire lors de leur complexe cycle de synthèse.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

A domain – adenylation domain
A_{core} – N-terminal subdomain of the A domain
ACP – acyl carrier proteins
AMP – adenosine monophosphate
AMPcPP – α,β-methyleneadeonsine 5'triphosphate
A_{sub} – C-terminal subdomain of the A domain
AT – acyltransferase domain
ATP – adenosine triphosphate
AVS – adenosine vinylsulfonamide inhibitor
C domain – condensation domain
CAT – chloramphenicol acetyltransferase
CMP-Pse5Ac7Fo – CMP-5-N-acetyl-7-N-formyl-pseudaminic acid
CoA – Coenzyme A
CTD – C-terminal domain
Cy domain – cyclization domain
E domain – epimerization domain
F domain – formylation domain
FAS – fatty acid synthase
FMT – formyl methionine tRNA formyltransferase
FT – formyltransferase
Grs – gramicidin (Soviet A) synthetase
KS – ketosynthase domain
Lgr – linear gramicidin synthetase
MLP – MbtH-like protein
MT domain – methyltransferase domain
N10-fTHF – N10-tetrahydrofolate
N5-fTHF – folinic acid; N5-tetrahydrofolate
NH-CoA – amino-CoA
NRP – nonribosomal peptide
NRPS – nonribosomal peptide synthetase
PCP domain – peptidyl carrier protein domain
PKS – polyketide synthase
PPE – phosphopantetheine
PPTase – phosphopantetheinyl transferase
R domain – reductase domain
Sfp – phoshopantetheinyl transferase from *Bacillus subtilis*
Smallecule – small molecule
T domain – thiolation domain
Te domain – thioesterase domain
TGH – transglutaminase homologue
THF – tetrahydrofolate

Preface

This is a manuscript-based thesis comprising of one published article, one article currently under review and one manuscript in preparation. The introduction (Chapter 1) and General Conclusions (Chapter 5) were partially adapted from:

Reimer JM*, Haque AS*, Tarry MJ*, Schmeing TM (2018) Piecing Together Nonribosomal Peptide Synthesis. Curr Opin Struct Biol, 49: 104-113.

Chapter 2

Reimer JM*, Aloise MN*, Harrison PM, Schmeing TM (2016). Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. Nature, 529 (7585):239-242.

Chapter 3

Reimer JM*, Harb I*, Ovchinnikova OG*, Jiang J, Whitfield C, Schmeing TM. Structural insight into a novel formyltransferase and evolution to a nonribosomal peptide synthetase tailoring domain. *Under review at Cell Chemical Biology.*

Chapter 4

Reimer JM, Harb I, Eivaskhani M, Schmeing TM. Structures of a dimodular nonribosomal peptide protein. *Manuscript in preparation.*

*denotes co-author

Contributions of Authors

Chapter 2

Together with Martin Aloise, we equally performed cloning, protein purification, crystallization and structure determinations. I performed the CoA syntheses and prepared figures. Paul Harrison conducted the bioinformatics analyses. T. Martin Schmeing designed the experiment and wrote the manuscript with input from other authors.

Chapter 3

Jessie Jiang initially cloned, purified and crystallized PseFT. I solved the initial structure of apo PseFT, and conducted soaking experiments and structure determination of ligand-soaked structures of PseFT with Ingrid Harb. Ingrid cloned and purified PseB and PseC. Olga G. Ovchinnikova performed bioinformatics analyses, LCMS assays and small molecule identification experiments. T. Martin Schmeing and I designed the experiments and wrote the manuscript together with input from other authors.

Chapter 4

I performed all experiments relating to F_1-A_1-PCP_1-C_2. I solved and built all the presented structures. Ingrid Harb provided significant help in cloning, purification and crystallization of F_1-A_1-PCP_1-C_2-A_2 and F_1-A_1-PCP_1-C_2-A_2-PCP_2 proteins. T. Martin Schmeing and I designed the experiments and wrote the manuscript with input from other authors.

Original Contributions of Knowledge

Chapter 2 | Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase

- Four independent crystal structures were determined of the initiation module of the linear gramicidin synthetase, with accompanying small angle X-ray scattering and bioinformatics analyses.
- This is the first structure of an initiation module that has been solved, the first time a single NRPS has been visualized in so many of its conformational states, and the first time the structure of an *in cis* tailoring domain has been determined.
- I show how a tailoring domain has been adopted into the architecture and synthetic cycle of an NRPS.

Chapter 3 | Structural insight into a formyltransferase involved in 7-N-formylpseudaminic acid synthesis and evolution to a nonribosomal peptide synthetase tailoring domain

- We identified and characterized a novel formyltransferase, PseFT, in a newly described pathway for the synthesis of CMP-7-formamidopseudaminic acid, used for glycosylation of flagellin. PseFT is representative of the pre-transfer formyltransferase prior to gene fusion with the linear gramicidin NRPS.
- I solved structures of PseFT alone, with cofactor, substrate and products. These structures reveal key insights into the adaptions that were needed to co-opt and evolve a sugar formyltransferase into a functional and useful NRPS domain.

Chapter 4 | Structures of a dimodular nonribosomal peptide synthetase protein

- I solved the first complete dimodular crystal structure of an NRPS, which shows the previously unobserved condensation state.
- I determined four additional crystal structures of shorter dimodular LgrA constructs that together show 6 different conformations of the protein. These structures show that NRPSs have a flexible and dynamic modular organization, and do not adopt any super-modular architecture.

Acknowledgements

Where do you start to thank people when you've been surrounded by an incredible support system of colleagues, fellow grad students, friends and family for almost seven years? Well, at the beginning of course!

I would like to sincerely thank my supervisor, Dr. Martin Schmeing, for taking me on as Grad Student #3 in the lab. Your mentorship has been pivotal to both the success of my research and my own personal development as a scientist. Also, thank you for teaching me the importance of having a side-project! The LgrA project has been quite the ride.

Completing a PhD in science is not a solo endeavour, but one that requires the finest of lab mates by your side. Thank you to all the past and present members of the Schmeing lab for your constant support, friendship and comradery. To Kris Bloudoff and Diego Alonzo - you quickly became my science family – thank you for always having my back, in and outside of the lab. To Ingrid Harb – you were the finest minion-turned-fellow-grad-student a grad student could ever ask for; thank you for your willingness to help. To 'Camster' Camille Fortinez – I will take 'le noms' with me and spread it to my future labs. To Maximillian Eivaskhani – thanks for joining the craziness that is Team LgrA. To Martin Aloise i.e. the original member of Team LgrA – thank you for your hard work and dedication to the project. Thank you to Michael Tarry, Asfarul Haque, Frederik Hansen, Clarisse Chiche-Lapierre and Itai Sharon for the many helpful scientific discussions and advice. Thank you to Alexei Gorelik for translating the abstract into French. I would also like to thank the 4th Floor Bellini crew, specifically Yazan Abbas, Pat Kim Chiaw, Jon Labriola, Shane Caldwell and Juliana Munoz for all the lab shenanigans and help over the years.

Many additional people have provided invaluable help along the way and deserve credit. Drs. Bhushan Nagar and Karine Auclair, for serving on my Research Advisory Committee and contributing constructive and beneficial advice. The wonderful chemists of Zamboni Chemical Solutions, with Honourable Mentions going to John Colucci and Mika Guerard – thank you for always opening your fume hoods to me and all my invisible molecules. Kurt Dejgaard and Alexander Wahba, for helping carry out mass spectrometry experiments. Shaun Labiuk and the entire CMCF team at the Canadian Light Source, for being indispensable in data collection. The

CCP4 school, specifically Nukri Sanishvili, at the Argonne National Laboratory for giving me a solid foundation in crystallography.

I am indebted to all the generous funding sources that have supported me over the years and have made my research possible: the McGill University CIHR-Chemical Biology scholarship, the J.P. Collip Fellowship, the Maysie MacSporran Graduate studentship, the GRASP student award, and the NSERC Graham Alexander Bell Canada graduate scholarship.

I wish I had adequate words to express how grateful I am to my entire family. Without their constant support and encouragement, I would not have had the courage or determination to fulfil this dream. To my parents, Heinz and Marlene, thank you for believing that I could do this even when I had one too many failed experiments saying otherwise.

Finally, to Dan. You quickly learned that being with a scientist wasn't all liquid nitrogen tricks and shooting X-rays. Thank you for your unwavering support, for learning that when I say things like 'I solved the structure of the fattest of cats!" means it's time to celebrate, and for conquering the elephant with me one bite at a time.

CHAPTER 1 | INTRODUCTION TO NONRIBOSOMAL PEPTIDE SYNTHETASES

1.1 Nature's assembly line

Nonribosomal peptide synthetases (NRPSs) are a family of microbial megaenzymes that produce natural products that are useful to society as therapeutics (antibiotics, antivirals, antitumours, and immunosuppressants) and green chemicals (agricultural agents, emulsifiers, siderophores, and research tools) (Figure 1.1) (Felnagle et al., 2008). Found in bacteria and fungi, genes for NRPSs are commonly organized in biosynthetic clusters, and are flanked by genes for resistance and transport of the secondary metabolite (Wang et al., 2014a). NRPSs typically synthesize their products through amide bond formation between aminoacyl (or other acyl) monomers. Their architecture is unrelated to the more famous peptide maker, the ribosome.

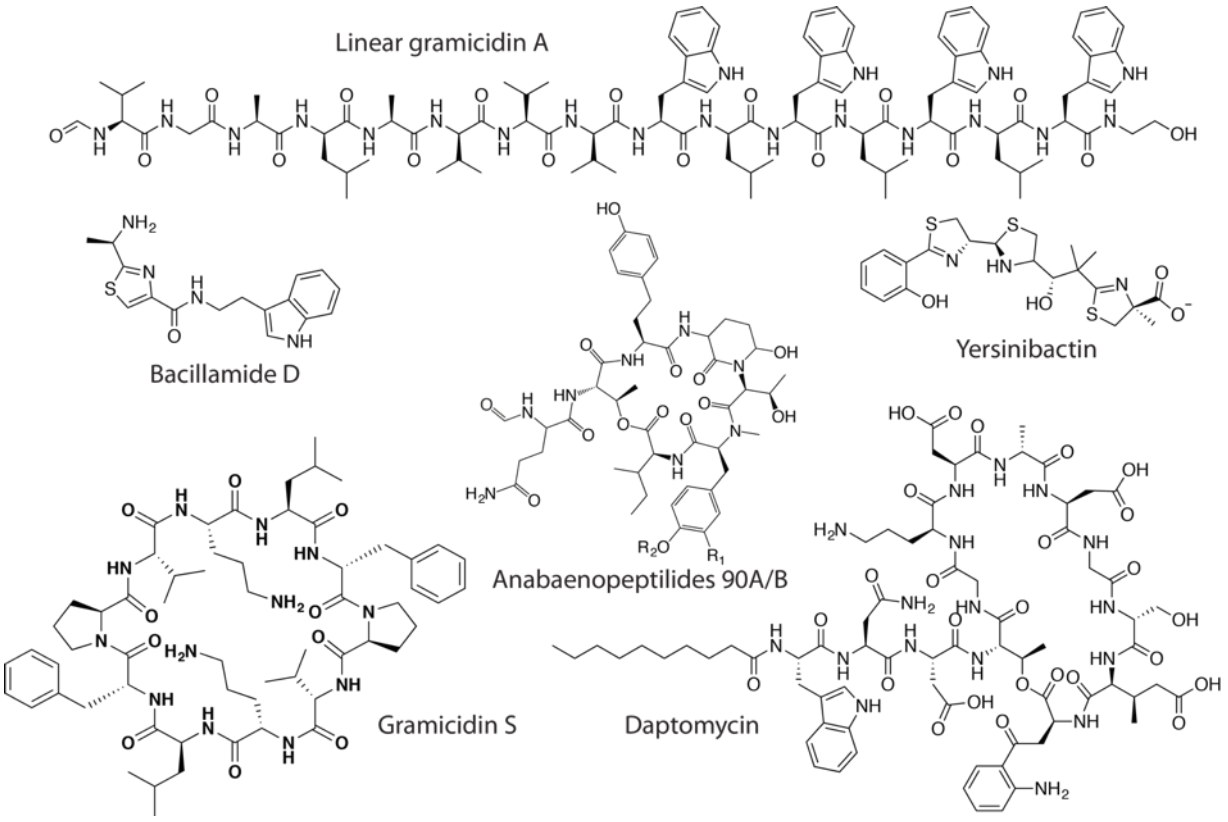


Figure 1.1 | Nonribosomal peptides.
Some examples of nonribosomal peptides: linear gramicidin A (topical antibiotic), bacillamide D (anti-algael), anabaenopeptilides 90A/B (micocystin), yersinibactin (siderophore), gramicidin S (topical antibiotic and spermicide), daptomycin (antibiotic, trade name Cubicin).

Whereas ribosomes use the same active sites for each amino acid added to the ribosomal peptide, NRPSs usually employ a dedicated set of enzyme domains for each amino acid added to the nonribosomal peptide. This set of domains is termed a module, and the synthetic strategy dictates that, normally, the number and specificity of the modules correspond to the length and sequence of amino acids in the peptide product. NRPSs can consist of a single polypeptide of between 1 and 18 modules, with a mass of ~220 kDa – 2.2 MDa, or be split over multiple proteins that assemble non-covalently (Walsh, 2004, Weissman, 2015, Schwarzer et al., 2003).

Within a module, the domains work together to incorporate the incoming amino acid into the growing peptide (Weissman, 2015, Gulick, 2016). A basic elongation module contains three core domains: a condensation (C) domain, an adenylation (A) domain and a peptidyl carrier protein (PCP) domain (Figure 1.2B). The A domain selects and adenylates the cognate amino acid, then attaches it by a thioester link to a prosthetic phosphopantetheinyl (PPE) group on the PCP domain. The PCP domain transports the amino acid to the C domain, which catalyzes amide bond formation between this amino acid and the peptide attached to the PCP domain of the preceding module, elongating the peptide by a single residue. Next, the PCP domain brings the elongated peptide to the downstream module, where it is passed off and further elongated in the next condensation reaction. Once a PCP domain has donated its peptide, it can accept a new amino acid from the A

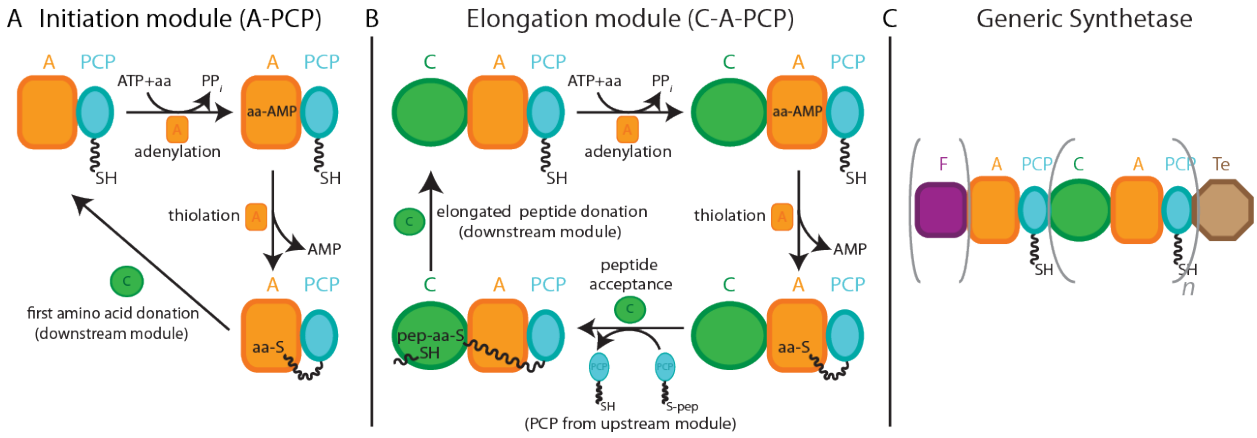


Figure 1.2 | NRPS schematic.
A schematic diagram of the initiation (**A**) and elongation (**B**) cycles of a canonical initiation and elongation module, respectively. Aa, amino acid. **C**, A generic synthetase: (F)-A-PCP-(C-A-PCP)$_n$-Te. F domain, formylation domain (Note that most NRPSs do not contain an F domain, the tailoring domain which formylates the N-terminal amino acid. It has been included as the NRPS under investigation in this thesis contains an F domain).

domain and participate in the next cycle of assembly-line synthesis. Initiation modules lack the C domain, with minimal initiation modules containing only A and PCP domains (Figure 1.2A). However, it is not uncommon for NRPSs to start with more complex initiation modules. Lipopeptides synthetases for peptides like daptomycin contain starter C domains that use acyl-PCPs as donor substrates; depsipeptide synthetases for valinomycin and cereulide contain ketoreductase domains; and kolossin A and linear gramicidin synthetases contain formylation (F) domains. Termination modules usually contain a thioesterase (Te) domain, which releases the peptide by cyclization or hydrolysis. A canonical organization of a basic NRPS is A-PCP-$(C\text{-}A\text{-}PCP)_n$-Te where *n* denotes the number of elongation modules in the synthetase (Figure 1.2C). Additionally, NRPS modules usually have tailoring domains, and the action of these domains is incorporated into the catalytic cycle of the module (Sundaram and Hertweck, 2016). NRPSs can alternatively end in a reductase (Gahloth et al., 2017) or terminal C domain (Zhang et al., 2016, Bloudoff et al., 2017). This wide range of tailoring domains, combined with the over five hundred monomers that can be used as substrates, including D-amino acids, aryl acids, hydroxy acids, and fatty acids, allows nonribosomal peptides to occupy a diverse area of chemical space (Caboche et al., 2008).

1.2 Structure and function of core domains

1.2.1 The adenylation domain

The A domain is the most well-characterized NRPS domain and is part of the ANL superfamily of adenylating enzymes, including firefly luciferases and acyl-CoA ligases (ANL – acyl-CoA synthetases, NRPS adenylation domains, luciferase enzymes) whose reaction cycle proceeds through an adenylate intermediate. The A domain from the gramicidin S synthetase[1] (GrsA) initiation module was the first structure of both an A domain and an NRPS domain, and was solved in complex with its substrate phenylalanine and adenosine monophosphate (AMP) (Conti et al., 1997). This structure revealed not only how A domains bind their cognate substrates, but using

[1] Note: The gramicidin S synthetase is discussed multiple times throughout Chapter 1. The synthetase and its product, gramicidin S, are distinct (Figure 1.1) and unrelated to the linear gramicidin synthetase.

GrsA as a template, sequence alignments were used to establish a ~10 amino acid "specificity determining code" that predicts the amino acid substrate for any bacterial A domain (Stachelhaus et al., 1999). Since then, the code has been refined by including phylogenetic information and can accurately predict the A domain specificity for over 30 acyl monomers (Challis et al., 2000). The emergence of the specificity code has made a significant impact on the NRPS field as products from uncharacterized NRPSs can be dependably predicted and the code can be used to bioengineer an A domain to accept non-cognate substrates (Eppelmann et al., 2002, Thirlway et al., 2012), thus creating a novel nonribosomal peptide product.

The A domain has a large A_{core} (~450 amino acids) portion with an active site for binding ATP and substrate amino acid, and a small A_{sub} (~100 amino acids) portion that changes position depending on functional state and provides catalytic residues to the adenylation reaction (Figure 1.3A) (Gulick, 2009, Conti et al., 1997, May et al., 2002, Gulick et al., 2003, Du et al., 2008, Yonus et al., 2008). (A_{core} and A_{sub} are also called the large/N-terminal and small/C-terminal subdomains (Gulick, 2017).) A domains share a common fold with ten conserved signature sequences: elements a1-a7 are assigned to the A_{core} and a8-a10 to the A_{sub} domain (Schwarzer et al., 2003). The adenylation cycle proceeds through a two-step reaction involving adenylation and

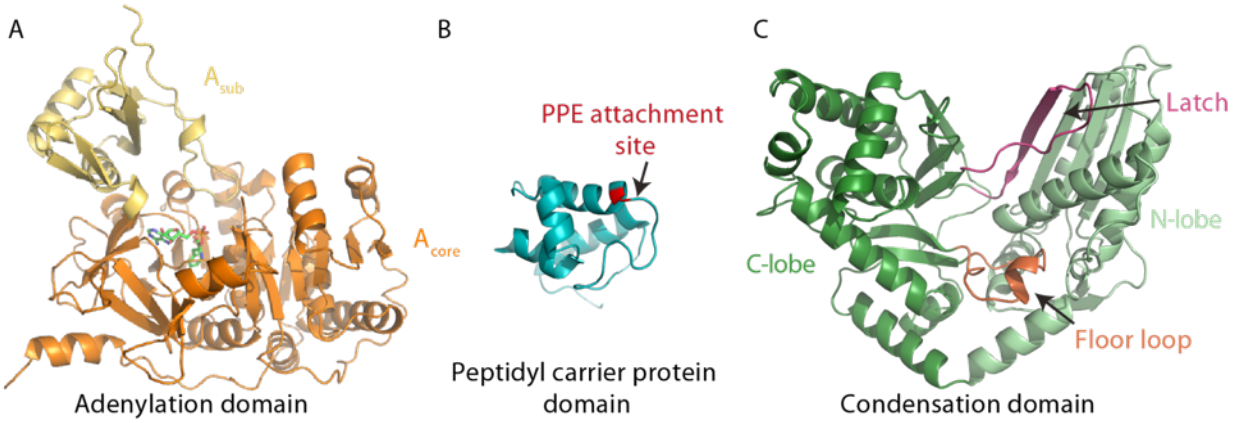


Figure 1.3 | Canonical core domains of an NRPS.
A, Adenylation (A) domain from the gramicidin S synthetase shown with substrates (PDB 1AMU) (Conti et al., 1997). A_{core} and A_{sub} are coloured orange and yellow-orange, respectively. **B**, The NMR structure of the peptidyl carrier protein from the tyrocidine synthetase 3 (TycC-PCP_3) (PDB 1DNY) (Weber et al., 2000). The serine modified with the PPE arm is shown in red. **C**, The VibH condensation domain from the vibriobactin synthetase (PDB 1L5A) (Keating et al., 2002). The C domain forms a pseudo-dimer of two lobes. The latch and floor loop are shown in raspberry and orange, respectively.

thioesterification, and was initially dissected using a combination of structures from the ANL family (Gulick, 2009). The A domain starts in the "open" conformation where the A_{sub} does not make significant contact with the A_{core} and is oriented away from the active site to allow binding of amino acid (or acyl monomer) and ATP (Conti et al., 1996). Following binding, the A_{sub} rotates 30° to adopt the "closed" conformation, enclosing the A domain active to catalyze amino acid adenylation using a conserved lysine to coordinate the α-phosphate of ATP. This reaction triggers a ~140° rotation of the A_{sub} to allow release of pyrophosphate, PCP domain binding and transfer of the amino acid to the PPE arm (Reger et al., 2008, Yonus et al., 2008). With the substrate tethered to the PPE arm, the aminoacyl-PCP moves to the C domain so that substrate can be integrated into the growing nonribosomal peptide.

1.2.2 The peptidyl carrier protein domain

Substrates and the nascent peptide are transported between each active site in the NRPS through the essential action of the PCP domain. This domain is similar to acyl carrier proteins (ACPs) found in the fatty acid synthase (FAS) and polyketide synthases (PKSs), and is also referred to as the thiolation (T) domain. The PCP is the smallest NRPS domain (~90 amino acids) and like its homologues, is composed of a 4-helix bundle (Figure 1.3B) (Weber et al., 2000). Apo PCP is inactive and must be post-translationally modified by a phoshopantetheinyl transferase (PPTase) to convert it to its active holo form. The PPTase uses coenzyme A (CoA) to transfer 4'-phosphopantetheine (PPE) onto a conserved serine (Lambalot et al., 1996) found at the end of α-helix 2. Aminoacyl and peptidyl intermediates are tethered to the end of the PPE arm through a thioester linkage between the residue and the PPE thiol group. Some NRPS clusters contain a dedicated proof-reading Type II thioesterase domain to prevent mispriming of the PPE arm that can arise from either promiscuous activity of the PPTase using acyl-CoA, instead of CoA, or the A domain activating and thiolating a non-cognate substrate (Schwarzer et al., 2002, Leduc et al., 2007). The Type II thioesterase domain acts *in trans* and hydrolyzes the improper acyl group from the PPE arm.

During the synthetic cycle, the PCP domain of an elongation module is involved in a minimum of three catalytic stages, thiolation, substrate acceptance and substrate donation, and

interacts with at least 3 different domains (Figure 1.2B). The PCP_n first binds the A_n domain during the thiolation state for loading of the PPE arm. Aminoacyl-PPE-PCP_n then binds the acceptor site of the C_n domain (acting as the "acceptor PCP") for transfer of growing peptide attached to PCP_{n-1}. Following condensation, peptidyl-PPE-PCP_n shuttles the peptide to the downstream module, binding the C_{n+1} domain at the donor site (acting as the "donor PCP"). The number of binding partners increases with each additional tailoring and/or thioesterase domain found in the module. Despite the restricted surface area imparted by the small size of the domain, PCP domains make productive and specific transient interactions with each binding partner.

1.2.3 The condensation domain

The C domain is a ~450 amino acid, V-shaped pseudo-dimer of chloramphenicol acetyltransferase (CAT) folds, with an active site at the middle of a tunnel connecting binding sites for donor and acceptor PCP domains (Figure 1.3C, Figure 1.4). The pseudo-dimer is composed of an N-terminal lobe (N-lobe) and C-terminal lobe (C-lobe), and both lobes contain a core β-sheet surrounded by peripheral α-helices. There are two crossover areas between lobes, the 'latch' and the 'floor loop.' The latch extends from the C-lobe, and provides a single β-strand to the N-lobe β-sheet before crossing back over to the C-lobe (Bloudoff et al., 2013), enclosing the top of the active site. The role of the latch has not been determined. It has been suggested that the latch may move to accommodate the incoming peptide attached to the donor PCP (Samel et al., 2007), however, normal mode and molecular dynamics analyses using calcium-dependent antibiotic (CDA) synthetase C_1 domain indicate that the latch interactions are maintained (Bloudoff et al., 2013). The floor loop is a small α-helix that reaches out from the C-lobe to the N-lobe. An active site tunnel ~ 30 Å in length, corresponding to the length of approximately two PPE arms, is created by the inherent V shape of the C domain, floor loop and latch. The donor and acceptor sites were identified at the tunnel entrances using biochemical studies and apo structures of the C domain (Keating et al., 2002, Bloudoff et al., 2013). Several structures of C domains have now been determined and all exhibit varying degrees of 'openness,' where the angle between the N- and C-

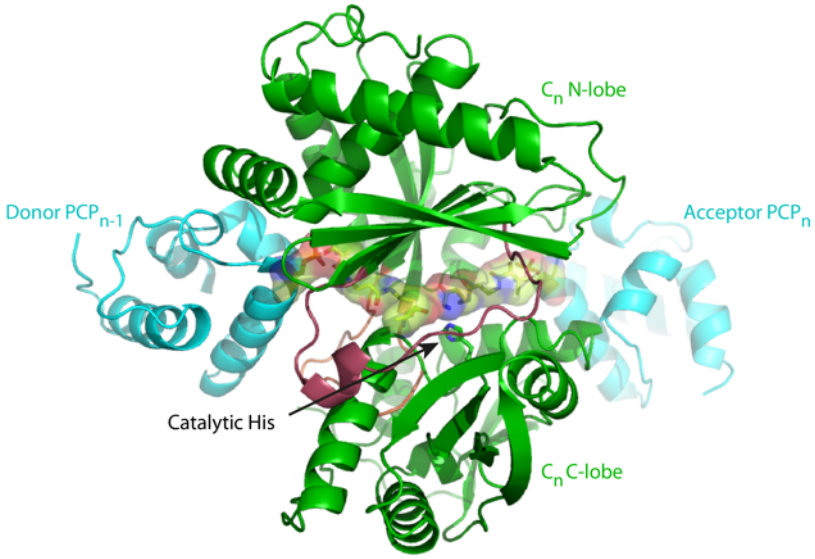


Figure 1.4 | PCP binding to the C domain.
Structure of the linear gramicidin LgrA C_2 domain with donor and acceptor PCP domains (Reimer JM *et al*, *manuscript in preparation*). Amino-acyl PPE moieties have been modelled in with surfaces shown to illustrate the substrate binding tunnel threading through the C domain. The catalytic histidine is shown in sticks in the middle of the tunnel. The latch is coloured raspberry and the floor loop orange.

lobes differs. It was proposed that C domains could transition between 'open' and 'closed' states to facilitate domain-domain communication throughout catalysis (Bloudoff et al., 2013), but further experimentation is needed to validate this theory.

The C domain was discovered over 3 decades ago (De Crecy-Lagard et al., 1995), but despite notable effort by multiple labs, the catalytic mechanism of the C domain has not been conclusively determined. The C domain contains a conserved catalytic motif, HHxxxDG, and it was originally thought that the second histidine, thus termed the catalytic histidine, of the motif would act like a general base, deprotonating the α-amino group of the acceptor aminoacyl-PCP to promote nucleophilic attack on the carbonyl carbon of the donor peptidyl-PCP. Several extensive mutagenesis experiments on the C domain were performed to probe the role of motif residues and other highly conserved residues surrounding the active site (Bergendahl et al., 2002, Roche and Walsh, 2003, Keating et al., 2002, Vater et al., 1997). The results of these studies reiterated the importance of the catalytic histidine and motif aspartic acid for maintaining condensation function, and identified other residues important for maintaining structural integrity. The catalytic histidine's role as a general base was called into question when it was found that some C domains

were still catalytically competent when the histidine was mutated with only a moderate decrease in activity (Keating et al., 2002, Marshall et al., 2002). Following, Samel et al. computed the pKa of the catalytic histidine in the structure of the TycC-C_6 domain and found it to be 11.8 (Samel et al., 2007); this indicated the histidine would be protonated under physiological conditions and unable to perform as a general base. Instead, they proposed that the histidine plays a positioning role by stabilizing the reaction intermediate in conjugation with the dipole moment of a proximal α-helix.

A breakthrough in unraveling the C domain catalytic mechanism came with a creative chemical biology approach undertaken by Bloudoff et al. Previous attempts by multiple groups, including our own, to solve a C domain structure with substrates, products and/or analogues were unsuccessful. A few structures captured the PPE arm in the active site tunnel (Drake et al., 2016, Chen et al., 2016), but substrates were never observed at the end of the PPE arm. Bloudoff et al. designed bromo-alkyl-aminoacyl chemical probes resembling the native amino-acyl acceptor substrate (Bloudoff et al., 2016). The chemical probes alkylated a cysteine engineered into the acceptor side of the CDA-C_1 domain active site tunnel, tethering and localizing the acceptor substrate mimic to the active site of the C domain. The structure of the alkylated CDA-C_1 was determined to high resolution, providing the first real glimpse into a C domain preparing for condensation. The sidechain of His157 is thought to be protonated and hydrogen bonds with the α-amino group of the amino-acyl acceptor substrate. This led to the conclusion that the role of the second histidine is to position the acceptor substrate α-amino group properly so that it can undergo nucleophilic attack on the donor thioester with the reaction cycling through a zitterionic transition state leading to the amide product (Yang and Drueckhammer, 2000). Similarly, the more famous peptide maker, the ribosome, was once thought to proceed through a general acid/base catalytic mechanism, but has now been shown to use substrate positioning of acyl-tRNA and solvent-mediated proton extraction to mediate peptide bond formation (Bieling et al., 2006, Kuhlenkoetter et al., 2011, Wallin and Aqvist, 2010).

1.2.4 The thioesterase domain

Located at the C-terminus of the NRPS, the Te domain is a ~275 amino acid α/β hydrolase domain with an active site topped by a variable 'lid' region. Te domains are related to serine

hydrolases and are homologous to their counterparts found in fatty acid biosynthesis and PKSs. Te domains use a Ser-His-Asp catalytic triad and proceed through a two-step mechanism. The PCP binds the edge of a crevice formed by the main core of the Te domain and the mobile lid segment (Frueh et al., 2008, Liu et al., 2011), and the PPE extends into the catalytic center. The PCP-Te interface buries 745 $Å^2$ of surface area, which is more than PCP binding to other domains. The configuration allows the Te domain active site serine to attack and accept the nascent peptide in the thioesterase first half-reaction, forming peptidyl-acyl-*O*-Te. The second half-reaction can be oligomerization, for which a presumably similar PCP-Te interaction occurs to provide additional copies of the peptide. The α-amino group from the newly synthesized peptidyl-PCP undergoes nucleophilic attack on the activated peptidyl-ester, resulting in oligomerization of the nascent peptide (as observed in gramicidin S and cereulide synthetases (Hoyer et al., 2007, Alonzo et al., 2015)). Alternatively, the second half-reaction can be hydrolysis or cyclization, for which the PCP domain presumably departs. The former release mechanism uses a water molecule to release linear peptide, while the latter uses an intra-peptidyl nucleophile leading to either head-to-tail (N-terminal nucleophile (Bruner et al., 2002, Kohli et al., 2002)) or branched (sidechain nucleophile (Konz et al., 1997)) cyclization. Furthermore, Te domains are also capable of regio- and stereo-specific cyclization, as exemplified in tyrocidine A and enterbactin synthetases (Keating et al., 2001). Te domains are typically specific for one form of peptide release, but this cannot be predicted as the discriminating factors between a hydrolyzing, cyclizing or oligomerizing Te are currently unknown.

1.3 Tailoring domains

It is exceedingly common for NRPSs to have optional tailoring domains, including oxidase, reductase, epimerization, ketoreductase, aminotransferase and methyltransferase domains, and the action of these domains must be incorporated into the catalytic cycle of the module where they act (Sundaram and Hertweck, 2016). NRPS tailoring domains were originally acquired by fusing genes for enzymes belonging to unrelated cellular processes to those which NRPSs perform (Lawrence and Roth, 1996), and are selected for incorporation based on their ability to perform unique chemical reactions outside of the NRPS synthetic scope. An NRPS will often employ several tailoring domains within its synthetic cycle, and the biological function of many nonribosomal

peptide products often relies on the successful modification of the peptide by a tailoring domain. There are three ways a tailoring domain can act on a nonribosomal peptide: *in cis* where the tailoring domain is incorporated into the NRPS architecture; *in trans* where the tailoring protein (or "accessory enzyme") is separate from the NRPS polypeptide but can act on intermediates covalently attached to the PCP domain; and post synthetically where the NRP is tailored following release from the NRPS (also called "maturation"). It is also worth noting that some NRPS systems include enzymes dedicated to introducing unusual chemistries into amino acids that are then selected by the NRPS as substrates. This increases the chemical diversity available to the nonribosomal peptide, but is formally a substrate generation process and not tailoring, so will not be discussed in detail here.

1.3.1 Cyclization domains

Some NRPs have unexpected thiazoline, oxazoline and methyloxazoline heterocyclic rings that are products of heterocyclization (Cy) domains. The C condensation domain can be replaced by a structurally-related Cy domain, which can perform both condensation and cyclization functions in modules that activate serine, threonine or cysteine residues (Figure 1.5). Following thiolation, the Cy domain will first mimic a C domain by condensing the upstream aminoacyl/peptidyl-PCP_{n-1} with serinyl/threoniyl/cysteinyl-PCP_n to generate an elongated peptide. The Cy domain then enables the hydroxyl or thio group of the Ser/Thr/Cys sidechain to undergo nucleophilic attack on the newly formed peptide carbonyl, followed by base catalysis to prompt final cyclodehydration (Duerfahrt et al., 2004). The oxidation state of the heterocycle can be furthered modified by additional tailoring oxidase (Du et al., 2000, Bloudoff et al., 2017) or reductase domains (Patel and Walsh, 2001, Reimmann et al., 2001) found *in cis* or *in trans*. The presence of the heterocycle is essential for the bioactivity of the NRP, as seen in bacitracin A (Konz et al., 1997), bleomycin (Shen et al., 2002) and bacillamide (Socha et al., 2007).

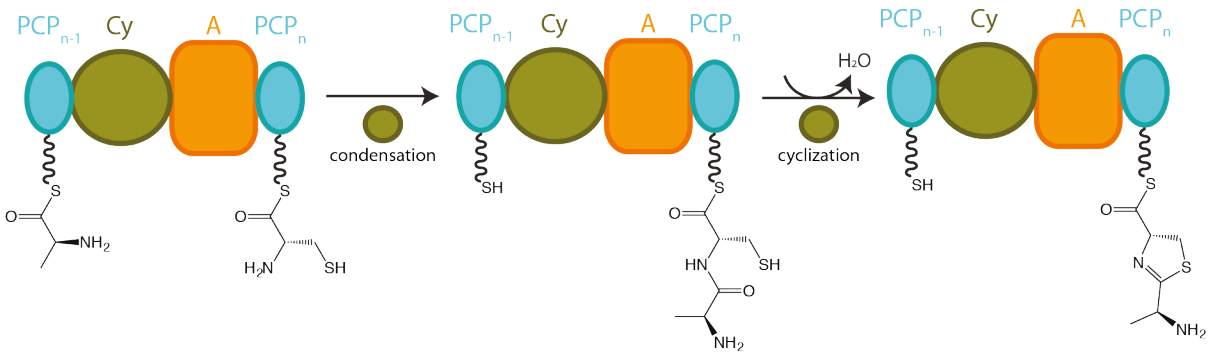


Figure 1.5 | Cyclization domains.
Condensation domains can be replaced with cyclization domains (Cy) that first condense substrates followed by cyclodehydration between cysteine, serine or threonine sidechains. Shown is a Cy-containing module from the bacillamide synthetase, adapted from Bloudoff K *et* al (Bloudoff et al., 2017).

Structures for the Cy domain were recently reported independently by two groups, the Cy domain from the epothilone synthetase protein, EpoB (Dowling et al., 2016), and the Cy domain found in BmdB of bacillamide synthetase (Bloudoff et al., 2017). The structures revealed the typical CAT fold characteristic of the C domain superfamily and provided foundational insights into the catalytic mechanism of Cy domains. Despite having condensation activity and structural homology to the C domain, Cy domains lack the C domain catalytic HHxxxD motif and instead, have their own aspartate-based DxxxxD motif (Konz et al., 1997). In the BmdB Cy structure, the two motif Asp sidechains occupy analogous positions to the first and last residues of the HHxxxD motif of the C domain. However, the sidechains are directed away from the active site and are used for structural integrity, not catalysis. Further, based on accompanying mutagenesis and bioinformatics studies, two new residues, D1226 and T1196, were identified to be integral for cyclodehydration. D1226 is poised to help orient substrates in the correct position and act as a general acid/base catalyst in cyclodehydration, while T1196 is within proximity to donate a proton to D1226. Additional structures with substrates or substrate analogues are needed to fully delineate the Cy domain catalytic mechanism.

1.3.2 Epimerization domains

Epimerization (E) domains are very common tailoring domains found in NRPS systems, and are embedded in the module following the PCP domain (C-A-PCP-E) (Figure 1.6). E domains are the most common route for integration of D-configured residues into NRPs, but three alternative routes exist. In one route, specific racemases enzymes associated with the NRPS clusters can provide D-amino acid which are recognized as cognate substrates by the A domain (Dittmann et al., 1994, Hoffmann et al., 1994, Cheng and Walton, 2000). The second route utilizes a tailoring dual functioning C/E domains which take the place of a C domain in a module and perform both condensation and epimerization reactions (Balibar et al., 2005). In a third, recently-reported route,

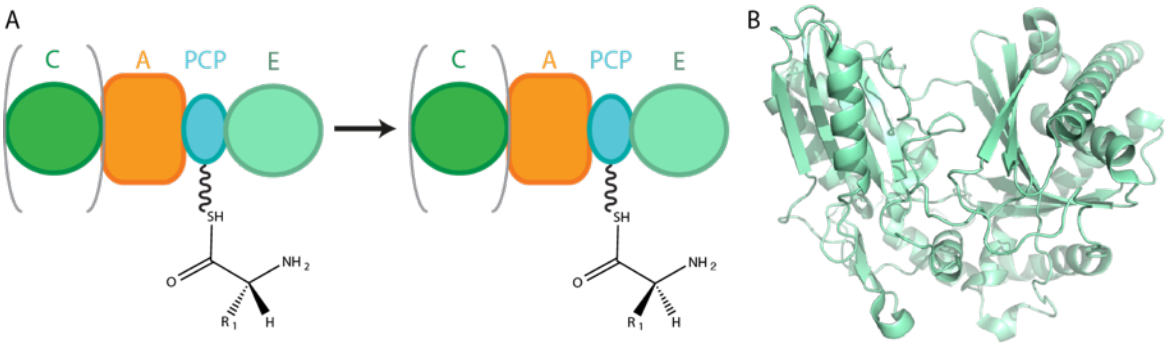


Figure 1.6 |Epimerization domains.
A, Epimerization (E) domains catalyze the epimerization of L-aminoacyl-PCP or peptidyl-PCP. **B**, The tyrocidine synthetase A (TycA) E domain, PDB 2XHG (Samel et al., 2014).

a unique Te domain epimerizes the peptide prior to hydrolysis from the NRPS scaffold (Gaudelli and Townsend, 2014). Nonproteinogenic D-amino acids confer advantages to NRPs by increasing conformational variability and providing resistance to the destructive action of cytosolic proteases (Radkov and Moe, 2014).

Surprisingly, the E domain was discovered prior to the C domain when Yamada & Kurahashi purified a "phenylalanine racemase" from *Bacillus brevis* that is now known to be part of gramicidin S synthetase 1 (Yamada and Kurahashi, 1968). Since then, extensive biochemical and structural studies have been performed to characterize the role of E domains in NRP synthesis. The E domain only epimerizes amino acids attached to the PCP arm (Luo et al., 2001), and has specificity for either L-aminoacyl-PCP or peptidyl-PCP, depending on the type of module which contains the E domain. For initiation modules (A-PCP-E), the E domain is capable of epimerizing L-

aminoacyl-PPE, while in elongation modules (C-A-PCP-E), the E domain displays preference for peptidyl-PCP (Linne and Marahiel, 2000; Stein, 2006). The order of catalytic events can be mapped based on the specificity for peptidyl-PCP: following substrate loading, condensation must occur within module$_n$ between aminoacyl/peptidyl-PCP$_{n-1}$ and aminoacyl-PCP$_n$ to produce a competent peptidyl-PCP$_n$ substrate for the E domain. Although the E domain can also catalyze the reverse reaction, the equilibrium between L- and D- substrate is kinetically driven towards the D configuration (Stachelhaus and Walsh, 2000). Additionally, the downstream C domain has been to shown to have chiral selectivity (Clugston et al., 2003), and thus acts as a gatekeeper to prevent L-amino acid from being incorporated into the final NRP and against misinitiation as peptidyl transfer cannot occur prior to epimerization (Linne and Marahiel, 2000, Stein, 2006 #1501).

Phylogenetic and sequence alignment analyses predicted the E domain to have a have a similar structure and catalytic mechanism to that of the C domain. Indeed, the first reported E domain structure was of tyrocidine synthetase A (TycA) E domain and revealed the expected CAT-like fold characteristic of the C domain superfamily. E domains house the same HHxxxDG catalytic motif, and structural analysis of TycA-E aided with the previous mutagenesis work on GrsA (A-PCP-E) (Stachelhaus and Walsh, 2000) led to a new proposal for the catalytic mechanism of the E domain: a conserved glutamic acid (not found in the C domain) located opposing the catalytic histidine was identified and suggested to take the role as a catalytic general acid-base (Samel et al., 2014). This proposal was later supported with the structure of the PCP-E didomain from GrsA where the holo-PCP domain is docked at the E domain's donor face with the thio-PPE extended into the active site (Chen et al., 2016). The sidechains of the catalytic histidine (His753) and conserved glutamic acid (Glu892) point towards the thiol group and are close to the theoretical donor and acceptor substrate positions. Through its dipole moment, helix α4 is suggested to help position and stabilize the reaction intermediate, which was first proposed in the study of CDA-C$_1$ modified with chemical probes (Bloudoff et al., 2016)). Unfortunately, despite attempts to solve the structure of GrsA aminoacyl-PCP-E, Chen *et al* were unable to observe substrate at the end of the PPE arm, which would have provided integral insights into the catalytic mechanism of both the E domain and the related C domain.

1.3.3 Methyltransferase domains

Nonribosomal peptides are often modified by methyltransferase (MT) domains to produce essential *N-*, *C-, S-* and *O-* methylated amino acids, such as in cyclosporin (Weber et al., 1994), yersiniabactin (Perry et al., 1999), thiocoraline (Al-Mestarihi et al., 2014) and saframycin Mx1 (Li et al., 2008) biosynthesis, with *N*-methylation being the most common. *N*-methyltransferases (N-MT) are ordinarily *in cis* and are embedded within the A_{sub} domain to act on aminoacyl-PCP prior to condensation (Shrestha and Garneau-Tsodikova, 2016; Ansari, 2008). N-MTs, as well as tailoring ketoreductase, oxidase and monooxygenase domains, usually interrupt the A_{sub} domain between the a8 and a9 motifs (Labby et al., 2015). *In trans* stand-alone N-MTs have been shown to methylate the backbone of peptidyl-PCP (Shi et al., 2009).

MTs are involved in many cellular processes and have been grouped into five distinct structural classes. The methyl donor, *S*-adenylmethinione (SAM), is used by all types of MTs, and the binding mode varies considerably between MTs even within the same class (Schubert et al., 2003). MTs display a high level of sequence divergence and were originally identified in NRPS clusters based on the presence of a methylated product (Weber et al., 1994). Following an in-depth *in silico* analysis of MTs from NRPSs and polyketide synthetases (PKSs), characteristic sequence motifs for *N-*, *C-,* and *O-* MTs were identified to allow prediction and classification of newly discovered NRPS MTs (Ansari et al., 2008).

Current structural information for MTs associated with NRPSs only exist for N-MTs. The *in trans* N-MT, MtfA, methylates the heptapeptide core prior to glycosylation in the synthesis of chloroeremomycin. The protein functions as a dimer and uses an unusually long 2-stranded β sheet to facilitate dimerization (Shi et al., 2009). However, the stand-alone MtfA has minimal sequence homology with embedded MTs and is not informative for the function of the more common integrated N-MT. Recently, the structure of an A domain from TioS with an *in cis* N-MT inserted into the A_{sub} domain was determined (A_{core}-A_{sub}-MT-A_{sub}:MLP; MLP, MbtH-like protein) (Figure 1.7B) (Mori et al., 2018). TioS is involved in the production of thiocoralines and the MT is suspected to methylate the peptide backbone. This is the first structure depicting an interrupted A_{sub} domain and shows how a tailoring domain can be integrated into the A_{sub} domain without

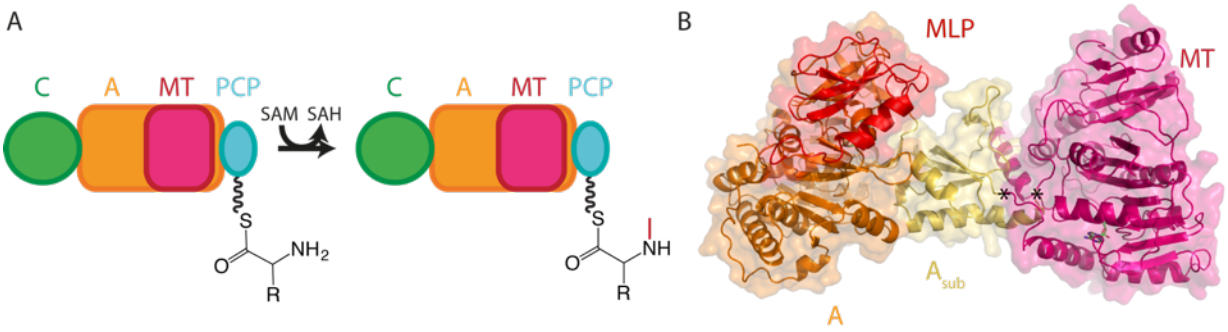


Figure 1.7 | N-methyltransferase domains.
A, N-MTs inserted into the Asub domain and N-methylate aminoacyl-PCP or peptidyl-PCP. **B**, The structure of an interrupted A domain from TioS with MLP bound (Mori et al., 2018). *demark MT insertion points. Colour code: A domain, orange; A_{sub} domain, yellow-orange; MT, hotpink; MLP, red.

disrupting A domain activity. In what is described as a 'dumbbell' structure, the N-MT caps the A_{sub} domain using extensive contacts and an elongated helix extending from the A_{sub} domain (α24, structural motif a8) into the MT anchors the two domains together. The A domain is in the thiolation state, and the PCP domain must bridge the ~ 60 Å distance between A and MT domain active sites for substrate to be methylated. The A_{core} domain also has a MLP bound to it. MLPs are necessary partners to some A domains, promoting activity in a manner that does not alter the average structure of the A domain (Miller et al., 2016). Interrupted A_{sub} domains are truly fascinating examples of NRPS architectural versatility, and it will be exciting to see future structures of embedded MTs or other tailoring domains in the context of an entire module to understand the conformational adaptations needed to proceed through the catalytic cycle.

1.3.4 Transglutaminase homologues

Andrimid is a potent inhibitor of the bacterial acetyl-CoA carboxylase and is produced by a particularly unique hybrid NRPS-PKS. This synthetase is extremely dissociated compared to normal NRPSs: 7 of the 12 proteins are stand-alone domains; the largest proteins, AdmO and AdmM, contain only 3 domains, none of which include all 3 canonical domains (C, A or PCP); and no more than one PCP domain is found per protein (Jin et al., 2006). The truly extraordinary feature of this synthetic system is that it does not include C domains or related domains to catalyze the formation of the first or second amide bonds. Instead, two free-standing transglutaminase homologues

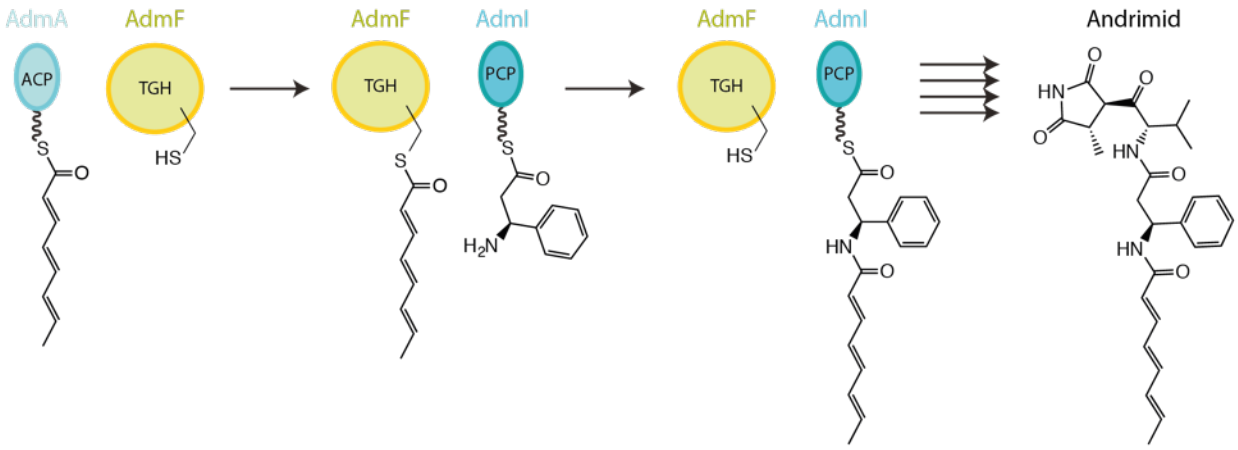


Figure 1.8 | Transglutaminase homologues.
AdmF is a transglutaminase homologue (TGH) found in the andrimid biosynthetic cluster, and catalyzes peptide bond formation between octatrienoyl-*S*-AdmA and (S)-β-Phe-*S*-AdmI.

(TGHs), AdmF and AdmS, are responsible for the first two amide bonds. Transglutaminases usually catalyze the formation of amide bonds between glutamine and lysine side chains for processes such as blood clotting (Cilia La Corte et al., 2011). That reaction is catalyzed by a catalytic triad, Cys-His-Asp, and proceeds through an acyl-enzyme intermediate. AdmF and AdmS contain the characteristic transglutaminase catalytic triad, but the rest of the protein does not show homology with any previously identified proteins. Other TGHs have been identified in orphan biosynthetic clusters, indicating the andrimid synthetase isn't the only NRPS to have co-opted this alternative peptide-making domain (Fortin et al., 2007).

AdmF, along with its carrier proteins, AdmA and AmdI, uses a donor fatty acid and an unusual β-amino acid as substrates to produce octatrienoyl-β-Phe-*S*-AdmI (Figure 1.8). (A dedicated aminomutase, AdmH, is present in the cluster to racemase L-Phe to β-Phe). Extensive biochemical characterization of AdmF determined that AdmF exhibits fairly promiscuous selectivity for donor and acceptor substrates with some preference for fatty acids over amino acids as the donor substrate (Magarvey et al., 2008, Fortin et al., 2007). AdmF also does not differentiate between the donor and acceptor PCP domains, AdmA and AdmI, *in vitro*, but does discriminate between its cognate carrier proteins over non-native carrier proteins.

AdmF/S represent a novel synthetic strategy for megaenzymes and hold promise for use in biocombinatorial experiments due to their somewhat relaxed substrate specificity. I, along with

several undergraduate students[2], made a significant effort to crystallize and solve the structure of AdmF and AdmS (and their homologues) with the overall goal of understanding how these unexpected amide bond-forming enzymes function. However, despite the dissociative nature of their biosynthetic cluster, AdmF and AdmS consistently exhibit a fervent disposition for aggregation and precipitate with abandon, thus rebuffing crystallization efforts and preventing further structural characterization.

1.3.5 Formylation domains

Although formylation is a key requirement in the synthesis of ribosomal peptides, formylation is a much less common modification in NRP synthesis. The first formylation (F) domain was identified in the anabaenopeptilide synthetase discovered in cyanobacterium *Anabaena* strain 90 by its sequence homology to methionyl-tRNA formyltransferase (Rouhiainen et al., 2000). The formylation gene was found in the initiation module of apdA, a dimodular protein subunit of the anabaenopeptilides 90A and 90B synthetases. Both anabaenopeptilides 90A and 90B (Figure 1.1) contain N-formylated glutamine, and using the A domain specificity code, the initiation module A domain was predicted to activate glutamine. Thus, it was proposed that the F domain formylates glutamine to provide the formyl modification observed in anabaenopeptilides.

Linear gramicidin synthetase is the second formylating NRPS to be identified and plays the star role in the following chapters. The antibiotic linear gramicidin was discovered in 1940 by soil microbiologist René Dubos, who subsequently recruited biochemist Rollin Hotchkiss to biochemically characterize the small molecule (Hotchkiss, 1940). Following linear gramicidin research focused on the bactericidal aspects of the molecule while research into its biosynthesis was largely ignored. It wasn't until over four decades later that Kessler *et al* located the linear gramicidin biosynthetic gene cluster in *Bacillus brevis* and identified the four protein subunits of linear gramicidin synthetase, LgrA-D (Kessler et al., 2004). A putative formylation domain was found in the initiation module of LgrA based on sequence alignments with methionine-tRNA formyltransferases and the presence of characteristic formyltransferase motifs, namely the catalytic triad and the SLLP formyltetrahydrofolate binding loop. Later, the F domain was

[2] Siraj Zahr, Laura Shen, Amanda Stanton, Ingrid Harb

biochemically characterized and was shown to use N^{10}-formyltetrahydrofolate (N^{10}-fTHF) (Figure 1.9) as a cofactor to formylate the first amino acid, valine, only when valine is presented to the F domain as valinyl-PPE-PCP. Further, formylation was required for synthesis of linear gramicidin to continue (Schoenafinger et al., 2006) and the bioactivity of the molecule (Wallace, 2000).

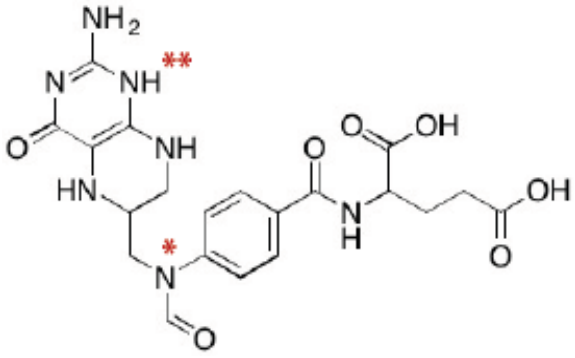


Figure 1.9 | F domain formyl donor.
N^{10}-formyltetrahydrofolate is used by F domains as a formyl group source in formylation reactions. The biologically relevant version is formylated on N10 (*). The commercially available analogue is formylated on N5 (**) and is often used in crystallographic studies to approximate the natural version.

1.3.6 Other tailoring domains

With the high occurrence of horizontal gene transfer in bacteria, the type of tailoring domains an NRPS can acquire is nearly endless. The synthetases for glycopeptide antibiotics (GPAs), including vancomycin and teicoplanin, contain an *in cis* X-domain that recruits four separate cytochrome P450 oxygenases to crosslink aromatic side chains to transform a linear heptapeptide into a complex aglycone structure (Haslinger et al., 2015). GPAs are further modified by dedicated glycosyltransferases to produce mature glycopeptide products (Losey et al., 2001). The cereulide and valinomyin synthetases have adopted a ketoreductase (KR) domain from their natural product producing cousin, PKSs, to produce notable depsipeptide products (Alonzo et al., 2015, Jaitzig et al., 2014). Like the MT domain, the KR domain is inserted into the A_{sub} domain and catalyzes stereospecific reduction of the α-keto monomer attached to the PPE arm (Magarvey et al., 2006). Halogenation of NRPs is executed by two types of halogenases: nonheme iron (II) halogenases that can activate aliphatic carbon centers to halogenate peptidyl-PPE using O_2, α-ketoglutarate and chloride, such as in barbamide (Galonic et al., 2006) and syringomycin E

(Vaillancourt et al., 2005) synthetases; and flavin-dependent halogenases that act on aromatic sidechains or heteroaromatic ring systems attached to the PPE arm in the presence of a dedicated flavin reductase, like in rebeccamyin (Yeh et al., 2006) and antibiotic C-1027 (Lin et al., 2007) synthetases. Many more tailoring domains exist, creating a vast pool of NRPs with diverse activities, and demonstrating the elegant ingenuity of the NRPS synthetic cycle.

1.4 Trapping NRPSs using chemical biology

The inherent flexibility of the A_{sub} and PCP domains has led to the development of several chemical biology tools for manipulating NRPSs to make them more amenable to crystallographic studies. Recombinant expression of PCP-containing NRPSs in *E. coli* will often lead to a heterogeneous sample of partially apo, holo and loaded PCP populations due to the relaxed specificity of EntD, the natural PPTase of *E. coli*. To obtain a homogenous sample suitable for both crystallographic and chemical biology experiments, two strains of *E. coli* have been engineered using homologous recombination: EntD was deleted from BL21 (DE3) cells to give BL21 (DE3) EntD- cells, yielding only apo PCP that can be modified by the promiscuous PPTase, Sfp, following purification (Chalut et al., 2006); and the *Bacillus subtilis sfp* gene was integrated into the *prp* operon of *E. coli* BL21 (DE3) cells under an IPTG-inducible T7 promotor giving rise to the *E. coli* BL21 (DE3) BAP1 (Pfeifer et al., 2001), producing holo PCP upon expression.

1.4.1 Phosphopantetheinyl analogues

The discovery of Sfp from *Bacillus subtilis* provided an innovative way to select specific catalytic states by targeting the PPE arm (Quadri et al., 1998). Sfp naturally uses CoA to post-

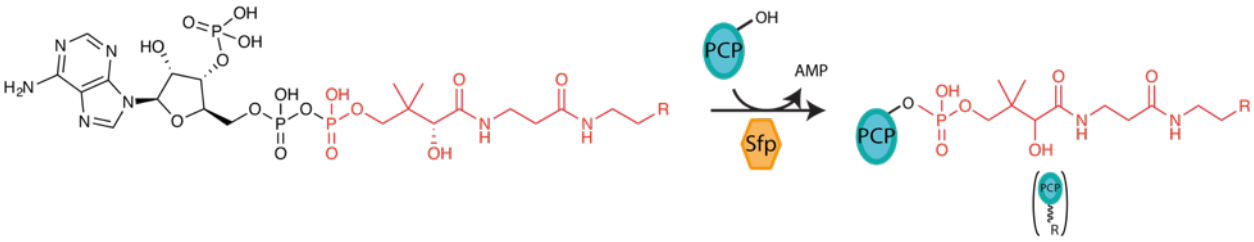


Figure 1.10 | Loading PCP domains with Sfp.
The phosphopantetheinyl transferase, Sfp, uses Coenzyme A and analogues to specifically load 4-phosphopantetheine onto a serine residue found in the PCP. The R group can be either (aminoacyl)-thio or (aminoacyl)-amino groups depending on the analogue.

translationally modify apo PCP to its holo form, but will also accept CoA analogues as viable substrates (Figure 1.10). As a result, a large arsenal of CoA analogues has been developed to probe the catalytic mechanism of different domains (Mishra and Drueckhammer, 2000). Many of the recent CoA analogue syntheses have used a chemo-enzymatic approach where they take advantage of the *E. coli* biosynthetic CoA enzymes (Worthington and Burkart, 2006, Nazi et al., 2004, Dai et al., 2001). In CoA biosynthesis, pantotheine is converted to CoA through the action of three enzymes, pantothenate kinase (PanK), phosphopantotheine adenyltransferase (PPAT) and dephosphocoenzyme A kinase (DPCK), in an ATP dependent reaction. Pantotheine can be chemically modified with a desired functional group and then converted to the CoA analogue in a one-pot reaction with purified PanK, PPAT and DPCK. Sfp can then use the functionalized CoA analogue to load the PCP domain for further studies. Alternatively, CoA itself can be functionalized with a myriad of different amino acids using PyBOP (Liu and Bruner, 2007) or N-hydroxysuccinimide (NHS) ester chemistry (Reimer et al., 2016a).

Substrates and products are linked to the PPE arm through a thioester bond, but the lability of the thioester bond can be problematic for biochemical and crystallographic experiments due to the high hydrolysis rate. To circumvent this, nonhydrolyzable analogues have been synthesized by functionalizing either amino-pantotheine (Liu and Bruner, 2007) or amino-CoA (Reimer et al., 2016a). This replaces the thioester bond with a more stable amide bond, albeit with altered geometry. However, it has been shown that A domains can still use the amino arm as a nucleophile to load substrate through the canonical adenylation reaction, generating aminoacyl-amide-PCP (Liu and Bruner, 2007). The use of amino-CoA analogues was essential for obtaining many of the crystal structures that are discussed in the following chapters.

1.4.2 Adenosine vinylsulfonamide inhibitors

Mechanism-based adenosine vinylsulfonamide inhibitors have become an important tool for studying A domains, both biochemically (Tarry et al., 2017, Tarry and Schmeing, 2015) and structurally (Tarry et al., 2017, Mitchell et al., 2012, Sundlov et al., 2012, Sundlov and Gulick, 2013, Drake et al., 2016, Miller et al., 2016). In the thiolation reaction, the A domain catalyzes nucleophilic attack of the PPE arm thiol on the aminoacyl adenylate. Adenosine vinylsulfonamide

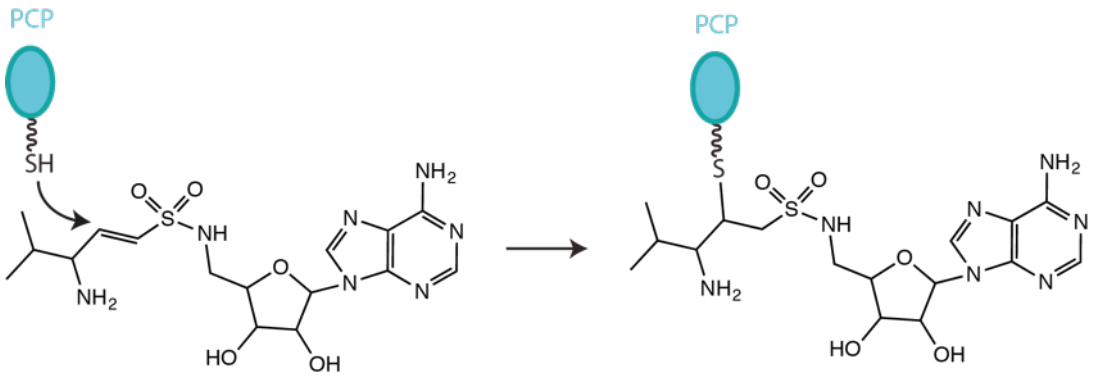


Figure 1.11 | Adenosine vinylsulfonamide inhibitor mechanism.

inhibitors contain a Michael acceptor and mimic the aminoacyl adenylate (Figure 1.11). Following binding of the inhibitor to the A domain active site, the A domain instigates the second half of the adenylation reaction, and the PPE arms becomes covalently bound to the inhibitor, forming a thioester intermediate mimic. Affinity for the adenosine vinylsulfonamide inhibitor is lower than that of the natural adenylate or other analogs of the thioesterification intermediate, but affinities are strong enough to markedly increase the interaction of the PCP domain and the A domain active site (Qiao et al., 2007b, Tarry et al., 2017)}.

PA1221 is a A-PCP didomain protein found in an NRPS orphan cluster and serves as an excellent example of the power of adenosine vinylsulfonamide inhibitors in structural studies of NRPSs. The structure of apo PA1221 was initially determined with the A_{sub} in the thiolation state, however, no electron density for the PCP was observed despite the A domain being in a competent conformation for PCP binding. Holo PA1221 was then inhibited using a valinyl-vinylsulfonamide inhibitor, and the subsequent structure revealed the PCP in the thiolation state with the PPE arm covalently linked to the vinylsulfonamide inhibitor in the A domain active site (Sundlov et al., 2012). Not only do mechanism-based inhibitors aid in visualizing flexible domains, they also provide mechanistic insights into reaction intermediates of the synthetic cycle.

With recent developments in the electron microscopy (EM) field, models of multi-modular NRPSs, an ambitious and lofty target (dream) for crystallographers, are starting to emerge using the same chemical biology tools developed for X-ray crystallography. Dimodular DhbF was recently examined using negative stain EM. To limit conformational heterogeneity in the sample, both A domains were stalled with vinylsulfonamide inhibitors. However, despite both modules being

locked into the thiolation state, DhbF was observed in multiple conformations (Further discussed in Section 1.5.4) (Tarry et al., 2017).

1.4.3 Electrophilic donor analogues

Bacteria contain three thio-templated systems, the fatty acid synthase (FAS), PKSs and NRPSs, which all use carrier proteins (CPs) to deliver substrates to various catalytic domains. Dead-end inhibitors have been developed to crosslink CPs to their binding partners to determine how they can specifically deliver substrates to each active site. This strategy was first employed using the ketosynthase (KS) domain found in the FAS of *E. coli*. The KS domain is analogous to the C domain in NRPSs as it condenses acyl substrates to extend the growing product. The active site contains a nucleophilic cysteine, where accepts the nascent product attached to the donor CP to form an acyl-*O*-KS intermediate. It was rationalized that if a CP was modified with an electrophilic β-chloroacryl moiety, upon reaction with the nucleophilic cysteine, β-chloro-elimination would occur, resulting in a permanent crosslink between the two domains (Figure 1.12A). β-

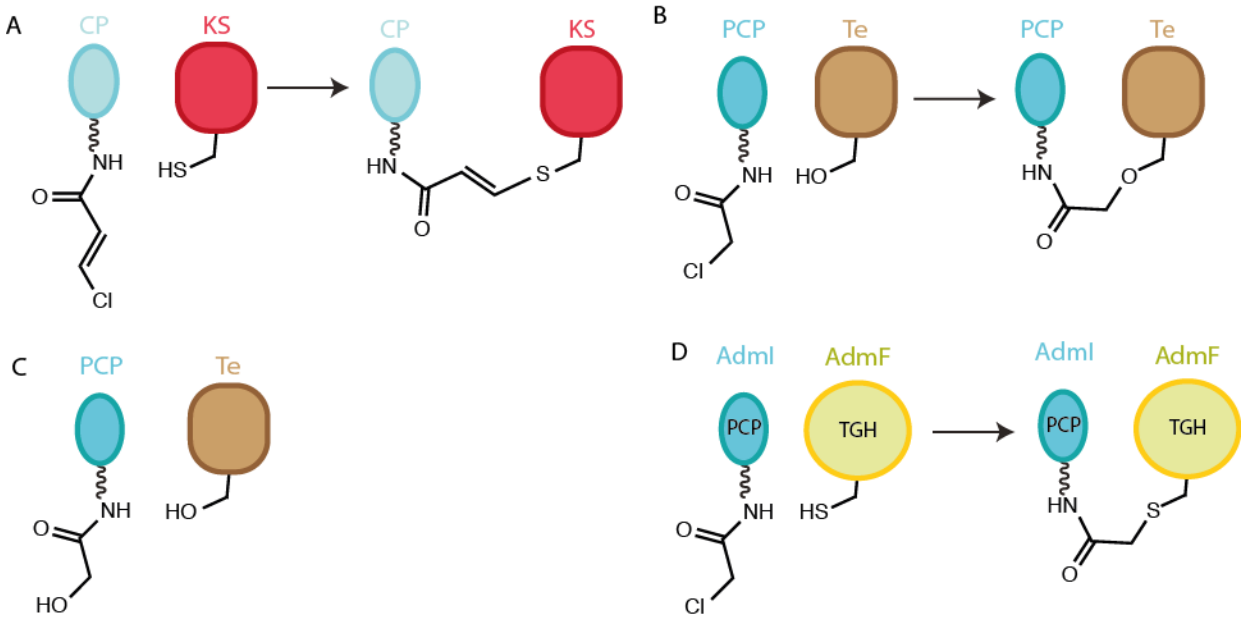


Figure 1.12 | Electrophilic donor analogues.
A, Strategy for crosslinking a CP to a KS domain using β-chloroacrylate-PPE (Worthington et al., 2006). **B**, Predicted reaction between α-chloro-acetyl-PPE-PCP and a Te domain. **C**, Observed reaction in the crystal structure of PCP-Te (Liu et al., 2011). **D**, Stand-alone domains can be crosslinked together using α-chloro-acetyl-CoA.

chloroacrylate-amide-CoA was synthesized, and after modification using Sfp, the CP domain became successfully irreversibly tethered to the KS domain of the *E. coli* fatty acid synthase (Worthington et al., 2006).

Similarly, the Te domain active site contains a nucleophilic serine that forms a peptidyl-acyl-*O*-Te intermediate prior to peptide release (Kohli and Walsh, 2003). Adopting an analogous strategy to β-chloroacrylate-amide-CoA inhibitor, Lui and Bruner synthesized the electrophilic donor analogue, α-chloro-acetyl-CoA, to tether the PCP domain to the Te domain of EntF (C-A-PCP-Te) (Liu and Bruner, 2007). The advantage of α-chloro-acetyl-CoA over β-chloroacrylate-amide-CoA is that the final crosslink only introduces a single extra carbon, and is thus a closer mimic to the natural intermediate state (Figure 1.11B). Following PCP domain modification with the inhibitor, Te domain activity assays showed significantly decreased activity, indicating the PCP domain had become successfully tethered to the Te domain. This tactic was later used to aid in the crystallization of the PCP-Te didomain construct of EntF (Liu et al., 2011). The structure revealed the PCP domain docked at the Te domain with the PPE arm extending into the active site. However, instead of observing the expected crosslink between the PPE arm and the active site serine, an unexpected reaction occurred resulting in the formation of α-hydroxy-acetyl-amide at the end of the PPE arm (Figure 1.12C). Despite this, the use of α-chloro-acetyl-CoA was successful in restricting the movement between the two domains, allowing structure determination.

The use of electrophilic donor analogues is especially useful for tethering stand-alone domains. Embedded PCP domains have the advantage that inter-domain linkers naturally keep the PCP domain in close proximity with their binding partners. In andrimid biosynthesis, the stand-alone PCP domains, AdmA and AdmI, must interact with AdmF to deliver the donor and acceptor substrates. To study these transient PCP-TGH interactions, we have shown that α-chloro-acetyl-CoA can be used to crosslink AdmA or AdmI to AdmF using the catalytic cysteine found in AdmF (unpublished) (Figure 1.12D). Electrophilic donor analogues have proven a power tool in studying these important PCP domain interactions.

1.4.4 Small molecules

The use of non-covalent small molecules can also promote a desired conformational state. It was shown that adenylating enzymes have higher affinity for their aminoacyl-adenylate intermediates than their starting substrates and can be used to partially inhibit A domain activity (Forrest et al., 2000). The inspiration for the above vinylsulfonamide inhibitors came from adenylate analogues, such as 5-*O*-*N*-(aminoacyl)sulfamoyl-adenosine, which were designed to trap the A domain in the closed conformation (Finking et al., 2003, Ferreras et al., 2005, Qiao et al., 2007a). Indeed, the crystallization of the LgrA construct, F-A-PCP-C (see Chapter 4) was greatly enhanced by co-crystallizing with the valyl-adenylate analogue, 5-*O*-*N*-valylsulfamoyl-adenosine. Similarly, the nonhydrolyzable ATP analogue, alpha, beta-methyleneadenosine 5'triposphate (AMPcPP), has been used in crystals structures of A domains to visualize the pre-adenylation state (Herbst et al., 2013, Chapter 4).

1.5 Structural characterization of multi-domain constructs

1.5.1 Towards understanding NRPS architecture

The first view of a complete module was that of the termination module of surfactin synthetase, SrfA-C, solved in the peptide-accepting state a decade ago (Figure 1.13) (Tanovic et al., 2008). With a domain architecture of C–A–PCP–Te, it represents both a minimal C–A–PCP elongation module and the most common type of bacterial termination module. The SrfA-C structure showed large distances between active sites in NRPSs indicating that substantial conformations changes would occur in the catalytic cycle. Another key finding was that the C-terminal lobe of the C domain and the A_{core} form a 'catalytic platform', burying a sizable (765 $Å^2$) area of surface. This interaction defines the overall rectangular shape of an elongation module, was by far the most extensive interdomain contact seen in the module, and was proposed to be a fixed interface.

Our knowledge of structures of elongation/termination modules was greatly enhanced with the recent determination of two more C–A–PCP–Te termination modules, of EntF (involved in enterobactin production) and AB3403 (from an uncharacterized pathway) (Figure 1.13) (Drake et al., 2016, Miller et al., 2016). Firstly, there was important insight regarding the predicted rigid

C:A interface. Despite similar interaction surfaces donated by the C domain and the A_{core} of SrfA-C, EntF and AB3403, a difference in relative angle of ~20° is propagated over the C domain, and combined with differences in conformation of the C domain lobes, analogous atoms in the far side of the C domains assume positions > 30 Å apart. This is not solely due to differences between synthetases, as two separate structures of the EntF module show somewhat shifted interfaces that cause a ~15 Å difference in relative orientation of the N terminus of the C domain. Thus, the C:A catalytic platform is more plastic than first thought. However, even with these movements, the C:A interface is by far the most constant of any between two canonical domains, and defines an NRPS module structurally.

A model of an elongation module cycle can be constructed using the catalytically-relevant states observed in the SrfA-C, EntF and AB3403 structures (Tanovic et al., 2008, Drake et al., 2016, Miller et al., 2016). As is the case in initiation modules, the elongation module starts its cycle with the A domain binding amino acid and ATP, and then adenylation. AB3403 shows the adenylation state of this module, with substrate analogues bound and the A_{sub} domain in the closed conformation (Figure 1.13, state i,ii). The next stage, thiolation, is captured in the structure of EntF stalled by a mechanism-based aminoacyl-adenosine-vinylsufonamide inhibitor (Figure 1.13, state iii). The aminoacyl-PCP then travels ~45 Å, rotating ~ 75°, to bind the acceptor site of the C domain (Figure 1.13, state iv) and elongate the nascent peptide in the condensation reaction. Both SrfA-C and AB3403 were visualized with PCP domain bound at the acceptor site of the C domain, but key differences do exist, highlighting the advantages of obtaining multiple similar structures (compare Figure 1.13, state i and ii). Firstly, the PCP domain of SrfA-C is unmodified, but in AB3403, the PPE arm was seen making specific hydrogen bonds with the side of the C domain tunnel. Furthermore, the PCP domain of AB3403 is rotated 30° relative to that of SrfA-C. The PCP domain of SrfA-C is unable to take the position seen in AB3403 because it would overlap with the A_{sub} domain and a loop of the C domain. One or the other position could be influenced by crystal packing, but the acceptor site of the C domain is even more shallow than the donor site, and it is likely that particular PCP-C domain pairs have their own preferred binding orientations, and that even within a particular PCP-C pair, multiple orientations should allow productive substrate delivery to the

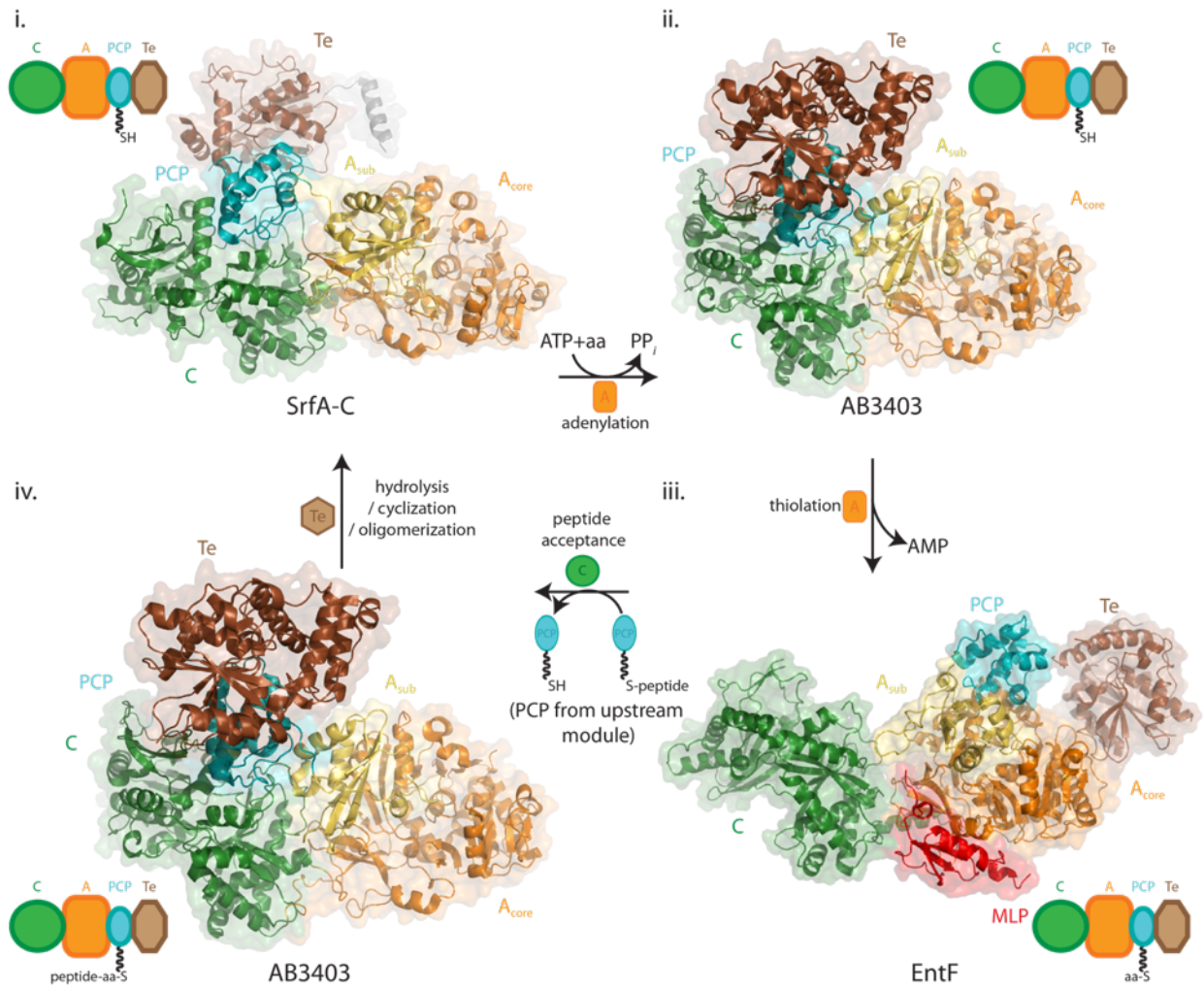


Figure 1.13 | Elongation and termination of nonribosomal peptide synthesis.
The catalytic cycle of the elongation and termination modules as illustrated by crystal structures of SrfA-C (PDB 2VSQ (Tanovic et al., 2008)), EntF (PDB 5JA2 (Miller et al., 2016)) and AB3404 (PDB 4ZXH (Drake et al., 2016)). The module selects and adenylates an amino acid (**i, ii**). (The SrfA-C A domain (**i**) is in a pseudo-open state and the AB3403 is in a canonical closed state). The A domain covalently tethers the amino acid to the PCP domain via thioester (**iii**). This aminoacyl-PCP accepts (**iv**) the peptidyl group from the upstream PCP, elongating the peptide. The PCP domains transports the peptide to the termination domain to release the peptide. Colour code: C domain, green; A_{core} domain, orange; A_{sub} domain, yelloworange; PCP domain, teal; Te domain, brown; MbtH-like protein (MLP), red; C-terminal affinity tag, grey.

catalytic center of the C domain. This delivery permits the aminoacyl-PCP to accept the peptidyl group from the donor, transforming it to a new, elongated peptidyl-PCP. Interestingly, the AB3404 structure shows that the first stages of the next cycle need not wait for condensation, as the

adenylation state is observed simultaneously with the peptide acceptance state (Drake et al., 2016). Accordingly, Figure 1.13, state ii and iv are the same structure.

1.5.2 Te domain flexibility

After the last elongation cycle, the PCP domain delivers the elongated peptide to the chain-terminating domain. In fungal termination modules, that is often a C_T domain, but for bacterial NRPSs, it is most commonly a Te domain and second most commonly a terminal reductase (R) domain. SrfA-C, EntF and AB3403 all have domains C–A–PCP–Te, yet their structures show markedly different Te domain positions (Figure 1.13). In SrfA-C, the Te domain floats near the acceptor side of the C domain and makes almost no contact with the other domains. The Te active site is facing away from the rest of the module and is partially blocked by the C domain, so both the Te domain and the PCP domain must move for productive binding. Te domain movement seems simplistic however, as in AB3403, the Te domain is displaced by >50 Å and only contacts the 'back face' of the PCP domain (another contact that would need to break for termination). Furthermore, in EntF the Te domain is >80 Å away from either of these positions, making modest contacts with the A domain. Electron microscopy also shows this domain to be very mobile domain (Drake et al., 2016) (Tarry et al., 2017). Thus, the Te domain is loosely tethered to the termination module via the mobile PCP.

1.5.3 PCP-C interaction

Interestingly, the first glimpses into the important PCP-C domain substrate donation conformation came from C domain homologues, an E domain (Chen et al., 2016) and a terminal cyclizing C (C_T) (Zhang et al., 2016). The donor PCP binding site was established by early apo structures of the C domain and with biochemistry (Keating et al., 2002, Bloudoff et al., 2013). The structure of an excised PCP-C didomain from TycC visualized the two domains require for donation, but was crystallized in a non-productive conformation as the PPE attachment site was oriented away from the C domain (Figure 1.14A) (Samel et al., 2007). The E domain catalyzes epimerization of the first amino acid residue of a peptiyl-PCP, and the C_T domain catalyzes internal cyclization and release of the peptide from peptidyl-PCP. Both the PCP-E structure form

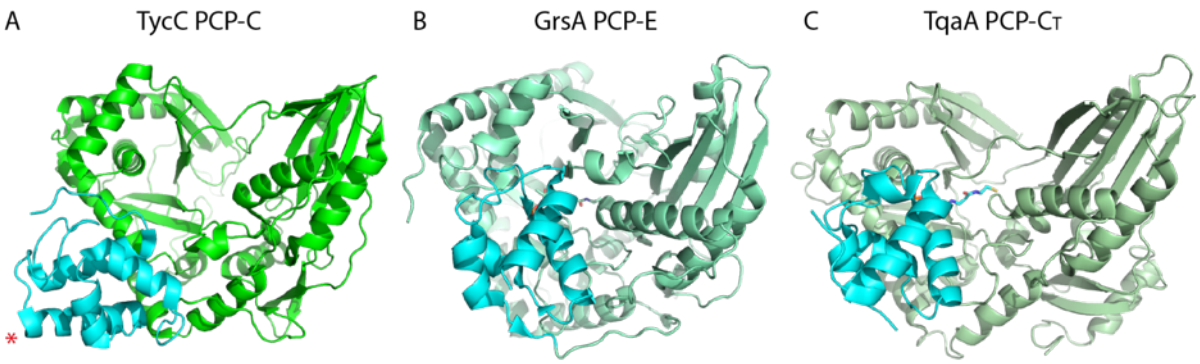


Figure 1.14 | Substrate donation to C domain and C domain homologues.
A, Structure of the TycC PCP-C didomain in an unproductive substrate donation state. * demarks modified serine. Substrate delivery is represented by didomain structures of PCP domain and C domain homologues, (**B**) E domain (Chen et al., 2016) and (**C**) C_T domain (Zhang et al., 2016). Colour code: PCP domain, cyan; C domain, green; E domain, teal; C_T domain, pale green.

GrsA (Figure 1.14B) and the PCP-C_T structure from TqaA (Figure 1.14C) contain PCP domains in the broad and somewhat shallow donor site, and visualize PPE arms lining the C domain tunnel toward the active sites. However, the two conformations of the PCP domains are rotated ~30° to one another and bind either the N-lobe or C-lobe of the broad donor site depression. The differences between the E and C_T domains and bona fide C domains are most pronounced in the area of the acceptor PCP binding site, as this position is blocked by domain-specific sequences in E and C_T domains. The donor face is not conserved and it is unclear whether the PCP bound to a canonical C domain would sample both the observed positions or assume a new orientation, or whether the presence of the rest of the upstream and downstream modules would influence the PCP:C interaction. These structures do certainly represent the broad strokes of what is happening when an upstream PCP delivers a donor substrate to be transferred in the peptidyl transferase reaction to the aminoacyl-PCP generated by an elongation module.

1.5.4 Bridging modules

The above-described structures, in context of excellent existing functional studies, provide a wealth of insight into each step of the NRPS cycle. They are less informative about how modules form intact NRPS megaenzymes. High resolution structures of multi-modular NRPSs are an outstanding goal in the field, and multiple such structures are required to answer questions of NRPS architecture, but a view of higher order structure is beginning to form.

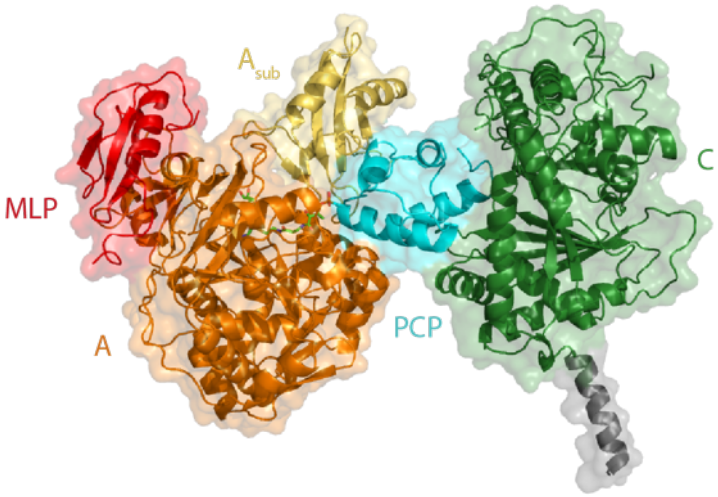


Figure 1.15 | Cross-module structure of DhbF.
The cross-module structure from the DhbF synthetase showing A_1-PCP_1-C_2 with bound MLP (Tarry et al., 2017). Colour code: C domain, green; A_{core} domain, orange; A_{sub} domain, yelloworange; PCP domain, teal; MbtH-like protein (MLP), red; C-terminal affinity tag, grey.

The structure of the cross-domain construct of bacillibactin synthetase protein DhbF was recently solved by the Schmeing group (Figure 1.15) (Tarry et al., 2017). The A_1-PCP_1-C_2 construct contained the A and PCP domains from the first module and the C domain from the second module. In this structure, the vinylsulfonamide inhibitor is present in the A domain active site and attached to the PCP domain, but the A_{sub} and PCP domain are somewhat shifted from a true thiolation conformation, likely due to crystal contacts. This structure (and the second structure of EntF shown above) both contain MLPs complexed to the A_{core} in the same position as in a fused A-MLP protein (Herbst et al., 2013).

A structure with the last large domain from module 1 and the first large domain from module 2 should allow visualization of intermodule interactions not formed by the PCP domain (which is unlikely to contribute to consistent higher-order structure because of its mobility). However, the structure showed no contact between the A_1 and C_2 domains. Instead, the only non-covalent interaction between module 1 and 2 was the back face of the PCP domain contacting the donor site of the C domain. A similar interaction was observed in the TycC PCP-C didomain structure (Figure 1.14A) and is echoed by the back face of the PCP domain interacting with the Te domain in AB3403. The lack of intermodular A:C contacts is in stark contrast to the extensive C:A contacts within a module, and implies that DhbF may not assume a single module-module

conformation. This was confirmed by negative stain electron microscopy (EM) of dimodular DhbF (C_1-A_1-PCP_1-C_2-A_2-PCP_2). Despite stalling with vinylsulfonamide inhibitors, DhbF was in multiple conformations. Five separate envelopes were reconstructed from the data and each envelope could be fitted with two models of C-A-PCP modules, as expected. However, there were large differences in the relative orientations of the two modules and no consistent module-module interface was observed, strongly suggesting that for DhbF at least, no regular, repeating supermodular architecture is present. This seems to also be the case for cyclosporin synthetase, a 12-module NRPS, micrographs of which show it to adopt either a 'ball of balls' or uneven 'balls on a string' morphology (Hoppert et al., 2001). It remains to be seen whether lack of higher-order architecture is a general theme for NRPS, or whether elegant and attractive regular architecture, which can be modelled based on single conformational states, can occur in some NRPSs (Marahiel, 2016).

1.6 Conformational dynamics

In addition to the fundamental structural work already done on NRPSs, NRPSs have been subjected to other biophysical techniques to delineate the order and timing of conformational changes that drive catalysis.

1.6.1 Adenylation reaction monitored by FRET

The conformational switch between the adenylation and thiolation steps in the adenylation reaction is referred to as the alternation mechanism. X-ray crystallography has definitively shown that these states exist within the A domain, but it is largely unknown how the transition between the A and PCP domain conformational states are coupled to the overall catalytic cycle. To answer this question, Alfermann *et al* used fluorescence resonance energy transfer (FRET) to monitor the adenylation reaction using GrsA (A-PCP, selective for Phe) as a model (Alfermann et al., 2017). They designed a FRET system where enhanced green fluorescent protein (EGFP) was attached C-terminally to the PCP domain and a reporting dye, AF546, was tethered to an engineered cysteine on the A_{core} domain near the active site. It was predicted based on the positions of the probes that a higher FRET signal would solely occur during the thiolation

state as the PCP only interacts with the A domain during this catalytic stage. However, after an exhaustive set of experiments using various substrates and ligands, they found that two states produced a high FRET signal: pre-transfer when the A domain has adopted the thiolation state but has not catalyzed thioesterification between holo PPE and Phe-adenylate; and post-transfer when the PPE arm has been loaded (Phe-PPE) but hasn't disengaged from the A domain active site. The latter is representative of product inhibition, and they propose that substrates for the next round of catalysis compete for the active site in a dynamic equilibrium state. Additionally, secluding Phe-PPE-PCP in the A domain active site could be used to protect the intermediate from premature hydrolysis if the C domain isn't ready to receive substrate. This contradicts an older alternative theory that the C domain binds acceptor aminoacyl-PCP substrate and waits until the donor substrate arrives before condensation can occur (Belshaw et al., 1999, Linne and Marahiel, 2000). It will be interesting to see the result of FRET experiments applied across entire module or multi-module systems to observe the effects of domain-domain interactions on catalysis.

1.6.2 Molecular dynamics

Molecular dynamics (MD) has become a powerful tool for investigating both the finer and grander movements required for catalysis using computational simulations. Despite their large size, NRPSs have been the subject of several MD simulations to investigate both intra-domain and inter-domain movements, with focus on condensing enzymes.

All C domain structures have been visualized in various degrees of 'openness' with up to ~12 Å between equivalent atoms at distal helices. Targeted MD simulations were performed on the CDA-C1 domain, the most closed C domain structure reported, and showed that CDA-C1 can plausibly transition to a more open state (as compared to the SrfA-C, TycC and VibH C domain structures) without any major clashes (Bloudoff et al., 2013). It is unknown if the C domain flexes between states as a communication mechanism with other domains or if the degree of openness is purely a variable structural feature of the C domain with no added mechanistic ramifications.

Docking domains, also known as communication (COM) domains, are very small domains that may be found in PKSs, NRPSs or hybrid PKS-NRPSs that are spread across multiple polypeptides. The natural product, epothilone, is produced by a hybrid PKS/NRPS in *Sorongium*

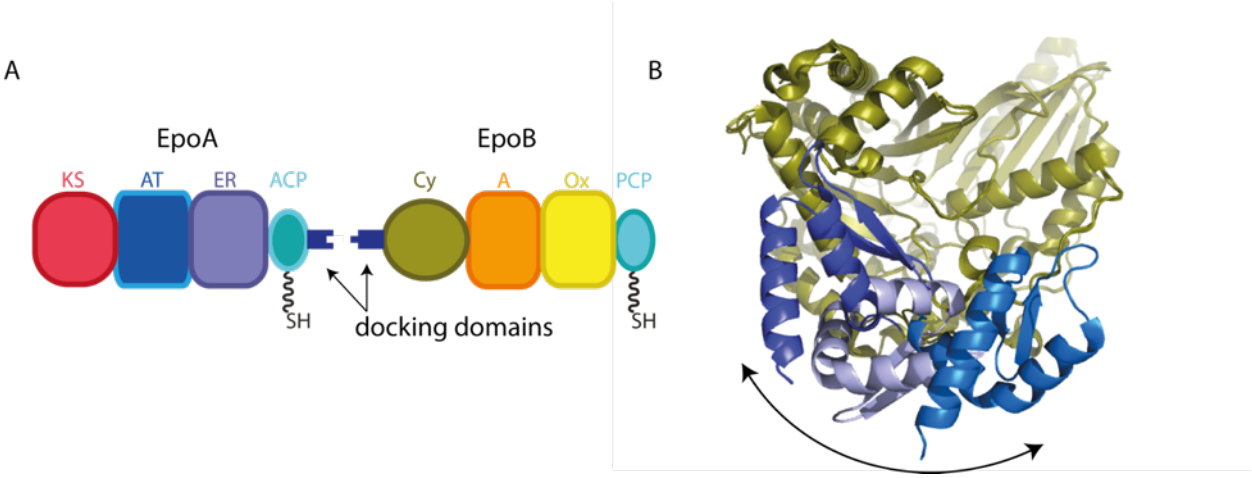


Figure 1.16 | Docking domains in epothilone biosynthesis.
A, Architecture of the first two proteins of the epothilone PKS/NRPS. Docking domains are used to promote interaction between EpoA and EpoB. KS, ketosynthase; AT, acetyltransferase; ER, -enoyl reductase; ACP, acyl carrier protein; Cy, cyclization; A, adenylation; Ox, oxidase; PCP, peptidyl carrier protein. **B**, Superimposed crystal structures of EpoB-Cy showing the three observed orientations of the docking domain. MD simulations recapitulated these movements (Dowling et al., 2016). PDB 5T81 and 5T7Z. Color code: Cy domain, olive; docking domain, various blues.

cellulosum and uses docking domains to facilitate substrate transfer from its PKS protein, EpoA, to its NRPS protein, EpoB (Figure 1.16A) (Julien et al., 2000): one docking domain is attached C-terminally to the PCP domain of EpoA and binds its complementary docking domain attached N-terminally to the starting cyclization domain of EpoB. Three structures of the Cy domain with its docking domain were determined with the docking domain observed in different orientations (Figure 1.16) (Dowling et al., 2016). MD simulations were carried out on the docking-Cy structures, and while the Cy domain remained relatively static, the docking domain showed a high degree of mobility. Furthermore, the MD simulations revealed that the active site tunnel, restricted in the crystal structure due to crystal packing, can dilate through minimal movements in sidechains to accommodate a fully loaded PPE arm.

1.6.3 Nuclear magnetic resonance of PCP domains.

The small size of the PCP domain has made it an excellent target for nuclear magnetic resonance (NMR) experiments. Early NMR studies focused on an excised PCP domain from the tyrocidine NRPS ($TycC_3$ PCP) and demonstrated that the PCP adopts three markedly different

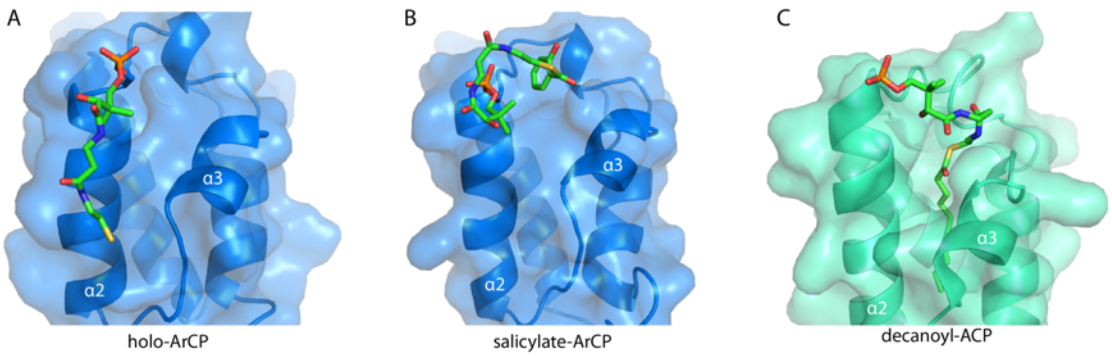


Figure 1.17 | Carrier protein dynamics.
Representative NMR states are shown for (A) holo-ArCP (PDB, 2N6Y (Goodrich et al., 2015)), (B) salicylate-ArCP (PDB, 2N6Z (Goodrich et al., 2015)) and (C) decanoyl-ACP (PDB, 2FAE (Roujeinikova et al., 2007)), demonstrating how ACPs bind the PPE arm depending on loaded state.

conformations in solution: an 'A state' found only in apo-PCP, an 'H state' adopted only by holo-PCP, and an 'A/H state' that exists for both apo- and holo-PCP (Koglin et al., 2006). Further NMR experiments with the phosphopantetheinyl transferase, Sfp, and a Te domain revealed that these two binding partners could discriminate between the A, H and A/H states to selectively bind a specific PCP conformation with the presence or absence of the PPE arm mediating the switch between states (Koglin et al., 2006, Koglin et al., 2008). The idea that the PCP can drive the direction of catalysis by switching its conformation to favour the next binding partner is an attractive theory. However, the theory has been challenged in recent years as the growing number of structures of PCP domains (Tufar et al., 2014, Samel et al., 2007, Liu et al., 2011, Mitchell et al., 2012, Sundlov et al., 2012, Tan et al., 2015, Reimer et al., 2016a), both in the context of a single domain or part of a larger construct, have all revealed the PCP in the A/H state, regardless PPE arm modification.

Carrier proteins in Type II (stand-alone domains) FAS and PKS systems have been shown by NMR and X-ray crystallography to protect their loaded substrates from cytosolic hydrolysis by sequestering them in a cleft between helices 2 and 3 (Roujeinikova et al., 2007, Ploskon et al., 2010, Crump et al., 1997) (Figure 1.17c). It remains unclear if NRPS PCPs use a similar strategy to protect substrate or modulate the catalytic cycle. PPE interactions with PtlL, a pyrrole type II PCP domain from the pyoluteorin synthetase, were investigated by NMR, and revealed that PtlL transiently binds either holo or loaded PPE in near identical modes with substrate or thio group residing between helices 2 and 3 (Jaremko et al., 2015). Subtle changes in the overall PtlL structure

were observed, but were insubstantial compared to the differences observed for the TycC-PCP domain (with A, H and A/H states). In a complimentary study, NMR was used to observe PPE arm interactions with the aryl carrier protein (ArCP) found in the yersiniabactin synthetase, and revealed the PPE arm transiently interacts with ArCP in two distinct conformations (Goodrich et al., 2015) (Figure 1.17A, B). In the holo form, the PPE arm was found parallel to helix 2 (Figure 1.17A) while in the loaded form, the PPE arm was curled backed onto itself and substrate rested between helices 2 and 3 (Figure 1.17B). Goodrich *et al* proposed that the different PPE arm conformations could help guide the PCP domain to its next binding partner as the PPE arm would hide or expose different PCP binding surfaces depending on its loaded state. However, ArCP has been excised from its native NRPS setting and it is unclear how being part of an NRPS module would affect the PPE conformations observed. An overlay between holo-ArCP with the structure of the LgrA initiation module in the thiolation conformation (Reimer et al., 2016a) shows that the PPE arm would clash with the A domain, indicating the non-covalent interactions between the PPE arm and PCP would have to break in the transition between adenylation and thiolation states of the A domain alternation mechanism. Indeed, the structures of multi-domain constructs that visualize the PPE arm in an active site show the PPE arm is extended and makes no contact with the PCP domain (Reimer et al., 2016a, Tarry et al., 2017, Drake et al., 2016, Miller et al., 2016). This is consistent with the 'swinging arm (Kittila et al., 2016, Felnagle et al., 2008)' model where the PPE arm does not actively interact with the PCP but instead, resembles a wrecking ball with substrate freely swinging around. Further, ArCP and PltL are not representative of canonical PCPs as they transport unusual cargos and have unique structural features. Additional studies of archetypal PCPs within their native NRPS setting are needed to better understand the role the PPE arm in guiding biosynthesis.

1.7 Bioengineering NRPSs

A key goal in the megaenzyme field is to harness synthetic systems to produce novel or improved natural chemicals through bioengineering endeavors. The modular nature of NRPSs have made them prime bioengineering targets as new products can theoretically be produced by

altering the specificity of a module or changing the NRPS architecture by domain or module swapping.

1.7.1 Precursor directed biosynthesis and mutasynthesis

Early bioengineering attempts focused on using either precursor directed biosynthesis or mutasynthesis to produce modified products (Figure 1.18A). In precursor directed biosynthesis, NRPS-producing cultures are supplied with alternative substrates that compete with the natural substrate for selection by an A domain for incorporation into the final product. This method relies on both the A and C domains having relaxed specificity to accommodate the unnatural substrate

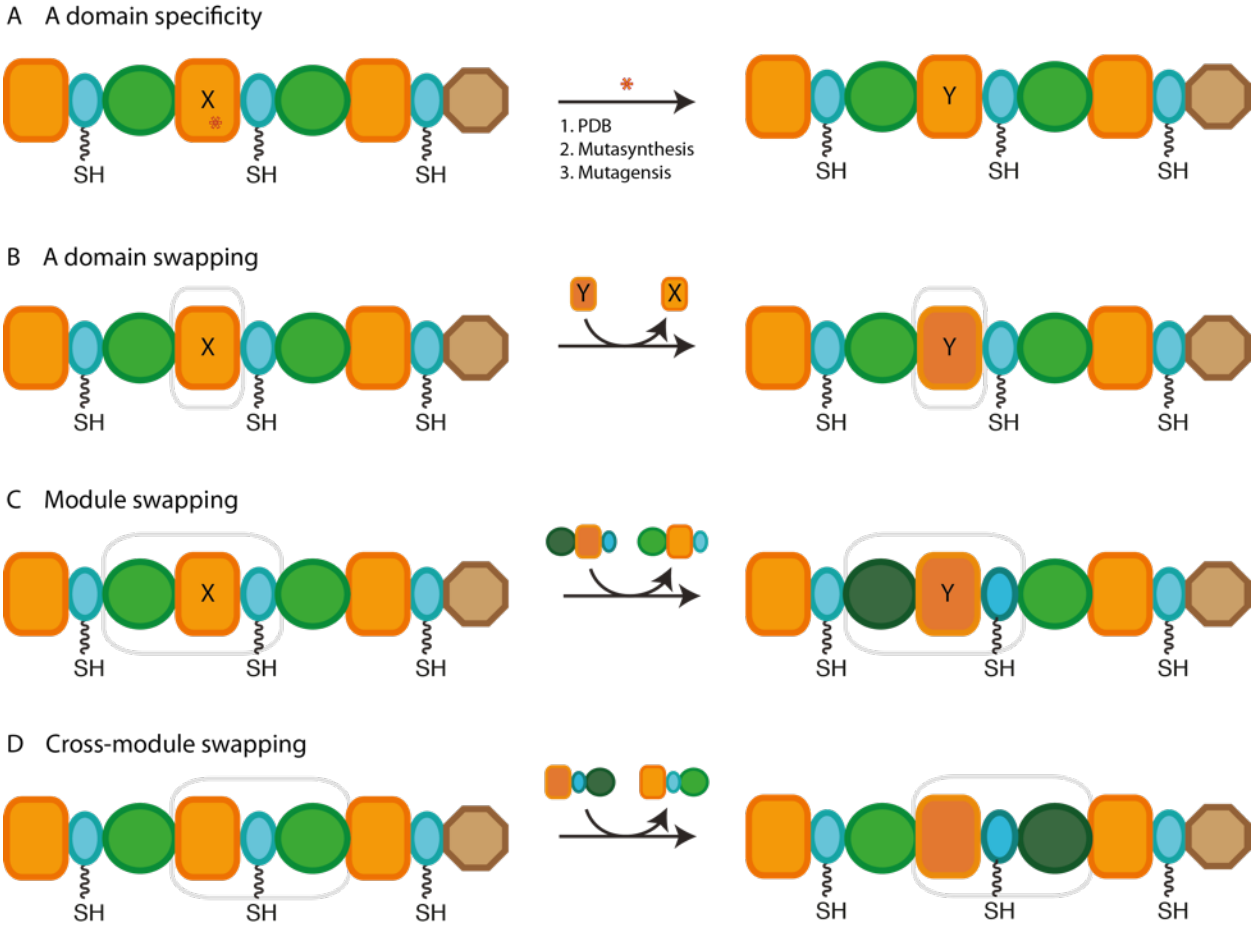


Figure 1.18 | Strategies for bioengineering NRPSs.
A, The substrate an A domain can be influenced by precursor directed biosynthesis or mutasynthesis, or altered by site-directed mutagenesis. **B, C,** New products can be created by domain or module swapping, respectively. **D**, Instead of module swapping, cross-module units can be exchanged to create new products.

into their catalytic cycles. Although precursor directed biosynthesis has only had limited success, it has the advantage that it does not require prior structural information on the target synthetase. Cyclosporin synthetase produces cyclosporins A, B, C, D and G, which are fungal undecappeptides and only differ in composition at position 2 of the peptide. Cyclosporin A is a clinically relevant immunosuppressant, and contains L-aminobutryic acid at position 2 (Tedesco and Haragsim, 2012). The relative ratios of cyclosporins being produced could be influenced using precursor directed biosynthesis. When cultures were supplemented with L-aminobutyric acid, production of all other cyclosporin forms was completely abolished, resulting in pure cyclosporin A (Kobel and Traber, 1982). A subsequent study created new forms of cyclosporin by introducing non-natural amino acids, L-β-cyclohexylalanine, DL-α-allylglycine and D-serine, at positions 1, 2 and 8, respectively, albeit with reduced yields (Traber et al., 1989). Mutasynthesis is a related technique and was developed to eliminate the problematic substrate binding competition inherent of precursor directed biosynthesis. The NRPS-containing organism is first genetically altered to remove a specific gene essential for production of the natural substrate, and then like precursor directed biosynthesis, cultures are supplemented with an alternative substrate to replace the missing natural substrate. The CDA synthetase in *Streptomyces coelicolor* produces a complicated cyclic lipopeptide containing a 2,3-epoxyhexanoyl fatty acid tail and several nonproteinogenic amino acid residues, including L-4-hydroxyphenylglycine (L-HPG). Adjacent to the CDA biosynthetic cluster are three genes, *hpgT, hmo* and *hmaS*, that are used to synthesize L-HPG. HmaS converts 4-hydroxyphenylpyruvate to L-4-hydroxymandelic acid, a precursor for L-HPG, and deletion of *hmaS* destroys CDA synthesis. When *S. coelicolor* Δ*hmaS* cultures were supplemented with L-4-hydroxymandelic acid derivatives, new variants of CDA were discovered with the arylglycine residue analogues incorporated into the L-HPG positions (Hojati et al., 2002).

1.7.2 Altering A domain specificity

After establishing the A domain specificity code, it was a logical progression to try to produce novel NRPs by altering the specificity of the A domain (Figure 1.18A). This was an attractive approach as it did not require disrupting the natural tertiary structure of the target

NRPS. As proof of concept, the specificities of the A domains in modules 1 and 5 of the surfactin A synthetase were modified using the specificity code as a guide, followed by recombinant production for *in vitro* activity assays (Eppelmann et al., 2002). A_1 domain, a natural Glu activating A domain, was successfully transformed into an A domain specific for Gln with catalytic rates equivalent to that of wildtype. Similarly, A_5 domain was switched from an Asp-activating A domain to an Asn-activating A domain. When the $A_{5\text{-}Asn}$ mutations were introduced back into the *Bacillus subtilis* surfactin biosynthetic cluster, a novel lipopeptide was synthesized *in vivo* with Asn in the wildtype Asp position. This strategy has been further expanded to facilitate the incorporation of non-proteinogenic amino acids into the final NRP: the Glu-activating A_{10} domain was altered to accept Gln and methyl-Gln residues in the CDA synthetase (Thirlway et al., 2012).

Despite success with altering A domain specificities by mutagenesis, the process is both labour and time intensive, making it unamenable to high-throughput screening. Evans *et al* aimed to overcome this bottleneck by using a directed evolution approach: using the andrimid synthetase as a model, they identified the 3 most divergent residues of the 10-residue specificity code and used limited saturation mutagenesis to target those 3 residues in AdmK, an A-PCP didomain protein (Evans et al., 2011). Clones were screened and identified using a multiplexed LC-MS/MS assay based on the specific fragmentation pattern of andrimid. This screening method has the additional benefit of simultaneously screening for productive interaction of the non-cognate substrate with downstream domains. Using this approach, four new andrimid analogues were identified in a total of more than 14 000 clones with activities comparable to the wildtype when combined with precursor directed biosynthesis methods.

1.7.3 Domain and module swapping

Domain and module swapping are simple in principal – a designer NRPS (or PKS) can be constructed by exchanging unwanted domains or modules with domains or modules that add a desired substrate or chemical modification to the engineered NRP (Figure 1.18B, C). This strategy has been referred to as the 'Lego-ization' of assembly line megaenzymes where domains and modules are the building blocks used to construct any NRPS and corresponding NRP imaginable (Sherman, 2005). However, these efforts have been hampered in some instances by the specificity

of the C domain with respect to the substrates (Belshaw et al., 1999), interactions with adjacent PCP domains (Kraas et al., 2012) and constrains of being embedded in a large NRPS (Duerfahrt et al., 2003). The daptomycin synthetase is an excellent example for module swapping as the synthetase has undergone extensive module swapping experiments. Daptomycin (trade name Cubicin) is a cyclic 13-mer lipopeptide antibiotic made from a three-subunit NRPS, DptA (Modules 1-5), DptBC (Modules 5-11) and DptD (Modules 12-13). In the first set of experiments, *dptD* was deleted, and complementary genes to *dptD* from similar lipopeptide synthetases, CDA or A54145, were added that maintained the Glu-selectivity of Module 12 but changed the substrate selected by Module 13 from kynurenine to Trp or Ile/Val, respectively (Miao et al., 2006, Coeffet-Le Gal et al., 2006). The daptomycin synthetase tolerated the subunit exchange and produced novel daptomycin analogues with the variant monomer at position 13, albeit with decreased efficiency and potency compared to the wildtype. A second set of experiments were designed to increase the potency of analogues by including an additional module swap with the original DptD swap: modules 8 and 11 were swapped with each other or with analogous modules 8 and 11 from CDA or A54145 (Nguyen et al., 2006). Again, although new analogues were synthesized by the chimeric daptomycin synthetases, none of the new products were more effective than daptomycin. The study was still an impressive example of the power of biocombinatorial bioengineering of NRPSs.

1.7.4 Cross-module swapping

A novel approach to module swapping was recently reported where instead of swapping C_n-A_n-PCP_n units, the swapping unit, defined as A_n-PCP_n-C_{n+1} , was constructed from neighbouring modules and was termed an exchange unit (XU) (Figure 1.18D) (Bozhuyuk et al., 2018). With several structures of full modules available, a linker connecting the C_n domain to the A_n domain was identified as a potential fusion point. NRPS chimeras were designed with the limiting condition that $XU_{A,n}$ could only be fused next to $XU_{B,n+1}$ if $XU_{B,n+1}$ activated the same amino acid as the natural downstream module of $XU_{A,n}$ (i.e. $XU_{A,n+1}$ and $XU_{B,n+1}$ activate the same residue), thus maintaining the specificity of the C domain and negating the potential deleterious effects of C domain specificity requirements. To test the plausibility of XUs, the XtpS NRPS, a four module synthetase that produces a xenotetrapeptide (a cyclic tetrapeptide named from the organism *Xenorhabdus*

nematophilia (Kegler et al., 2014)), was used as a model. The synthetase was reconstructed using XUs from parts of the GxpS, and KolS synthetases, as well as the terminal XtpS A-PCP-Te fragment. The fabricated XtpS produced 50% less product compared to the wildtype, which is still significantly higher than previous engineering attempts (Calcott and Ackerley, 2014, Stachelhaus et al., 1999, Kries et al., 2015, Winn et al., 2016, Mootz et al., 2002). Bozhüyük *et al* proceeded to create artificial NRPSs by fusing XUs from multiple sources together, which produced novel tetra- and penta-peptides.

1.8 Thesis objectives and overview

Nonribosomal peptide synthesis routinely includes modifications introduced through the action of tailoring domains, which results in chemically diverse products with unique structures and functions. An NRPS will often employ multiple tailoring domains throughout its synthetic cycle, yet despite their high prevalence within NRP synthesis, very little is known about how *in cis* tailoring domains are incorporated into the synthetic cycle and architecture of an NRPS. The location of a tailoring domain within a module can often be inferred through sequence alignments, but it cannot be predicted how it is integrated into the structural framework of the module. NRPSs must be highly adaptable to prevent synthesis from being disrupted, especially given the vast diversity of tailoring domain types.

Starting in Chapter 2, my initial aim was to solve the structure of the ~19 kDa formylation domain from the linear gramicidin synthetase LgrA initiation module. I was unable to crystallize the F domain from two different species, and altered my strategy to try solving the structure of the F domain within its native NRPS setting. I crystallized F-A and F-A-PCP in 4 different crystal forms and solved the structures of 5 different conformations of the initiation module that describe each catalytic step in the initiation synthetic cycle. The F and A domains are fused together using a hydrophobic patch and maintain an elongated conformation throughout each structure. The PCP and A_{sub} domains make large movements to transport substrate between the A and F domain active sites. Small-angle X-ray scattering was used to confirm that the movements of the A_{sub} and PCP domains were indeed possible in solution. Bioinformatics analysis identified the F domain to have evolved from a sugar formyltransferase (FT) source. Together, the entire synthetic cycle of a formylating initiation module is described using LgrA as a model.

With the F domain established as a sugar FT descendent, I wanted to retrace how a sugar formyltransferase could evolve into a productive tailoring domain within an NRPS setting. In Chapter 3, PseFT, a sugar formyltransferase from *Anoxybacilllus kamchatkensis*, was identified based on sequence homology as a putative pre-transfer sugar FT that would be prototypical of the original FT. The PseFT gene is located within a CMP-pseudaminic acid-like biosynthetic cluster where PseFT replaces the typical acetyltransferase. I characterized the activity of PseFT with its upstream enzymes, PseB and PseC, and show that PseFT is active on the precursor sugar UDP-4-amino-4,6-dideoxy-L-AltNAc using HPLC and NMR. I solved the structure of PseFT alone and with substrates, which are the first structures of a sugar FT that acts on nonulosonic acids. By learning how the F domain came to exist within an NRPS, future bioengineering experiments may be informed to create novel tailoring domains and thus designer NRPS products.

Lastly for Chapter 4, I continue the story of LgrA and follow the initiation module PCP as it donates formyl-valine to its adjacent elongation domain. Although it was suspected how a donor PCP would interact with a C domain, a structure depicting this important catalytic step did not exist. I solved the structure of dimodular LgrA with 5 accompanying structures of truncated constructs that contain the full initiation module and parts of the elongation module. Together with the Chapter 2 initiation module structures, the synthetic story of LgrA can be told showing each major catalytic step, including substrate donation and condensation. With multiple conformations of the modules observed within the same state, I proposed that NRPSs have a flexible and dynamic overall architecture. The specificity requirements of substrate donation were also probed using mutagenesis and LC-MS.

CHAPTER 2 | SYNTHETIC CYCLE OF THE INITIATION MODULE OF A FORMYLATING NONRIBOSOMAL PEPTIDE SYNTHETASE

Reimer JM, Aloise MN, Harrison PM, Schmeing TM. (2016). Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. Nature, 529 (7585):239-242.

2.1 Abstract

Nonribosomal peptide synthetases (NRPSs) are massive proteins that produce small peptide molecules with wide-ranging biological activities, including green chemicals and many widely-used therapeutics (Walsh, 2004). NRPSs are true macromolecular machines, with modular assembly-line logic, a complex catalytic cycle, moving parts and many active sites (Weissman, 2015), (Hur et al., 2012). In addition to the core domains required to link the substrates, they often include specialized tailoring domains, which introduce chemical modifications and allow the product to access a large expanse of chemical space (Hur et al., 2012, Walsh et al., 2001). It is still unknown how any of the NRPS tailoring domains are structurally accommodated into the megaenzymes or how they have adapted to function in nonribosomal peptide synthesis. Here we present a series of crystal structures of the initiation module of an antibiotic-producing NRPS, linear gramicidin synthetase (Kessler et al., 2004, Schoenafinger et al., 2006). This module includes the specialized tailoring formylation domain, and we capture states that represent every major step of the assembly-line synthesis in the initiation module. The transitions between conformations are staggering, with both the peptidyl carrier protein and the adenylation subdomain undergoing huge movements to transport substrate between distal active sites. The structures highlight the great versatility of NRPSs, as small domains repurpose and recycle their limited interfaces to interact with their various binding partners. Understanding tailoring domains is important if NRPSs are to be exploited for production of novel therapeutics.

2.2 Introduction, results and discussion

Tailoring domains embedded within NRPSs are vital for the production and bioactivity of these synthetases' nonribosomal peptide (NRP) products (Walsh et al., 2001). Tailoring domains exist in addition to the core NRPS adenylation (A), peptidyl carrier protein (PCP) and condensation (C) domains, which a module requires to add an amino acid to the growing NRP: the A domain selects, activates and transfers the substrate amino acid to the PCP domain, which transports it to the C domain for peptide bond formation (Walsh, 2004) (Fig. 2.1, Extended Data Fig. 2.1). Tailoring domains are common in NRPSs (Hur et al., 2012, Walsh et al., 2001). For example, cyclosporin synthetase contains methyltransferase domains (Lawen and Zocher, 1990); daptomycin

("Cubicin") synthetase, epimerization domains (Robbel and Marahiel, 2010); bactitracin ("BACiiM") synthetase, a heterocyclization domain (Konz et al., 1997); valinomycin synthetase, ketoreductase domains (Cheng, 2006); bleomycin synthetase, an oxidase domain (Schneider et al., 2003); and soframycin synthetase, a reductase domain (Koketsu et al., 2012). These domains impart key functionalities into the NRP, by, for example, providing protease resistance, enabling novel interactions, improving affinity by limiting NRP conformational flexibility, or allowing the NRP to assume its active conformation. Linear gramicidin synthetase (LgrA-D) was found by Marahiel

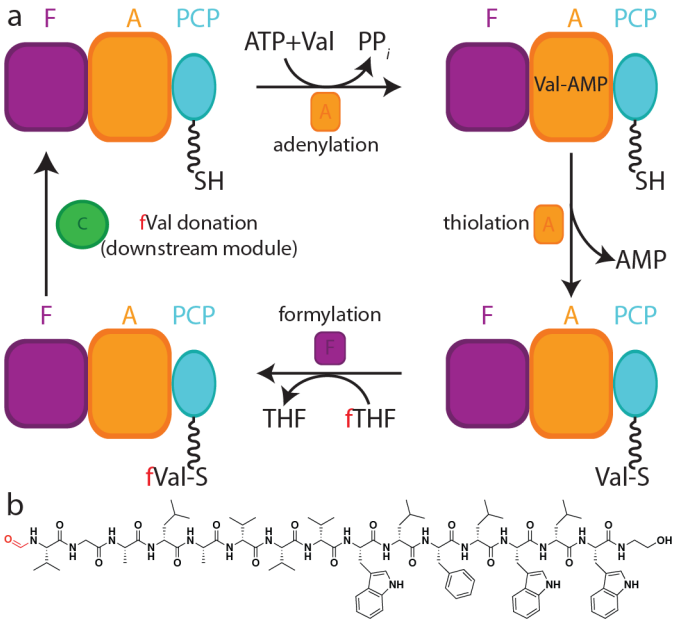


Figure 2.1 | A schematic of the action of the linear gramicidin synthetase initiation module.
a, The F-A-PCP initiation module is the first module of LgrA, the dimodular F-A-PCP-C-A-PCP-E*NRPS protein in the LgrA-E synthetic cluster (E*, inactive epimerization domain). The initiation cycle begins with valine selection and adenylation followed by thiolation onto the PPE arm of the PCP domain. The F domain formylates PCP-PPE-Val before it is brought to be the donor in the condensation reaction of the downstream module. **b**, Chemical structure of linear gramicidin A.

and colleagues to contain an active formylation (F) domain as the first domain of its F-A-PCP initiation module (Kessler et al., 2004, Schoenafinger et al., 2006) (Fig. 2.1). F domains are homologous to formyltransferase (FT) proteins that modify substrates in three diverse pathways: ribosomal translation (Schmitt et al., 1998), purine anabolism (Almassy et al., 1992) and bacterial

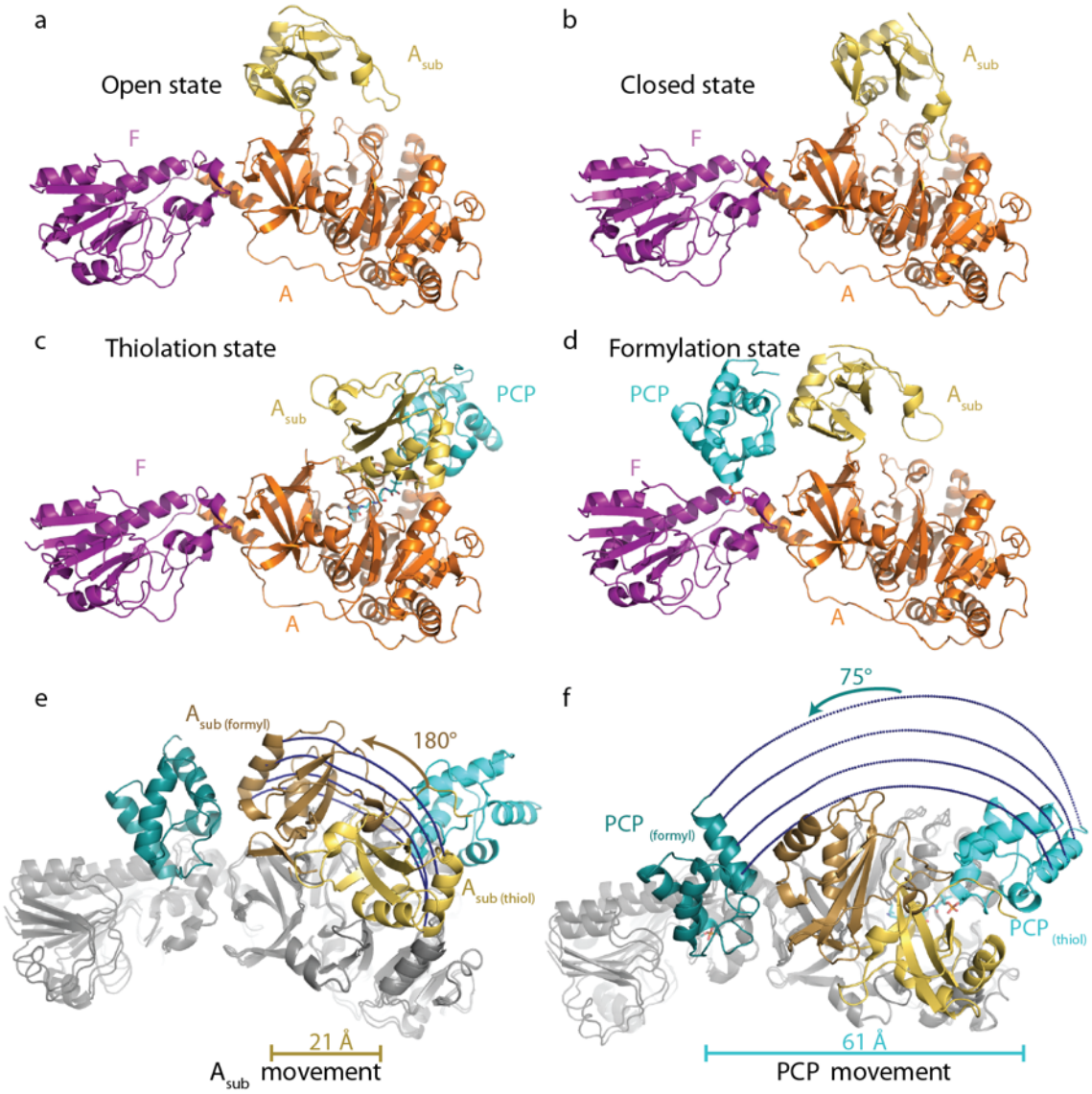


Figure 2.2 | Crystal structures representing the steps of the synthesis cycle in the LgrA initiation module.

a-d, The F-A-PCP LgrA initiation module in open (**a**), closed (**b**), thiolation (**c**) and formylation (**d**) states. (The PCP domain is not necessary for the open and closed states and is disordered in **b** and **c**.) The transition between thiolation and formylation states requires massive rigid body movements of both the A_{sub} and PCP domains. **e**, The PCP domain rotates 75° and translocates its center of mass by 61 Å. PPE arm attachment point, Ser729, moves 52 Å, and some residues move > 80 Å. **f**, The A_{sub} domain rotates 180° and translocates its center of mass by 21 Å.

outer membrane synthesis (Thoden et al., 2013). The LgrA initiation module must formylate its substrate for linear gramicidin synthesis to proceed (Schoenafinger et al., 2006) (Fig. 2.1), and this

formyl group is essential for gramicidin's clinically important antibacterial activity (Wallace, 2000). Gramicidin molecules form head-to-head dimers through the formyl group to make a β helical pore in gram-positive bacterial membranes. This pore freely allows passage of monovalent cations, destroying the ion gradient and killing the bacteria.

We have determined four independent crystal structures of the initiation module of LgrA, at 2.5, 2.6, 2.8 and 3.8 Å resolutions (Extended Data Table 2.1, Extended Data Fig. 2.2), that show four different functional conformations: the A domain open (substrate binding), A domain closed (adenylation), thiolation and formylation states (Fig. 2.2, Extended Data Fig. 2.3, Supplemental Video 2.1). This augments the excellent existing structural knowledge of NRPSs, (reviewed by Weissman, (Weissman, 2015)) by visualizing the structure of an NRPS module that includes a tailoring domain, showing how the tailoring domain is incorporated into and used as part of an NRPS, observing several functional states (open, closed, thiolation) in a single protein, rather than over different excised mono- and di-domains (Weissman, 2015, Conti et al., 1997, Yonus et al., 2008, Reger et al., 2008, Gulick, 2009, Mitchell et al., 2012), and visualizing a novel functional state (formylation).

The F domain is connected to the rest of the F-A-PCP LgrA initiation module through an interface with the A domain (Fig. 2.3) that buries 830 Å^2 of surface area. This is distinct from the C-A interface in C-A-PCP elongation and termination modules (Tanovic et al., 2008) (Extended Data Fig. 2.4). The F-A interface appears sufficient to maintain these domains in a very elongated conformation (Fig. 2.2). Across all nonequivalent molecules in the crystals, the relative orientation between the two domains varies only by ~5°, and our small angle X-ray scattering analysis indicates that this extended conformation is representative of the initiation module in solution (Extended Data Fig. 2.5). This architecture means that the adenylation active site and the formylation active site are always ~50 Å apart, necessitating that valine substrate travel a large distance between subsequent steps in synthesis. Accordingly, positions of the PCP domain and the A_{sub} domain (C-terminal portion of the A domain) change markedly in the module's progression through functional states.

The NRPS assembly-line process (Fig. 2.1, Extended Data Fig. 2.1, Supplemental Video 2.2) begins with ATP and valine binding to an open conformation of the A domain (Gulick, 2009) (Fig.

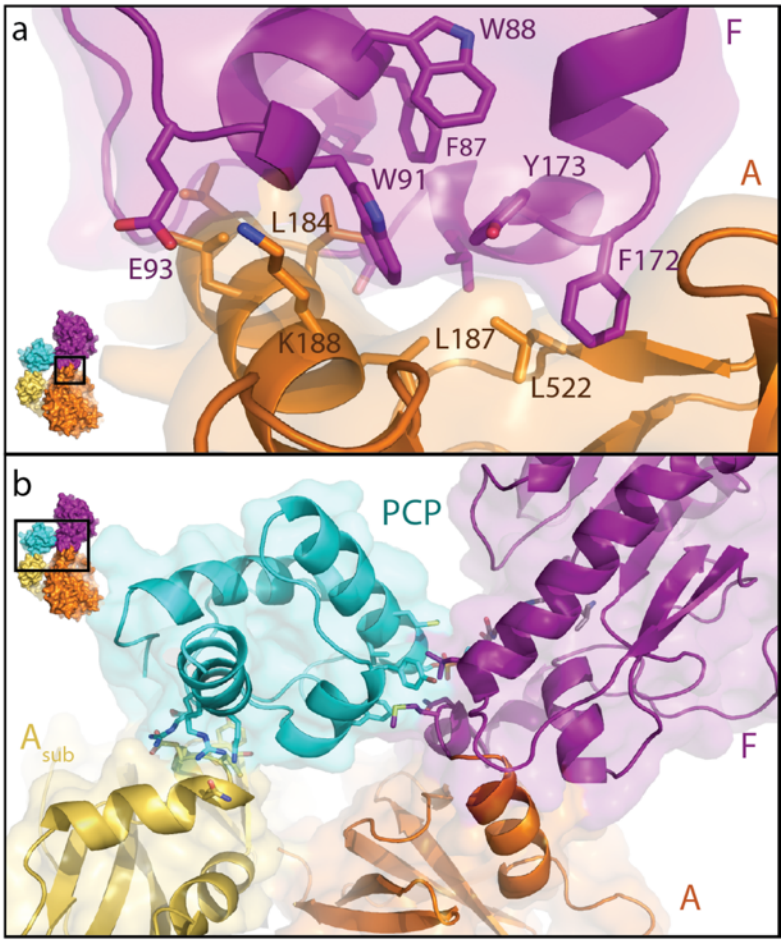


Figure 2.3 | Interdomain interfaces of the initiation module.
a, The F domain is fused onto the A domain and forms a small hydrophobic core (Extended Data Fig. 2.8). **b**, Interaction of the PCP domain with A_{sub} and F domains in the formylation state. The A_{sub} domain creates an electrostatic platform for the PCP domain. The PCP domain binds to F domain hydrophobic residues Leu127 (often Lys or Glu in FTs) and Met178 (in the C-terminus of the F domain that is not similar to formyltransferasres). The PPE phosphate interacts with Arg170 (often Glu, Ser or Asn in FTs) and Asn177 (usually Glu, Asp or Met in formyltransferases).

2.2a, Supplemental Video 2.1). The A domain closes upon substrate binding by rotating the A_{sub} by ~30° to catalyze formation of the valine adenylate (Conti et al., 1997, Yonus et al., 2008) (Fig. 2.2b). Next, the thiolation reaction transfers the valine from the adenylate to the thiol of the PCP domain's phosphopantetheine arm (PPE). We accessed this state by attaching a non-hydrolyzable analog (Liu and Bruner, 2007) of the product of the reaction, valine-NH-PPE, to the PCP domain. The resulting structure shows the known 140° rotation of the A_{sub} (Reger et al., 2008, Mitchell et al., 2012) and the product valine-NH-PCP still bound to the active site (Fig. 2.2c). The PCP must

now transport its valine 50 Å between A and F domain active sites to accept a formyl group. Our next structure (Fig. 2.2d) shows that to achieve this, the PCP domain makes a massive movement of a rigid ~75° rotation and 61 Å translocation (Fig. 2.2e). The ~10 residue linker between A and PCP domains is not nearly sufficient to span the 55 Å travelled by the first residue of the PCP domain; accordingly, the A_{sub} domain undergoes a full 180° rotation and 21 Å translocation to allow PCP domains to bind the F domain (Fig. 2.2f). There, valine-PCP accepts a formyl group from the donor cofactor formyl-tetrahydrofolate (fTHF) (Extended Data Fig. 2.2f) onto its amino group (Kessler et al., 2004, Schoenafinger et al., 2006). The PCP will then move the formyl-valine to the next module, where that module's condensation (C) domain will catalyze peptide bond formation between fVal-PCP and its glycine-PCP2, making the first peptide bond of linear gramicidin and liberating the PCP to participate in the next round of reactions (Schoenafinger et al., 2006) (Supplemental Video 2.2).

How did the F domain become a functional NRPS domain? LgrA's F domain was fused into an existing NRPS (Kessler et al., 2004), and we suggest that the pre-transfer source was a single domain FT from a distantly related bacterium with a signature of missing helix α2 and strand β3. As the high incidence of horizontal transfer (Extended Data Fig. 2.6) is consistent with conferring a competitive advantage, and bacteria possessing FTs similar to the F domain also have canonical tRNA and phosphoribosylglycinamide FTs, it is likely that the pre-transfer FT performed the remaining known FT function, sugar formylation for cell wall synthesis. After fusion, the F-A-PCP initiation module evolved rapidly (Extended Data Fig. 2.6, 2.7). The fold of the first 171 amino acids of LgrA is conserved with the sugar FTs, leaving only residues 172-179, including a single α-helix, as a new structural element and link to the A domain (Extended Data Fig. 2.8). A "landing pad" evolved to include a hydrophobic patch for binding the PCP domain, and positive residues and hydrogen bond donors to interact with the PPE phosphate (Fig. 2.3, Extended Data Fig. 2.8). The F-PCP interaction places the PPE attachment point, Ser729, an ideal 16 Å away from the fTHF in the conserved FT active site. Interestingly, this positions the valine-PPE exactly in the sugar-dTDP binding site of sugar FTs (Kessler et al., 2004, Thoden et al., 2013, Rouhiainen et al., 2000) (Fig.

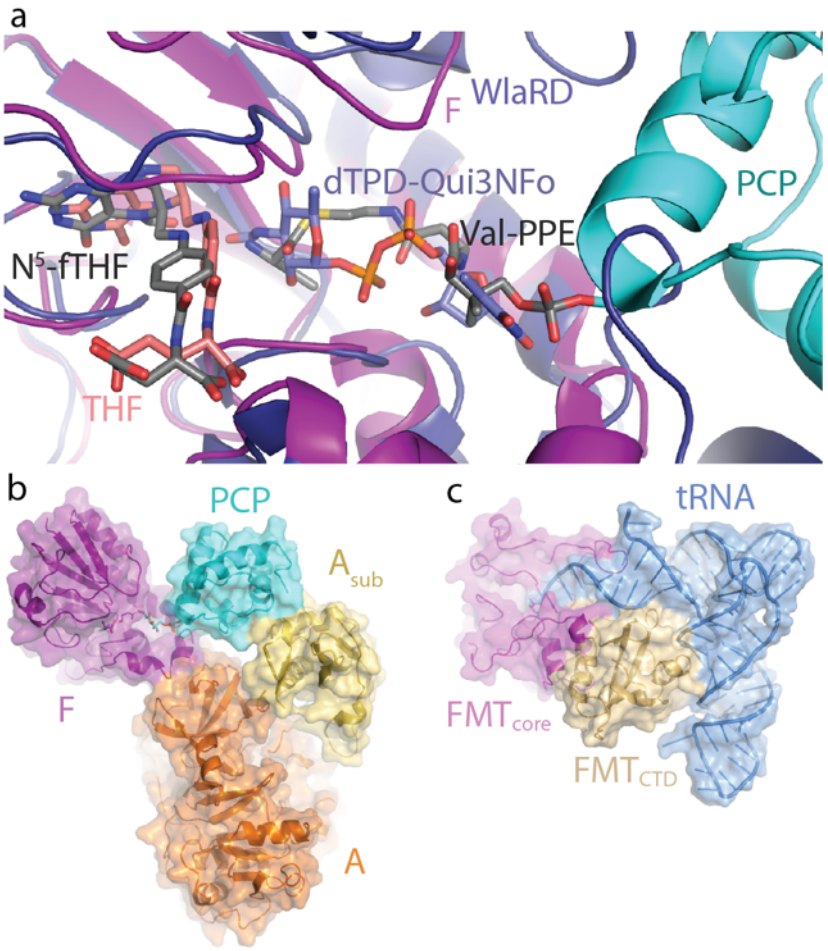


Figure 2.4| Comparisons of the F domain to sugar and tRNA formyltransferases.
a, The binding mode for the PPE arm to the F domain is similar to that of sugar-dTDP in sugar formyltransferases (protein WlaRD, PDB 4LY3 (Thoden et al., 2013)). Note that the valine and most of the PPE arm (carbons shown in grey) are modelled, as they are not visible in electron density maps at 3.8 Å resolution. **b, c**, The A_{sub} domain emulates the positioning role of the FMT_{CTD} in methionyl-$tRNA^{fMet}$ formyltransferases (PDB 2FMT (Schmitt et al., 1998)). Excluding the PPE arm, the PCP domain buries only 279 $Å^2$ of F domain surface. The A_{sub} provides an additional 345 $Å^2$ of interaction surface to position the PCP domain.

2.4a). The similar length and hydrophilic nature of the sugar-dTDP and valine-PPE likely enabled the F domain to formylate valine-PCP soon after the fusion event, before formylation was absolutely required for downstream peptide synthesis to proceed.

The PCP domain interaction with the F domain is quite minimal, and accordingly, the A_{sub} domain donates an additional binding interaction in the formylation state (Kessler et al., 2004) (Fig. 2.3b). This is very reminiscent of methionyl-$tRNA^{fMet}$ formyltransferase (FMT), the essential bacterial two-domain FT that uses its C-terminal domain (FMT_{CTD}) to present the methionyl-

$tRNA^{fMet}$ to the FT active site (Schmitt et al., 1998) (Fig. 2.4b,c). This functional convergent evolution presents yet another interesting parallel to the completely separate macromolecular system that synthesizes peptides, ribosomal translation. In both LgrA and the ribosome, a mobile carrier macromolecule (PCP domain / tRNA) covalently (through thioester / ester bonds) transports an amino acid to a formyltransferase enzyme (F domain/FMT), where the carrier is oriented by a positioning domain (A_{sub}/FMT_{CTD}) to allow formylation, before acting as the first donor substrate for a peptidyl transferase enzyme (C domain/large ribosomal subunit).

Observing the same protein all these conformations, including the novel formylation conformation, highlights and further demonstrates the great versatility of the small domains in NRPSs (Weissman, 2015, Hur et al., 2012, Yonus et al., 2008, Liu and Bruner, 2007) . The small ~100 residue A_{sub} has three distinct roles in the cycle: providing catalytic residues for the adenylation reaction (Yonus et al., 2008); positioning the PCP for the thiolation reaction and later for the formylation reaction, and bridging the distance between the active sites the PCP must visit (Mitchell et al., 2012, Tanovic et al., 2008). The A_{sub} uses different surfaces for each of these roles (Extended Data Fig. 2.9a). In addition, the F domain adds to the long list of partners with which the equally small PCP domain must interact (A, F, C, TE, all tailoring domains), and it performs its tasks with overlapping surfaces (Lohman et al., 2014) (Extended Data Fig. 2.9b).

Adapting a formyltransferase has further increased the functionality of NRPSs. The formyl functionality seems to be useful in nonribosomal peptides, as F domains have been incorporated into NRPSs multiple independent times: the F domains in kolossin A synthetase (Bode et al., 2015), anabaenopeptilide synthetases (Rouhiainen et al., 2000), the oxazolomycin synthetase (Zhao et al., 2010) and a dozen orphan NRPSs and NRPS-PKSs arose from a separate fusion event with an FMT and display a different FMT-FMT_{CTD} -$C_{partial}$-A-PCP domain sequence in their initiation modules (Kessler et al., 2004). Sampling additional chemical space can lead to novel or improved activity in nonribosomal peptides, which inspires many bioengineering experiments on NRPSs (Clardy et al., 2006) aimed to meet the dynamic challenges to human health. NRPSs are well placed to be engineered for production of new compounds because their synthetic scheme is conceptually straightforward, and NRPSs already naturally produce many therapeutics, as well as promising NRPs like teixobactin (Ling et al., 2015) and piperidamycin (Hosaka et al., 2009), two recently

discovered, first-in-class compounds with strong antibacterial activity. The structures presented here reveal the interface between the F and A domains and show all the interactions that the PCP domain makes in the LgrA initiation module. This knowledge could substantially facilitate our ability to introduce an F domain into a foreign NRPS, and make formylation an accessible tool in the NRPS bioengineering toolkit.

2.3 Acknowledgements

We thank Thierry Mintya, Daina Avizonis, Marie-Christine Tang and Alexandra Furtos for experimental support, Robert Zamboni, John Collucci and Kimiaka Guerard for small molecule synthesis assistance, Jessie Jiang, Diego Alonzo and Dmitry Rodinov for experimental advice and assistance, Shaun Labuik and Pawel Grochulski (Canadian Light Source) for diffraction data collection, Richard Gillian (CHESS SAXS beamline G1), Monica Pillon and Alba Guarne for SAXS help, Kristjan Bloudoff, Michael Tarry, Asfarul Haque and Albert Beghuis for helpful discussions and Jerry Pelletier, Nancy Rogerson and Ali Nahvi for critical reading of the manuscript. This work was supported by CIHR grant 106615, a HFSP CDA and a Canada Research Chair in Macromolecular Machines to T.M.S.. J.M.R. is supported by an NSERC Alexander Graham Bell studentship, and M.N.A. by a studentship from the CIHR Training Grant in Chemical Biology.

2.4 Author Information

Crystallography data have been deposited in the Protein Data Bank under accession codes 5ES5, 5ES6, 5ES7, 5ES8 and 5ES9.

2.5 Materials and Methods

2.5.1 Cloning of linear gramicidin synthetase initiation module constructs

Genomic DNA was isolated from *Brevibacillus brevis* ATCC 8185 (Cedarlane Laboratories) using a GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich). Gene constructs comprising F and A domains (F-A) and all three domains (F-A-PCP) were amplified by PCR from the *lgrA* gene using the following primers, designed using sequence alignment with A and PCP domains of known structure and the study of Marahiel and coworkers: FA_fwd

AATCATCCATGGGAAGAATACTATTCCTAACAACATTTATGAGCAAAG; FA_rev AATCATCTCGAGTTACGCATCGGCCTGCACGTCT; FAT_fwd TGACTACCATGGGGAGAATACTATTCCTAACAACATTTATGAGC; FAT_rev CGTTGAGCGGCCGCTTGCTCCGTAAGCAGACGTTT. PCR product for F-A-PCP was digested using NcoI and NotI (New England Biolabs) and ligated into a pET21-derived vector containing an N-terminal octa-histidine tag with a tobacco etch virus (TEV) protease cleavage site. PCR product for F-A-PCP was cloned between NcoI and NotI restriction sites into a pET21-derived vector containing an N-terminal TEV cleavable octa-histidine tag and a C-terminal TEV cleavable calmodulin binding peptide (CBP) tag. Point mutations were introduced into the construct of F-A-PCP using the QuikChange (Agilent) site-directed mutagenesis kit.

2.5.2 Expression and purification of proteins

The F-A protein was expressed in E. coli BL21 (DE3) cells. A 10 ml aliquot of overnight culture was used to inoculate 1 L of LB medium supplemented with 350 µg ml-1 kanamycin. The culture was grown at 37 °C to an OD600 of 0.6, before inducing protein expression using 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and reducing the temperature to 16 °C for 18h. Cells were harvested by centrifugation at 4 °C and resuspended in nickel binding buffer (2 mM imidazole, 150 mM NaCl, 0.25mM TCEP, 50 mM Tris-HCl, pH 7.0). The cells were lysed by sonication on ice and centrifuged for 30 min at 20 000 x g at 4 °C. Clarified lysate was loaded onto a HiTrap IMAC FF column (GE Healthcare). F-A was eluted using a gradient of 2-250 mM imidazole. Fractions containing F-A were pooled, diluted 10-fold with ion exchange binding buffer (0.25 mM TCEP, 20 mM Tris, pH 8.0), loaded onto a HiTrap Q HP column and eluted using a gradient to 100% elution buffer (1M NaCl, 0.25 mM TCEP, 20 mM Tris-HCl, pH 8.0). The eluted protein was concentrated using a 10k MWCO Amicon Ultra-15 filtration unit (EMD Millipore) and subjected to gel filtration chromatography using a HiLoad 16/600 Superdex 200 column (GE Healthcare) equilibrated with S200 buffer (150 mM NaCl, 0.25 mM TCEP, 20 mM Tris, pH 7.0). Protein purity was confirmed using SDS-PAGE and native PAGE. Pure F-A was concentrated in storage buffer (25% glycerol, 150 mM NaCl, 0.25 mM TCEP, 20 mM Tris, pH 7.0), flash frozen with liquid nitrogen and stored at -80 °C for later use.

F-A-PCP was expressed in E. coli BL21 EntD-(DE3) cells using the same protocol as above. Cells were pelleted, resuspended in CBP binding buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM imidazole pH 8.0, 2 mM CaCl2, 2 mM β-mercaptoethanol (βME), and 0.1 mM phenylmethanesulfonyl fluoride (PMSF)), sonicated and clarified by centrifugation for 30 min at 20 000 x g at 4 °C. Clarified lysate was loaded onto a 30 ml calmodulin sepharose 4B column (GE Healthcare). F-A-PCP was eluted with elution buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM EGTA, 2 mM βME, and 0.1 mM PMSF). For biochemical assays of F-A-PCP and mutants, protein was pooled, concentrated in storage buffer, flash-frozen in liquid nitrogen and stored at -80 °C. For crystallographic studies, F-A-PCP was further purified as follows. Protein was dialyzed against binding buffer for a minimum of 4 hours before being loaded onto a 5 ml HiTrap IMAC FF column (GE Healthcare) charged with Ni2+ and equilibrated in nickel binding buffer. F-A-PCP was eluted using a 60 mL gradient of 0 – 250 mM imidazole. Fractions containing F-A-PCP were pooled and affinity tags were removed by cleavage with TEV protease at room temperature overnight using a 1:4 mg ratio of TEV to F-A-PCP. Cleaved F-A-PCP was passed back over the nickel and calmodulin affinity columns, with the flow-through collected, concentrated, and applied to a HiLoad 16/600 Superdex 200 (GE Healthcare) in S200 buffer. Pure F-A-PCP was concentrated to 5.0 mg ml-1 in storage buffer, flash-frozen in liquid nitrogen and stored at -80 °C.

2.5.3 Substrate syntheses

Amino-coenzyme A (amino-CoA) (Liu and Bruner, 2007) was prepared enzymatically starting from amino-pantetheine (WuXi AppTec) using a previously published protocol (Nazi et al., 2004) with the following modifications: one-pot synthesis was carried out at pH 9.0, the amounts of DPCK and ATP were doubled to 9.8 mg and 30 mM, respectively; and the enzymes were removed using a 10k MWCO Amicon Ultra-15 filtration unit (EMD Millipore). An ATP regeneration system using 0.1 mg/ml pyrophosphatase (Roche), 30 mM phosphoenolpyruvate, and 0.1 mg/ml pyruvate kinase (Roche) was also included. The filtrate containing amino-CoA was purified on a preparative reverse-phase C18 HPLC (35 ml/min; 0-4 min, 0% B; 4-9 min, 0-98% B, where A is 0.1% trifluoroacetic acid (TFA, Sigma-Aldrich) in H_2O and B is 0.1% in acetonitrile (ACN, Sigma-Aldrich)). Amino-CoA eluted at 7 min and was lyophilized to dryness.

Valine-amino-coenzyme A (Val-NH-CoA) (Liu and Bruner, 2007) was synthesized by coupling 1 molar equivalent of amino-CoA with 8 molar equivalents of tert-Butoxycarbonyl-L-valine-N- hydroxysuccinimide ester (Boc-Val-OSu, Sigma-Aldrich) in N,N-dimethylformamide (DMF, Sigma-Aldrich) with 4 molar equivalents of N,N-diisopropylethylamine (DIPEA, Sigma-Aldrich) overnight with stirring. Boc-Val-NH-CoA was purified using the above chromatographic profile and lyophilized to dryness, then deprotected using 1.5 ml 95% TFA/2.5% H2O/2.5% triisopropylsilane (TIPS, Sigma-Aldrich). The deprotection mix was agitated for 2 h at 25 °C in a thermomixer at 700 rpm before being transferred to 20 ml ice-cold diethyl ether and incubated at -20 °C for 2 h. The solution was centrifuged and the pellet was redissolved in 5% aqueous ACN solution and purified with the same protocol as amino-CoA. Compound identity was verified by mass spectrometry and NMR (Supplemental Data 1).

2.54 Loading phosphopanetheinylates on the PCP domain

Unmodified F-A-PCP was converted to valine-NH-F-A-PCP by incubating 25 µM apo-F-A-PCP with 5 µM Sfp, 0.25 mM valine-NH-CoA, 10 mM $MgCl_2$ and 25 mM Tris pH 7.0 for a minimum of 4 h at 25 °C. To remove Sfp for subsequent crystallization trials, the reaction mix was loaded onto a Superdex S75 10/300 GL (GE Healthcare Life Sciences) equilibrated in 25 mM Tris pH 7.5, 150 mM NaCl, and 2 mM βME.

2.5.5 SAXS

Inline SEC-SAXS data was collected on the G1 beamline at the Macromolecular Diffraction Facility at the Cornell High Energy Synchrotron Source (Acerbo et al., 2015) (Skou et al., 2014) at 9.963 keV (1.244 A) at $7.89x10^{11}$ photons/s. The X-ray beam was collimated to 250x250 µm and the sample cell path length was 2 mm. The G1 beamline was outfitted with a GE AKTA purifier with a GE Superdex 200 5/150 GL column and 50 µl sample loop. The column was equilibrated in 25 mM Tris pH 7.5, 150 mM NaCl, and 2mM BME and the samples were centrifuged for 10 minutes prior to sample injection. Images were recorded on a Pilatus 100K-s detector and normalized using beam stop photodiode counts. F-A-PCP eluted in a single monomeric peak and eleven peak exposures were averaged using BioXTAS RAW software (Nielsen et al., 2009). A buffer scattering

curve was created by averaging the first eleven exposures after injection, and this scattering curve was subtracted from the F-A-PCP scattering curve to yield the corrected scattering curve for F-A-PCP. *Ab initio* models were generated by first creating pairwise distribution functions (P(r)) with GNOM (Svergun, 1992), leading to twenty independent bead models produced by DAMMIF (Franke and Svergun, 2009). Models were aligned, averaged, and filtered using DAMAVER (Volkov and Svergun, 2003) assuming P1 symmetry. All DAMMIF models were included in the final DAMAVER model. They had a mean NSD value of 0.82 +/- 0.052. CRYSOL (Svergun et al., 1995) was used to check how well the final model fit with our crystal structures. Flexibility was analyzed using EOM (Tria et al., 2015, Otwinowski and Minor, 1997), whereby crystal structures of F, A_{core}, A_{sub}, and the PCP were used to generate a pool of 10,000 models.

2.5.6 Crystallography

To obtain the crystal structures described in this study, genes from 4 species, of up to 4 domain constructs each (F, F-A, F-A(ΔA_{sub}), F-A-PCP) were cloned and assayed for heterologous expression. Purification was undertaken for all well-expressing proteins and crystallization trials were performed, including trials using protein with affinity tags removed or retained, and in the presence or absence of a variety of ligands (ATP, AMPcPP, AMP, valine, THF, N^5-fTHF, phosphopantetheine, valine amino phosphopantetheine, valine vinyl sulfonamide adenylate, dead-end THF analog). Up to 4032 crystallization conditions were assayed per protein sample, and gave a total of ~50 "hits", 6 of which were successfully optimized to allow structure determination. Together, 4 of these crystal structures (F-A in crystals of space group $P4_12_12$, F-A-PCP in R3:H, F-A-PCP-PPE-NH-Val in $P2_1$ and F-A-PCP-PPE in $P3_22$), plus an additional structure including ligands soaked into F-A $P4_12_12$ crystals, captured the states that represent every major step of the assembly-line synthesis in the LgrA initiation module and are presented here.

The final crystallization conditions were optimized in 24-well sitting drop plates, with 2 µl protein sample plus 2 µl reservoir solution in the drop and a 500µl reservoir volume and are as follows: "F-A" and "F-A soak": protein LgrA F-A (10 mg/ml) was crystallized using a precipitant solution of 2 M Na-formate, 0.1 M Na-Acetate pH 5.3 into space group $P4_12_12$. "F-A-PCP" (open and closed states): protein LgrA F-A-PCP (5 mg/ml) was crystallized using a precipitant solution of

0.92 M $AmSO_4$, 0.1 M bis-Tris pH 5.5, 1% PEG 3350 into space group R3:H. "F-A-PCP-NH-Val" (thiolation state): protein F-A-PCP-PPE-NH-Val (4.7 mg/ml) was crystallized using a precipitant solution of 12% PEG 20 000, 0.1 M MES pH 6.7 into space group $P2_1$. "F-A-PCP-PPE" (formylation state): protein F-A-PCP-PPE (5.5 mg/ml) was crystallized using a precipitant solution of 1 M $AmSO_4$, 0.1 M bis-Tris pH 5.5, 3% PEG 3350 into space group $P3_22$.

Solutions of mother liquor with increasing amounts of glycerol (5%, 10%, 25%) were used to replace the drop solution for cryoprotection. For soaking with the N^5-fTHF, valine and AMPcPP, 10mM of each was included in the final cryoprotection solution and incubated for 30 min. (Marahiel and coworkers showed that LgrA uses commercially available N^5-fTHF in addition to its natural substrate, N^{10}-fTHF (Schoenafinger et al., 2006).) Crystals were flash-cooled in liquid nitrogen and diffraction data sets collected at 200 K using beamline 8 of the CMCF at the Canadian Light Source (λ= 0.979 Å) in Saskatoon, Canada.

All datasets were integrated and scaled using the programs HKL-2000 (Otwinowski and Minor, 1997) and iMosflm (Leslie and Powell, 2007). Structure determination of F-A in the $P4_12_12$ space group was performed by molecular replacement using a search model of the A domain from gramicidin Soviet synthetase (1AMU (Conti et al., 1997); note that linear gramicidin and gramicidin Soviet are made by different NRPSs) with the A_{sub} subdomain removed and side chains trimmed to the beta carbon, in the program Phaser (McCoy et al., 2007). Density for the F domain was visible in the resulting maps. Iterative building in the program COOT (Emsley et al., 2010) and refinement in the program Phenix (Adams et al., 2010) produced the final F-A structure. This structure was then used as a search model to determine the structure of F-A-PCP in space groups $P3_22$, R3:H, and $P2_1$ by molecular replacement using the program Phaser, followed by iterative building in the program COOT and refinement in the programs Phenix and CNS (Brunger, 2007). The highest resolution shell CC* values are: $P4_12_12$ – 0.845; $P4_12_12$ (soak) – 0.897; $P3_22$ – 0.883; R3:H – 0.822, and $P2_1$ – 0.822. The quoted resolution of each structure represents the CC 1/2 of the diffraction data (Leslie and Powell, 2007, Evans and Murshudov, 2013).

2.5.7 Bioinformatics

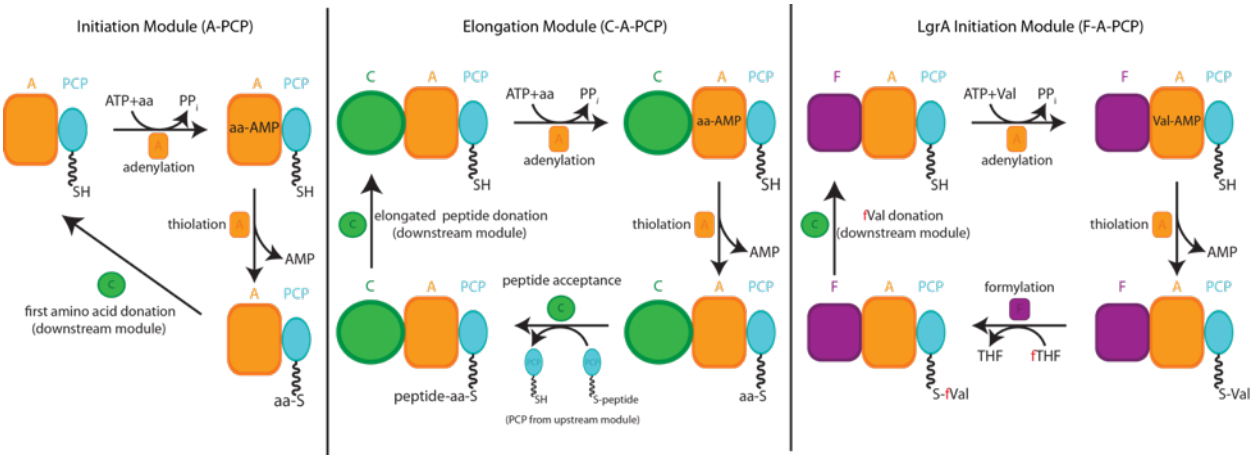
Multiple sequence alignments (MSAs) were constructed using Clustal Omega (Sievers et

al., 2011) (ebi.ac.uk/Tools/msa/clustalo) and PROMALS3D (Pei and Grishin, 2014) (prodata.swmed.edu/promals3d), following database searches using BLAST (Altschul et al., 1990) (ncbi.nlm.nih.gov/blast). MSAs were drawn/edited using Jalview (Waterhouse et al., 2009) (www.jalview.org). PHYLIP (evolution.genetics.washington.edu) was used to make neighbour-joining trees bootstrapped with 100 replicates, and FigTree (tree.bio.ed.ac.uk) was used to draw them. WebLogo (Crooks et al., 2004) (weblogo.berkeley.edu) was used to draw sequence logos of residue groupings of interest. AmiGO (Ashburner et al., 2000) (amigo.geneontology.org) was used to check for experimentally characterized proteins.

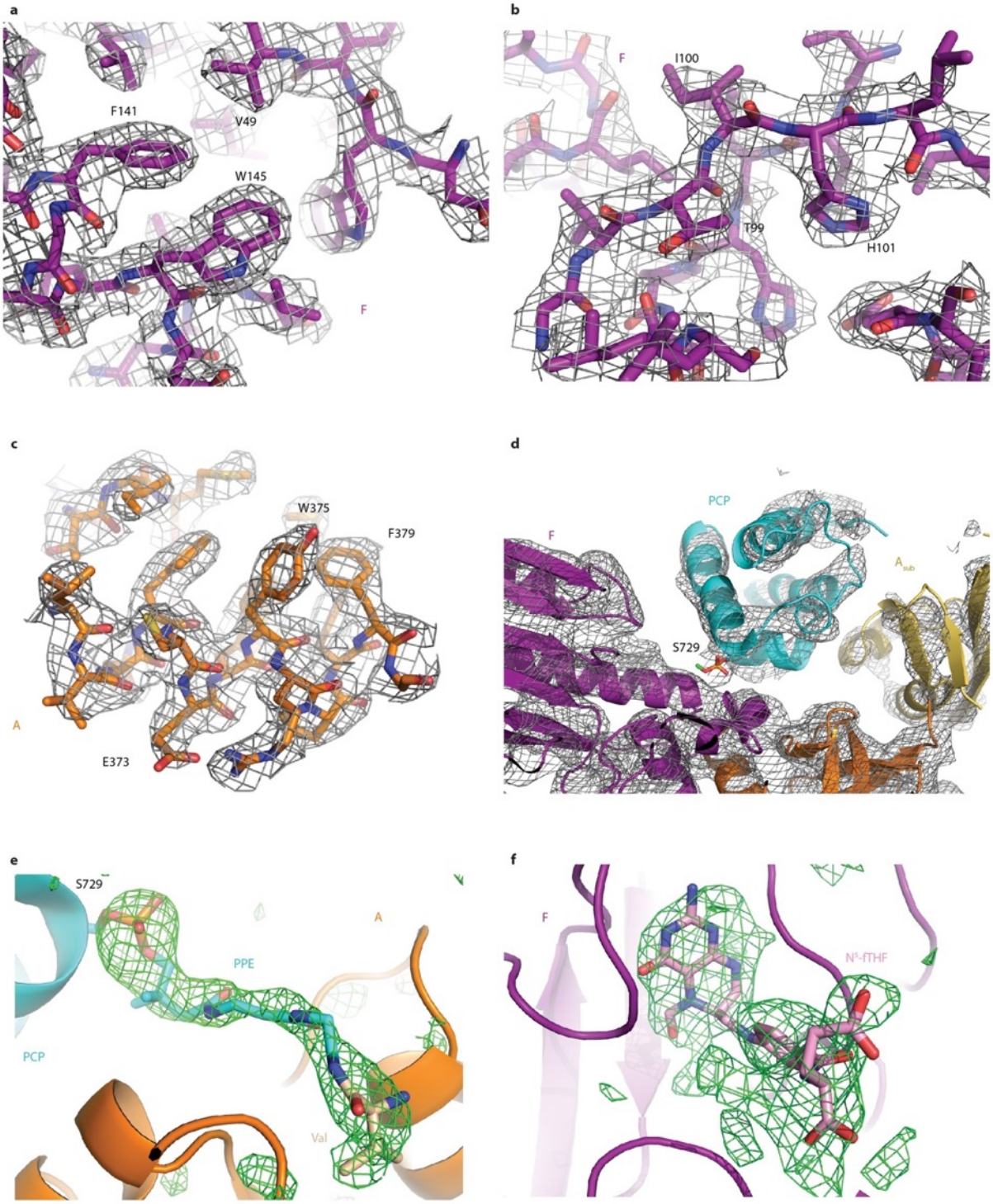
2.5.8 Analysis of synthesized Val–NH-CoA

Val–NH-CoA was verified by both mass spectrometry (calculated *m/z* [MH+]: 850.2304; measured *m/z* [MH+]: 850.2299) and 1H NMR [1H NMR (600 MHz, H2O) δ 8.69 (d, *J* = 16.2 Hz, 1H), 8.46 (s, 1H), 6.25 (d, *J* = 5.9 Hz, 1H), 4.97–4.89 (m, 1H) 4.75–4.73 (s, 1H), 4.65–4.58 (m, 1H), 4.30–4.21 (m, 2H), 4.04 (s, 1H), 3.91–3.82 (d, *J* = 5.7 Hz, 1H), 3.75 (dd, *J* = 8.8 Hz, 1H), 3.68–3.61 (d, *J* = 5.9 Hz, 1H), 3.55–3.42 (m, 3H), 3.42–3.23 (m, 4H), 2.48 (m, 2H), 2.21–2.14 (m, 1H), 1.03–0.99 (m, 7H), 0.87 (s, 3H), 0.44 (s, 3H)].

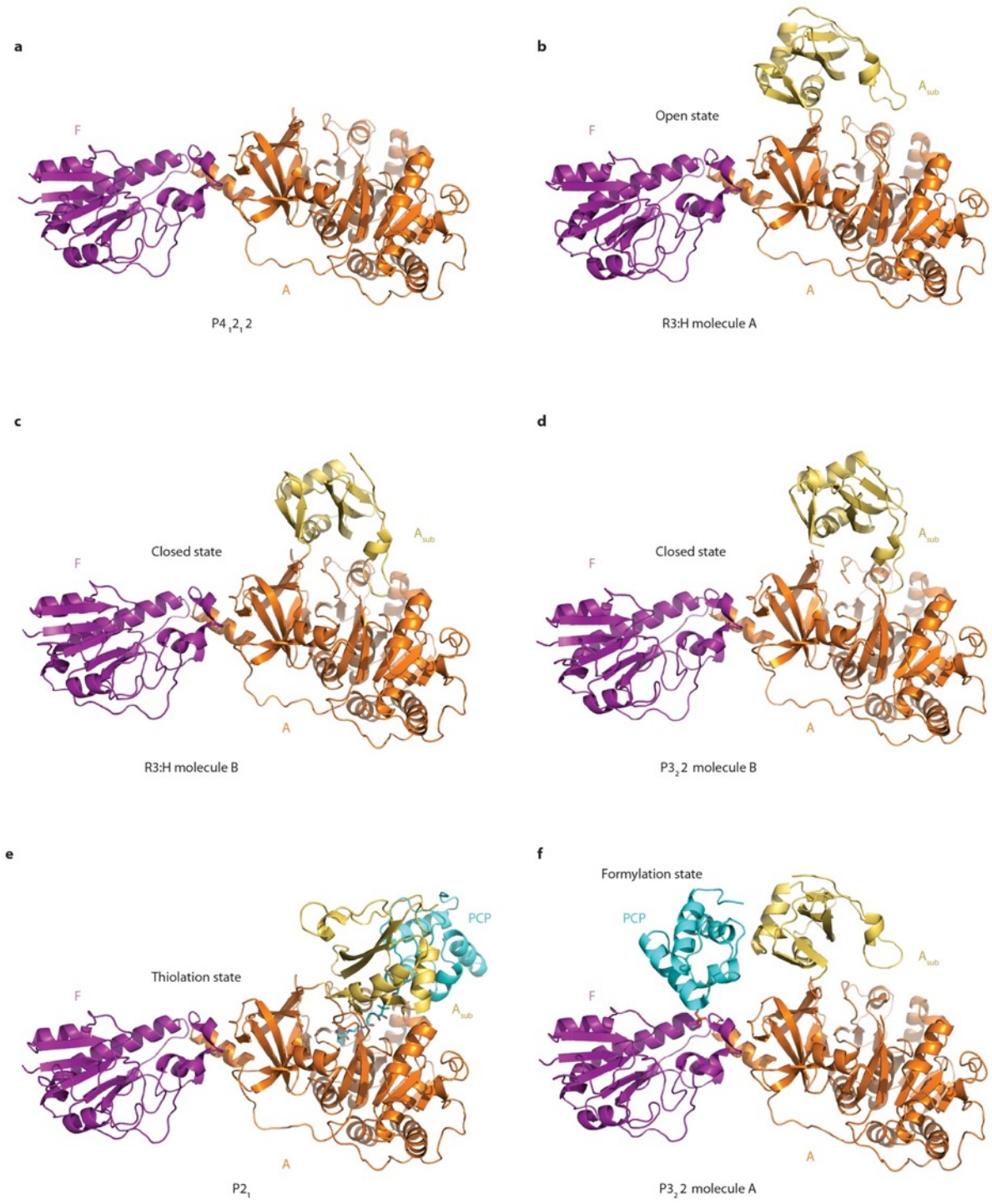
2.6 Supplementary Information



Extended Data Figure 2.1 | Synthetic cycles in canonical initiation, canonical elongation and LgrA initiation modules.
Schematic diagrams comparing the synthetic cycle in canonical initiation and elongation modules with that in the LgrA initiation module.

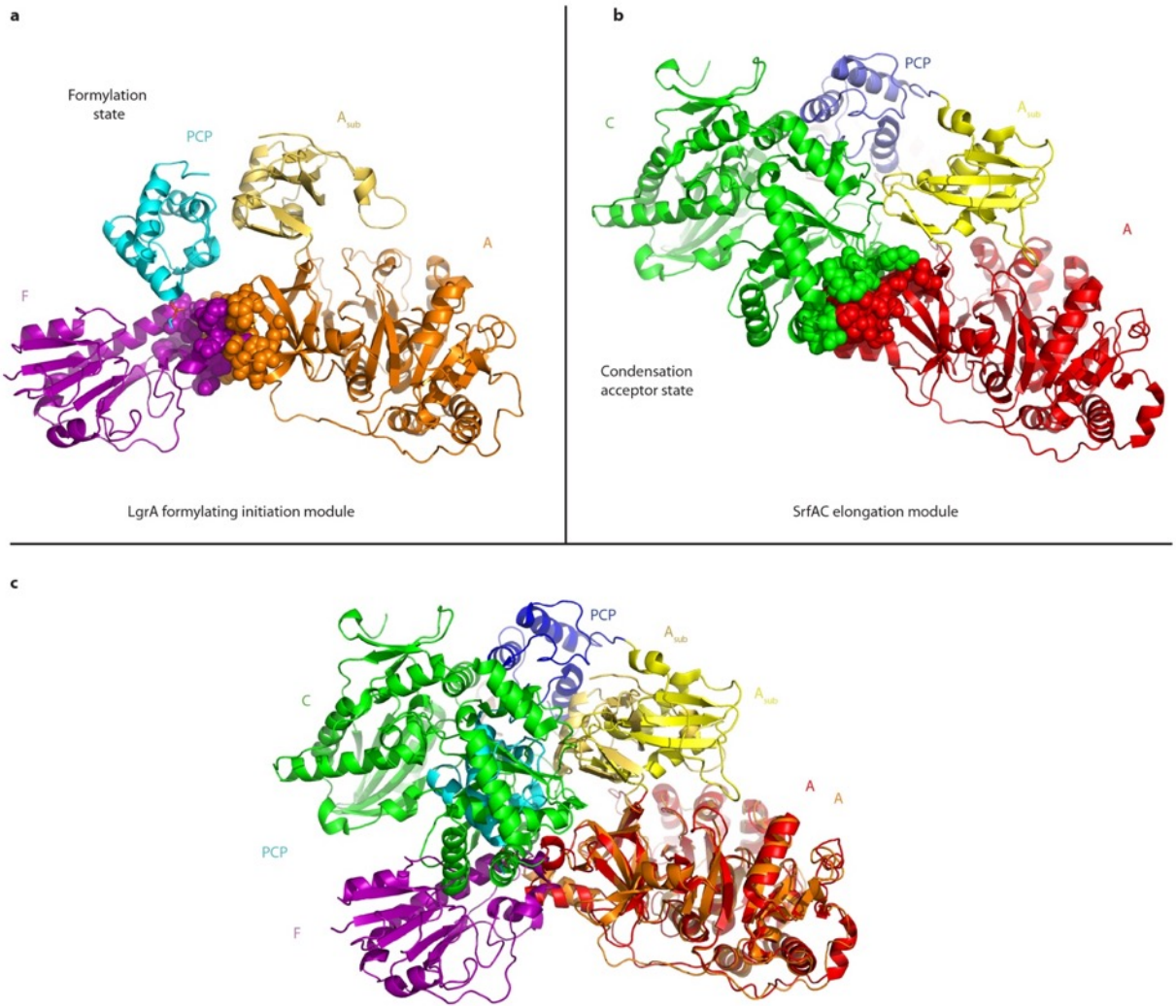


Extended Data Figure 2.2 | Representative electron density.
a–**d**, 2*F*o − *F*c density maps for protein in *P*41212 (**a**), *R*3:H (**b**), *P*21 (**c**) and *P*322 (**d**) crystal forms contoured at 1σ. **e**, **f**, Unbiased *F*o–*F*c density maps for the PPE–NH–Val arm in the *P*21 (thiolation state) contoured at 3.3σ (**e**), and a *P*41212 crystal soaked with N5–f THF, AMPcPP and valine contoured at 2.5σ (**f**).



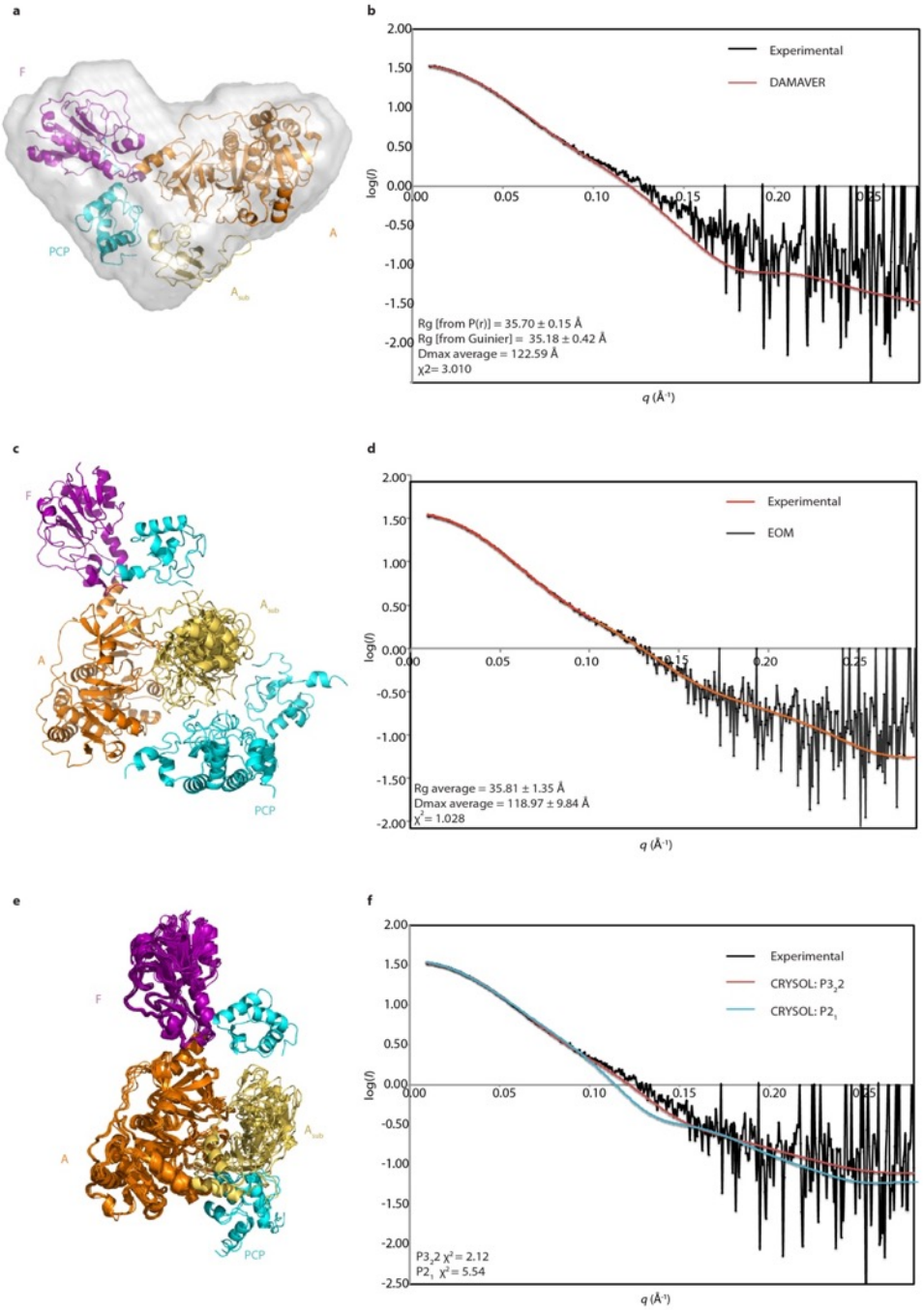
Extended Data Figure 2.3 | Crystal structures of the initiation module of linear gramicidin synthetase.

a–f, Models of F–A (Asub disordered) (**a**), F–A–PCP (PCP disordered) (**b–d**) and F–A–PCP from the four independent crystals structures determined (**e, f**). The crystal with space group *P*322 diffracted anisotropically to ~3.8 Å resolution, but the other higher resolution structures enabled the building of high quality models shown in **d** and **f**.



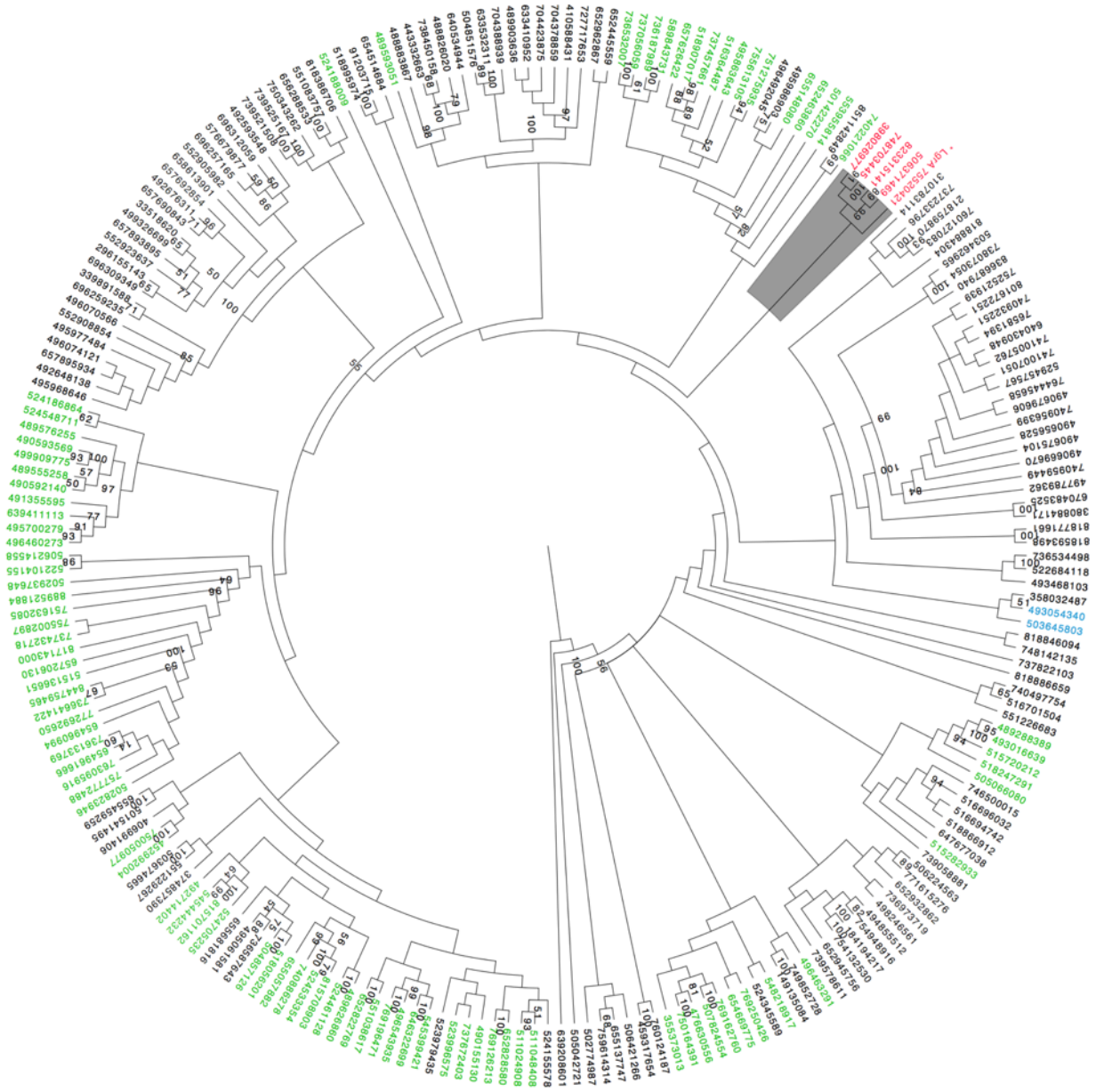
Extended Data Figure 2.4 | Comparison between the LgrA initiation module and the SrfA-C termination module.

a, b, The LgrA initiation module in the formylation state (**a**) and the termination module of surfactin synthase subunit 3 (SrfA-C) (Tanovic et al., 2008) (**b**) in the state where aminoacyl-PCP would be positioned to act as an acceptor substrate in the condensation reaction (PPE arm not present). The F and C domains are each positioned directly N-terminal of their A domains and bury similar amounts of A domain surface area (829 $Å^2$ and 903 $Å^2$; contributing residues shown in spheres), each forming 'stable platforms' (Tanovic et al., 2008). Both modules use very large movements of their PCP and A_{sub} domains to bring the aminoacyl-PCP of the module to distant active sites to act as the acceptor substrate in an amide bond forming reaction. **c**, However, the F–A and C–A interfaces are distinct, and, if the A domains are superimposed, the F and C domains are only partially overlapping. This places their active sites in dissimilar locations, necessitating that A_{sub} and PCP assume different positions to deliver their substrate. The PCP domain in the formylation state completely overlaps with the position of the C domain.



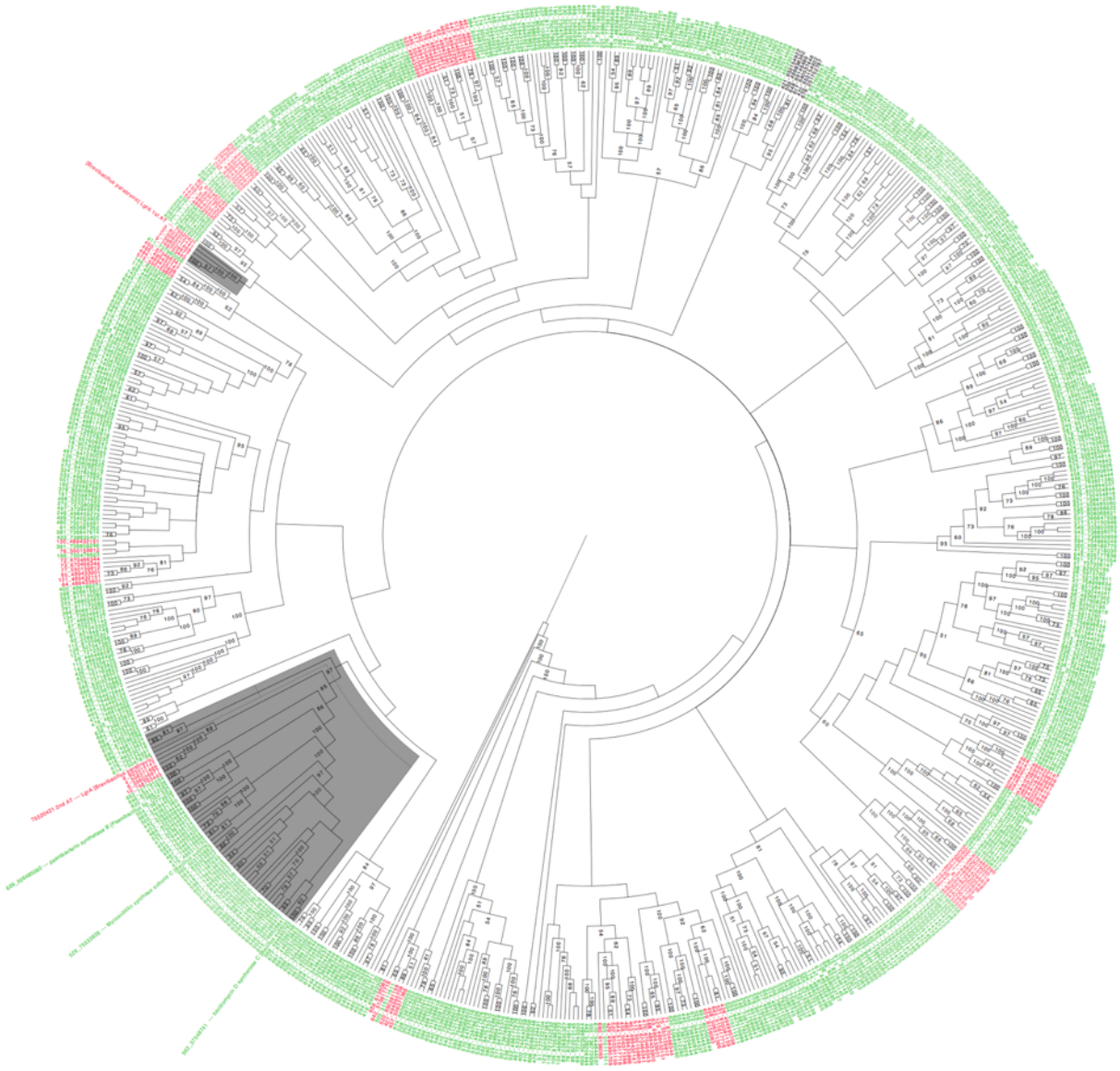
Extended Data Figure 2.5 | Small-angle X-ray scattering analysis of F-A-PCP.
Caption on following page.

Extended Data Figure 2.5 | Small-angle X-ray scattering analysis of F–A–PCP.
a, The crystal structure in the formylation state is shown superposed on the averaged filtered *ab initio* small-angle X-ray scattering model generated with DAMAVER (Volkov and Svergun, 2003), with a NSD value of 0.819 ± 0.052. **b**, The calculated scattering curve for the DAMAVER is overlaid with the experimental scattering with χ^2 = 3.010, where *I* represents scattering intensity and *q* is equivalent to $4\pi\sin(\theta)/\lambda$. **c**, To understand the flexibility of F–A–PCP better, EOM (Tria et al., 2015) was performed and generated five different ensembles. The ensemble resembling the formylation state structure represented over 60% of the optimized models generated, while the remaining <40% resembled the thiolation state structure. **d**, The calculated scattering of the EOM model has a χ^2 = 1.028, which demonstrates that F–A–PCP has flexibility. The data are consistent with extreme flexibility for A_{sub} and PCP domains, and limited flexibility in F–A_{core}. **e**, All independent molecules from the crystal structures were overlaid to further illustrate the flexibility of the system. **f**, CRYSOL (Svergun et al., 1995) was used to generate predicted scattering curves for the formylation state and thiolation state crystal structures with χ^2 = 2.12 and χ^2 = 5.54, respectively.

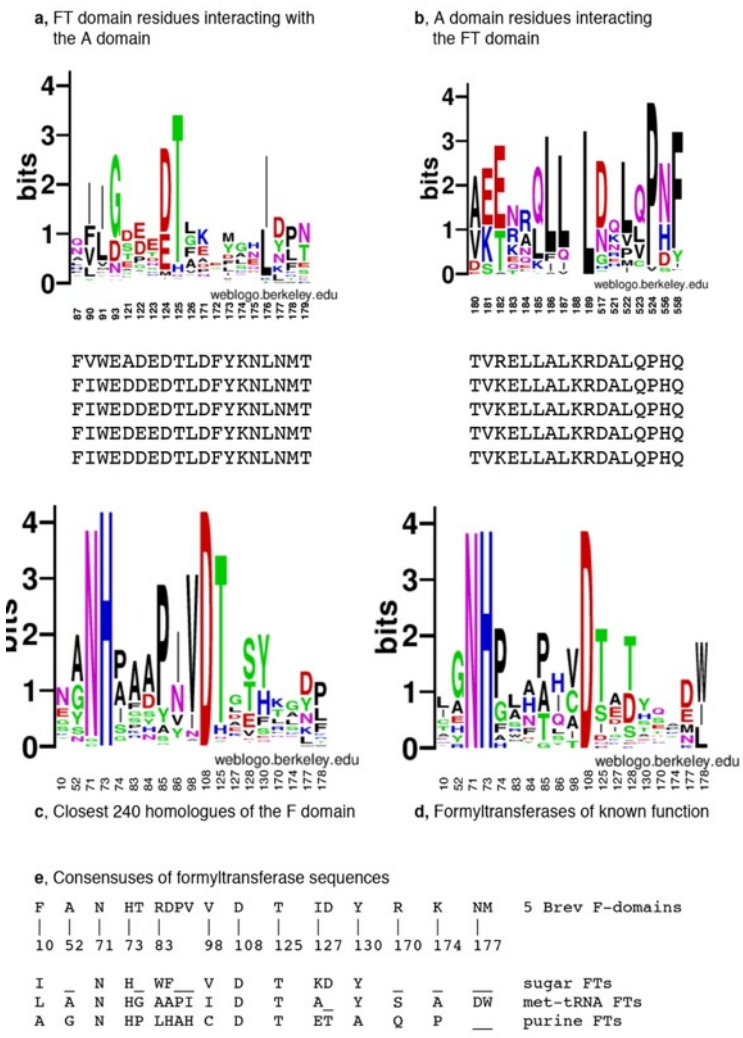


Extended Data Figure 2.6 | Neighbour-joining tree of LgrA F domain and homologues.
Caption on following page.

Extended Data Figure 2.6 | Neighbour-joining tree of LgrA F domain and homologues.
This neighbour-joining tree of the LgrA F domain and homologues was made using PHYLIP (http://evolution.genetics. washington.edu) based on an initial Clustal Omega (Sievers et al., 2011) alignment of the closest 220 homologues of the LgrA F domain (Blast (Altschul et al., 1990) BLAST E-value <1 × 1014). The most similar formyltransferases to the F domain share ~45% identity, and all of these 220 formyltransferases have only inferred function. The tree was drawn using the program FigTree (http://tree.bio. ed.ac.uk). The sequences are named with their GenInfo Identifier (GI) numbers. Colouring: red, Brevibacilli; green, other Firmicutes; black, other bacteria; blue, Archaea. The clade of the LgrA F domain is highlighted in grey. Only nodes with bootstraps of >50% are shown. Several horizontal transfer events are evident where Firmicute and non-Firmicute proteins cluster together with high bootstrap values (for example, >70%). The several horizontal transfer events of formyltransferase domains between Firmicutes and other bacterial groups suggest the LgrA F domain likely originated from horizontal transfer.

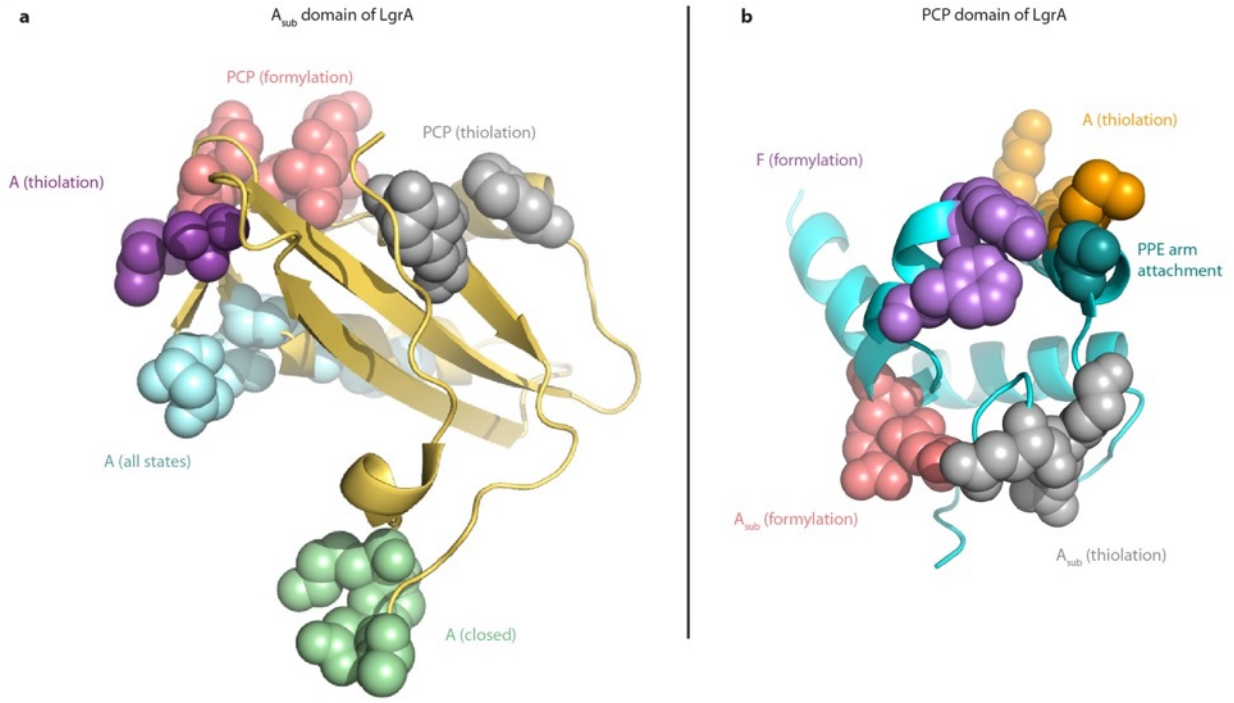


Extended Data Figure 2.7 | Neighbour-joining tree of LgrA A–PCP and homologues.
This neighbour-joining tree of LgrA A–PCP didomains and homologues was made for the 500 closest homologues (BLAST E-value <1 × 1014). The sequences are named with their GI accession codes. Colouring: red, Brevibacilli; green, other Firmicutes; black, other bacteria. The significant clades of the LgrA A–PCP domains are highlighted in grey. Only nodes with bootstraps of >50% are shown. Three functionally characterized homologues of LgrA that are shown to be directly related are labelled. The A–PCP portion of the initiation module is quite divergent, but the second module of LgrA clearly shares a common origin with functionally characterized NRPSs in Bacilli and other Firmicutes.



Extended Data Figure 2.8 | Conservation and variation of residues involved in the interaction interfaces.

a, b, Sequence logos made using the WebLogo server (http://weblogo.berkeley.edu) (Crooks et al., 2004) show conservation and variation as found in multiple sequence alignments of F domain residues that interact with the A domain (**a**) and A domain residues that interact with the F domain (**b**). Below each logo are the corresponding residues in the LgrA proteins from the five Brevibacillus species, with the crystallized LgrA on the first line. FT, formyltransferase. **c, d**, Sequence logos indicate the conservation and variation in F domain residues involved in binding and interaction with PCP–PPE–Val across the closest 240 homologues of LgrA (**c**) and all of the functionally or structurally characterized formyltransferase proteins (**d**) (reduced for redundancy so that no two sequences have >50% sequence identity). **e**, Consensus sequences for the five Brevibacillus LgrA homologues and for the formyltransferases of known structure for each of three formyltransferase types. Catalytic residues are His73, Asn71 and Asp108.



Extended Data Figure 2.9 | Interaction surfaces in PCP and A_{sub} domains.
The A_{sub} (**a**) and PCP (Weissman, 2015, Lohman et al., 2014, Allen and Gulick, 2014) (**b**) domains must maximize the use of their limited surfaces to interact with their many binding partners. Shown are the surfaces observed in this study, and many excellent previous studies have also documented interaction surfaces biochemically or structurally. This includes, for example, the equivalent of PCP domain residues Met249, Phe264 and Ala268, which are required for interaction with the C domain in the acceptor site55 and form hydrophobic interactions with the C domain (Tanovic et al., 2008) in a very similar manner and using an overlapping surface, as the PCP domain does to interact with the F domain. Furthermore, partially overlapping surfaces in PCP domains have been proposed to interact with their (acyl-) PPE arm to protect thioester intermediates (Jaremko et al., 2015)or to promote binding to the appropriate partner domain (Goodrich et al., 2015). These interactions might occur during PCP domain transit, but they would have to be broken before productive binding to partner domains. Several of these PPE interactions are incompatible with the productive domain– domain interactions (Goodrich et al., 2015), and in catalytic configurations seen here and previously, the PPE arms extend into the partner domain and make little contact with the PCP domain.

Extended Data Table 2.1 | Crystallographic statistics.

	F-A	F-A-PCP	F-A-PCP-PPE-NH-Val	F-A-PCP-PPE	F-A soak
Data collection					
Space group	$P4_12_12$	R3:H	$P2_1$	$P3_22$	$P4_12_12$
Cell dimensions					
a, *b*, *c* (Å)	161.3, 161.3, 138.2	278.7, 278.7, 82.8	77.9, 101.2, 139.6	162.1, 162.1, 208.9	160.8, 160.8, 137.6
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 120.0	90.0, 91.1, 90.0	90.0, 90.0, 120.0	90.0, 90.0, 90.0
Resolution (Å)	87.97-2.46 (2.52-2.46)	80.44-2.80 (2.88-2.80)	81.97-2.60 (2.66-2.60)	83.73-3.80 (4.01-3.80)	87.66-2.80 (2.90-2.80)
R_{merge}	0.097 (1.521)	0.072 (1.13)	0.173 (1.64)	0.110 (2.13)	0.127 (1.88)
$I/\sigma I$	12.9 (1.4)	10.2 (1.5)	5.5 (1.4)	10.8 (1.3)	10.3 (0.9)
Completeness (%)	100 (100)	100 (100)	100 (99.9)	100 (100)	100 (100)
Redundancy	9.7 (8.4)	3.9 (3.9)	3.8 (3.8)	11.2 (11.3)	14.7 (14.9)
Refinement					
Resolution (Å)	63.95-2.46 (2.50-2.46)	46.05 -2.80 (2.84-2.80)	47.47-2.55 (2.58-2.55)	48.50-3.77 (3.88-3.77)	49.98-2.81 (2.87-2.81)
No. reflections	66519	59106	65089	32124	61365
R_{work}/ R_{free}	0.233/0.254	0.227/0.263	0.225/0.274	0.308/0.331	0.236/0.260
No. atoms					
Protein	4644	10887	12180	11504	9625
Ligand/ion	9	10	56	5	65
Water	66	9	19	0	78
B-factors					
Protein	73.22	135.30	83.33	261.02	101.14
Ligand/ion	71.23	169.13	82.85	282.78	111.65
Water	75.11	111.99	64.75	N/A	110.47
R.m.s deviations					
Bond lengths (Å)	0.002	0.003	0.007	0.005	0.004
Bond angles (°)	0.660	0.540	1.145	0.913	0.670

*Highest resolution shell is shown in parentheses.

†One crystal was used for each structure.

2.7 Segue to Chapter 3

The modifications provided by tailoring domains are primarily what chemically distinguish a nonribosomal peptide from a ribosomal peptide. The mechanism by which new tailoring domains are acquired into NRPSs has only been briefly studied by evolution analyses. Based on our bioinformatics of the LgrA F domain, I wanted to investigate the sugar formyltransferase origins of the F domain using a biochemical and structural approach. Three free-standing sugar formyltransferases with high sequence similarity to the LgrA F domain were initially identified, and together with Jessie Jiang, an undergraduate student in the lab, we were able to successfully crystallize one of the formyltransferase candidates, PseFT from *Anoxybacillus kamchatkensis*. Following structure determination, we collaborated with Chris Whitfield's group at the University of Guelph to biochemically characterize the role of PseFT in its biosynthetic pathway of a sugar-nucleotide. We also solved 3 additional structures of PseFT with different ligands to gain insight into the evolutionary changes a prototypical sugar FT had to undergo to become a functional formylating domain in an NRPS.

CHAPTER 3 | STRUCTURAL INSIGHT INTO A NOVEL FORMYLTRANSFERASE AND EVOLUTION TO A NONRIBOSOMAL PEPTIDE SYNTHETASE TAILORING DOMAIN.

Janice M. Reimer[1,a], Ingrid Harb[1,a], Olga G. Ovchinnikova[1,b], Jesse Jiang[a], Chris Whitfield[b,*], and T. Martin Schmeing[a]. Structural insight into a novel formyltransferase and evolution to a nonribosomal peptide synthetase tailoring domain. *Under review at Cell Chemical Biology.*

[a]Department of Biochemistry, McGill University, Montréal, QC, H3G 0B1, Canada

[b]Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, N1G 2W1, Canada

[1]Equal contribution

3.1 Summary

Nonribosomal peptide synthetases (NRPSs) increase the chemical diversity of their products by acquiring tailoring domains. Linear gramicidin synthetase starts with a tailoring formylation (F) domain, which probably originated from a sugar formyltransferase (FT) gene. Here we present studies on an *Anoxybacillus kamchatkensis* sugar FT representative of the pre-horizontal gene transfer FT. Gene cluster analysis reveals that this FT acts on a UDP-sugar in a novel pathway for synthesis of a 7-formamido derivative of CMP-pseudaminic acid. We recapitulate the pathway up to and including the formylation step *in vitro*. Our X-ray crystal structures of the FT alone and with ligands unveil contrasts with other structurally characterized sugar FTs and show close structural similarity with the F domain.

3.2 Introduction

Nonribosomal peptide synthetases (NRPSs) are large enzymes that synthesize secondary metabolites with wide-ranging chemical and pharmaceutical properties (Walsh, 2004, Weissman, 2015). NRPSs use a modular, thio-templated synthetic scheme whereby repeating sets of NRPS domains (modules) add one amino acid or other monomer building block to the growing peptide chain (Schwarzer et al., 2003). The core NRPS domains required to make up a simple nonribosomal peptide are the adenylation (A) domain, the peptidyl carrier protein (PCP) domain and the condensation (C) domain. The A domain selectively binds cognate amino acid from the cellular pool, activates the amino acid by adenylation, and then transfers it to the prosthetic pantetheine (PPE) moiety on the PCP domain. The PCP domain transports covalently-bound intermediates between active sites, and the C domain catalyzes peptide bond formation.

Chemical modification, or tailoring, of nonribosomal peptides increases the chemical diversity of the bioactive secondary metabolites produced (Walsh et al., 2001, Hur et al., 2012). The chemical moieties introduced through tailoring are often required for synthesis to proceed and are indispensable for the biological or chemical activity of the small molecule product. A nonribosomal peptide can be modified in three different ways: post-synthetically, by “maturation proteins” which act on the peptide after release from the NRPS; co-synthetically *in trans,* by “accessory proteins” that act on synthetic intermediates covalently attached to PCP domain of the

NRPS; or co-synthetically *in cis,* by tailoring domains that are integrated into the NRPS polypeptide. NRPSs commonly contain *in cis* tailoring domains, including methyltransferase, epimerization, ketoreductase, oxidase and heterocyclization domains.

Tailoring domains, which are not evolutionarily related to core NRPS domains, arose though incorporation of foreign gene sequences into the NRPS coding region. The high prevalence of horizontal gene transfer in bacteria suggests that such fusion events are not rare (Lawrence and Roth, 1996). Advantageous fusions are likely to be maintained in the genome, while deleterious fusions are eliminated, and neutral fusions could be selected against by the relatively high bacteria genomic deletion rates (Cooper, 2014). A maintained post-fusion NRPS is thus likely to make a novel nonribosomal peptide that is beneficial to the host organism, either directly upon fusion or fairly shortly thereafter (Fischbach et al., 2008). By characterizing NRPSs containing tailoring domains and the stand-alone proteins which have homology to the tailoring domains, we can learn how nonribosomal peptide synthetases incorporate novel functionalities into their architecture.

We recently characterized the structure of the initiation module of linear gramicidin synthetase, which contains a tailoring formylation (F) domain as the first domain (Reimer et al., 2016b, Reimer et al., 2016a). The F domain catalyzes N-formylation of Val-PPE during an early step of linear gramicidin synthesis. This formylation was reported to be required for the steps that follow in the synthesis of linear gramicidin (Schoenafinger et al., 2006, Kessler et al., 2004), and indispensable for the bioactivity (Townsley et al., 2001). The N-formyl group mediates non-covalent head-to-head dimerization of two linear gramicidin peptides, which insert into the membrane of gram-positive bacteria, forming a pore for monovalent cations, disrupting the ion gradient and leading to bacterial death.

The F domain in linear gramicidin synthetase was originally recognized by sequence similarity to characterized formyltransferase (FT) proteins (Kessler et al., 2004). Three well characterized groups of proteins share a homologous FT catalytic core domain and use N^{10}-formyltetrahydrofolate (N^{10}-fTHF) as the formyl donor: glycinamide ribonucleotide transformylases, which catalyze a step of purine biosynthesis (Almassy et al., 1992); methionyl-tRNA formyltransferases, which formylate initiator Met-$tRNA^{fMet}$ for bacterial translation (Schmitt et al., 1998); and sugar formyltransferases, which formylate sugar-nucleotide substrates. We

propose that linear gramicidin synthetase arose from a horizontal gene transfer event which fused an NRPS with an FT which formylates sugar-nucleotide substrates (Reimer et al., 2016a).

The first structure of a sugar FT was of the N-terminal domain of the bifunctional enzyme ArnA. ArnA acts on the pentose substrate UDP-4-amino-4-deoxy-L-arabinose (UDP-L-Ara4N) in a pathway to make L-Ara4FN, a precursor for the sugar, L -Ara4N, frequently found in the lipid A moiety of bacterial lipopolysaccharides (Whitfield and Trent, 2014, Williams et al., 2005, Gatzeva-Topalova et al., 2005). Since then, the structures of seven other sugar FTs that catalyze N-formylation of aminodideoxyhexoses have been determined (Table 3.1). All display a mixed α/β topology including a Rossman fold and an Asn-His-Asp catalytic triad. The final 6-deoxyamino sugar products are all used in the glycosylation of the virulent *O*-antigen component of lipopolysaccharides (Holden et al., 2016), except for Rv3404c, which is suspected to be used in the glycosylation of the *M. tuberculosis* H37Rv cell wall (Dunsirn et al., 2017). These FTs are typically part of multi-domain proteins, where the conserved N-terminal domain contains the formyltransferase active site and the C-terminal domain (CTD) has varied functions, including structural support, protein dimerization, substrate binding and a completely different catalytic activity.

Table 3.1 | Sugar formyltransferases with determined structures and their products.

Formyltransferase name	Formyltransferase product
ArnA (Breazeale et al., 2005)	UDP-4-deoxy-4-formamido-L-arabinose (UDP-L-Ara4NFo)
WlaRD (Thoden et al., 2013)	dTDP-3,6-dideoxy-3-formamido-D-glucose (dTDP-D-Qui3NFo)
WbtJ (Zimmer et al., 2014)	dTDP-4,6-dideoxy-4-formamido-D-glucose (dTDP-D-Qui4NFo)
WbkC (Riegert et al., 2017)	GDP-4,6-dideoxy-4-formamido-D-mannose (GDP-D-Rha4NFo)
QdtF (Woodford et al., 2015)	dTDP-3,6-dideoxy-3-formamido-D-glucose (dTDP-D-Qui3NFo)
FdtF (Woodford et al., 2017)	dTDP-3,6-dideoxy-3-formamido-D-galactose (dTDP-D-Fuc3NFo)
VioF (Genthe et al., 2015)	dTDP-4,6-dideoxy-4-formamido-D-glucose (dTDP-D-Qui4NFo)
Rv3404c (Dunsirn et al., 2017)	dTDP-4-formamido-4,6-dideoxyglucose (dTDP-Qui4NFo)

Here, we identify an FT from *Anoxybacillus kamchatkensis* likely to be similar to the pre-transfer ancestor of the F domain. The FT gene is located within a gene cluster encoding all enzymes required for the biosynthesis of CMP-5-acetamido-7-formamido-3,5,7,9-tetradeoxy-L-

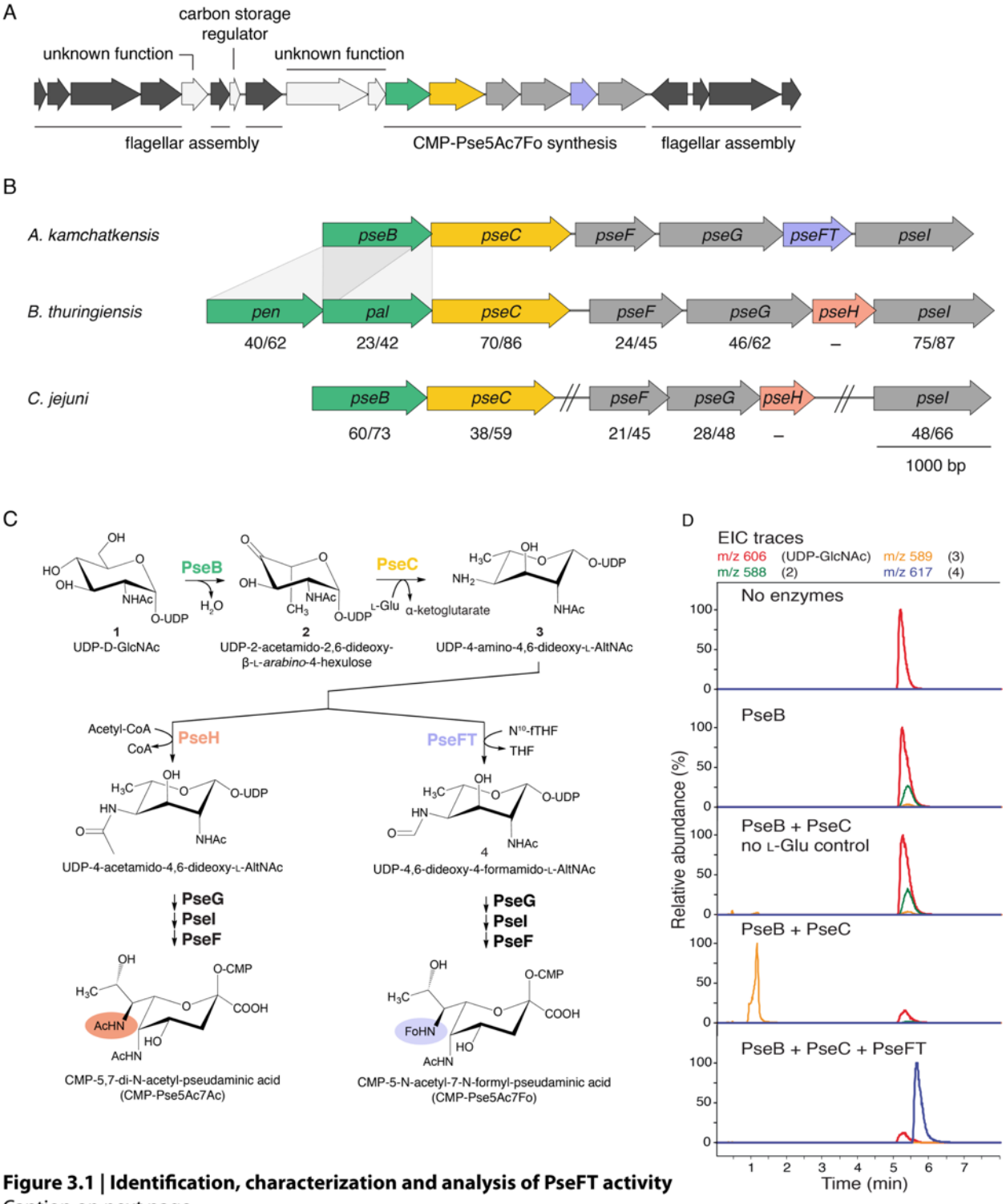


Figure 3.1 | Identification, characterization and analysis of PseFT activity
Caption on next page.

Figure 3.1 | Identification, characterization and analysis of PseFT activity.
A, Organization of CMP-Pse5Ac7Fo biosynthesis cluster in *A. kamchatkensis* (GenBank NZ_ALJT01000017.1) with neighboring genes, and (B) compared to CMP-Pse5Ac7Ac synthesis in *C. jejuni* and *B. thuringiensis*. The percentage of amino acid identity/similarity for the corresponding homologues proteins is indicated. *A. kamchatkensis* and *C. jejuni* pathways differ by a single N-acyltransferase enzyme (PseFT versus PseH). *B. thuringiensis* uses two homologous enzymes, Pen and Pal, to perform the UDP-GlcNAc 4,6-dehydration role of PseB. C, The first three steps of the proposed CMP-Pse5Ac7Fo biosynthesis pathway, leading to formation of product 4, were reconstituted *in vitro* with purified *A. kamchatkensis* PseB, PseC and PseFT. D, LC-ESI-MS analysis in negative ion mode of enzymatic reaction products. Shown are overlapped extracted ion chromatograms (EIC) of the reaction mixtures containing either no enzymes, or one, two or three enzymes, as well as a control reaction without amino donor. UDP-GlcNAc *m/z:* 606 (in red); UDP-2-acetamido-2,6-dideoxy-L-*arabino*-4-hexulose *m/z*: 588 (in green); UDP-4-amino-4,6-dideoxy-L-AltNAc *m/z:* 589 (in yellow); UDP-4,6-dideoxy-4-formamido-L-AltNAc *m/z*: 617 (in dark purple).

glycero-L-*manno*-non-2-ulosonic acid (CMP-5-N-acetyl-7-N-formylpseudaminic acid, CMP-Pse5Ac7Fo) (Figure 3.1A). The cluster itself is flanked by the genes involved in flagellar biosynthesis, which suggests that CMP-Pse5Ac7Fo is used for flagellin glycosylation. We cloned, purified and expressed the *A. kamchatkensis* FT (here called PseFT) and the two putative upstream enzymes (PseB and PseC) in the biosynthetic pathway, and confirmed the activities of all three proteins *in vitro*. To visualize the structural differences between PseFT and other reported sugar FTs, or the NRPS F domain, we determined the structure of the PseFT alone and in complex with cofactor, substrate or products. PseFT has interesting differences from other characterized sugar FTs, but is remarkably similar to the F domain in structure.

3.2 Results

3.2.1 Identification and characterization of PseFT

Sequence similarity suggested an evolutionary relationship between the linear gramicidin synthetase F domain and free-standing FTs. Therefore, we performed a search of the NCBI non-redundant protein sequence database using BLAST with the protein sequence of the linear gramicidin synthetase subunit A (LgrA) F domain as the input query. The search returned several linear gramicidin synthetase proteins from *Brevibacillus* species, and thousands of free-standing FTs from bacteria and archaea (Reimer et al., 2016a). The top FTs were >40% identical and >60%

similar to LgrA over the entire F domain. The genomic environments of several of these FT genes were investigated, including those of the ten proteins with highest identity and similarity to LgrA F domain. Seven of these ten FT genes are found clustered with at least some genes predicted to be involved in the synthesis of CMP-pseudaminic acid. The CMP-Pse5Ac7Ac biosynthetic pathway has been characterized in Gram-negative and Gram-positive bacteria and converts UDP-GlcNAc to CMP-Pse5Ac7Ac using six and seven enzymes, respectively (Schoenhofen et al., 2006, McNally et al., 2006, Li et al., 2015) (Figure 3.1B). In the *A. kamchatkensis* cluster (Figure 3.1B), the gene coding for PseFT (protein WP_019417499), which is highly similar to the F domain, presumably takes the place of the gene for the N-acetyltransferase PseH. PseH catalyzes acetylation of the C4 amino group of UDP-2-acetamido-4-amino-2,4,6-trideoxy-L-altrose (UDP-4-amino-4,6-dideoxy-L-AltNAc) to produce UDP-4-acetamido-4,6-dideoxy-L-AltNAc (Schoenhofen et al., 2006). We postulated that PseFT acts on the same substrate generating UDP-4,6-dideoxy-4-formamido-L-AltNAc (Figure 3.1C).

3.2.2 Analysis of *A. kamchatkensis* PseB activity and product

To analyze the pathway and evaluate the ability of PseFT to formylate UDP-4-amino-4,6-dideoxy-L-AltNAc, we expressed and purified PseB, PseC and PseFT. The activity of purified enzymes was assessed by examining *in vitro* reaction products by reverse-phase HPLC (Figure S3.1) and LC-MS analysis (Figure 3.1D). A control reaction mixture containing the initial substrate, UDP-GlcNAc **1,** incubated in the absence of enzymes eluted as a single peak at 10.5 minutes and gave an $[M–H]^-$ ion peak at *m/z* 606.07. Adding PseB (a predicted UDP-GlcNAc 5-inverting 4,6-dehydratase (Ishiyama et al., 2006)) to the reaction mixture containing **1** resulted in the formation of novel product(s), evident as a slower migrating broad asymmetric peak. PseB uses $NADP^+$ as a cofactor in a three-part oxidation-dehydration-reduction reaction that regenerates the oxidized form of the cofactor (Morrison et al., 2008). Addition of exogenous

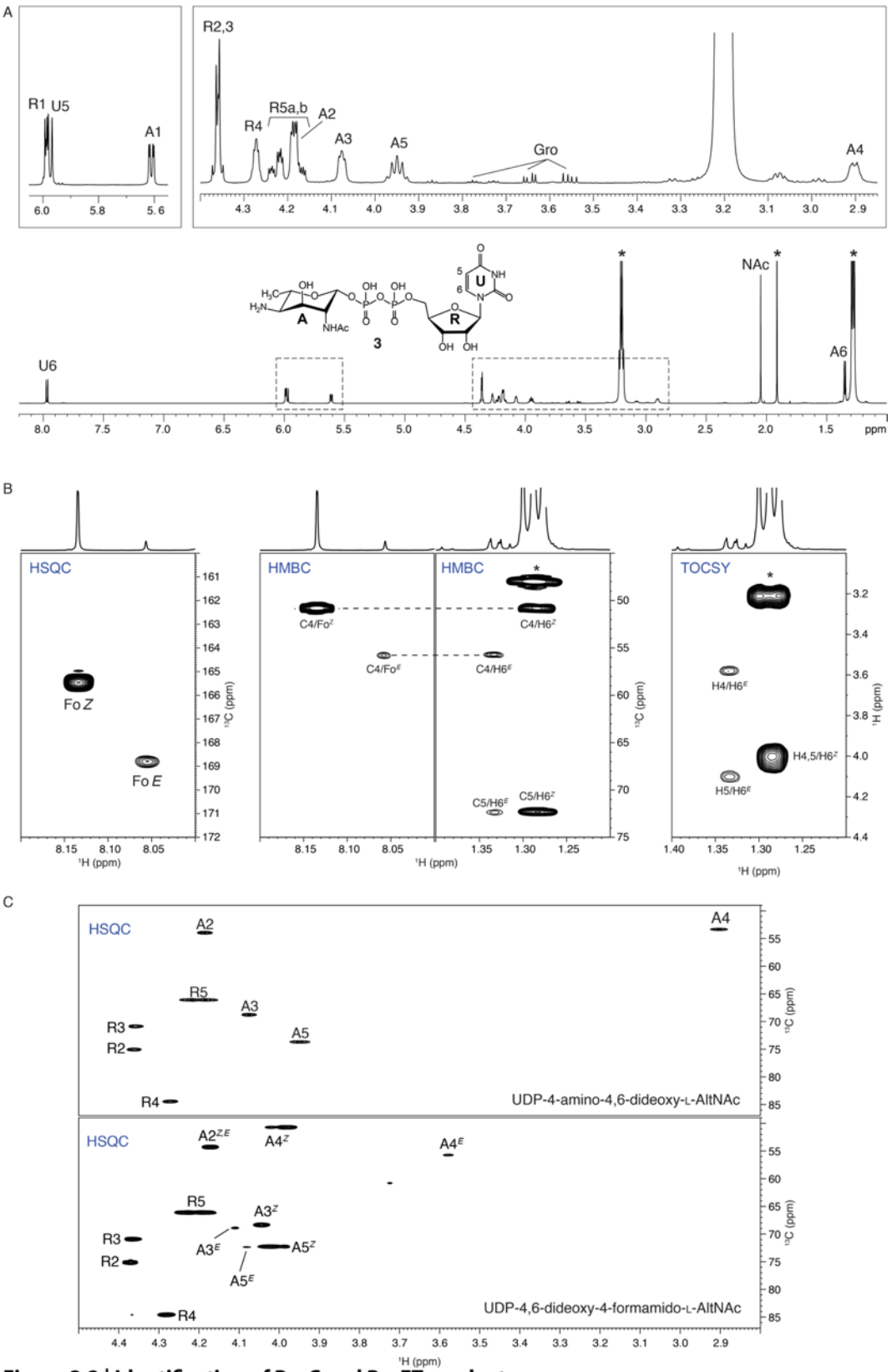


Figure 3.2 | Identification of PseC and PseFT products
Caption on next page.

Figure 3.2 | Identification of PseC and PseFT products.
A, The structure and ^{1}H NMR spectrum of the purified PseC product. The expanded regions of the spectrum showing the resonances of sugar ring protons are given in the inserts. The signals arising from residual triethylammonium acetate from the HPLC buffer are marked with asterisks. Glycerol (Gro) is a contaminant from the purified enzymes. **B**, NMR analysis of the purified PseFT product. Shown are selected parts of ^{1}H,^{13}C HSQC, HMBC and ^{1}H,^{1}H TOCSY spectra. The two series of signals for the 4,6-dideoxy-4-formamido-L-AltNAc moiety originate from the presence of *Z*- and *E*-stereoisomers, which result from restricted rotation around the C–N amide bond of the N-formyl group. Contaminating signals of triethylammonium acetate from the HPLC buffer are marked with asterisks. **C**, Parts of ^{1}H,^{13}C HSQC spectra of the PseC product (top) and the PseFT product (bottom). Arabic numerals refer to cross-peaks in sugar residues, labelled as "A" (altropyranose derivative) and "R" (ribose). N-formylation of the amino group at position 4 causes considerable downfield shift of the signal A H-4.

$NADP^{+}$ was not required for the reaction, suggesting that the enzyme was purified with the cofactor tightly bound. LC-MS of the PseB reaction mixture showed the presence of $[M–H]^{-}$ ion peak at *m/z* 588.06. The loss of 18.01 atomic mass units in **2** compared to **1** confirmed the dehydratase activity of PseB.

UDP-hexose-4-uloses are unstable metabolites that partially decompose upon lyophilization from buffered solutions, making purification of these compounds challenging. To identify PseB products in reaction mixture, the enzymatic reaction was performed in an NMR tube in deuterated buffer and monitored in real time by ^{1}H NMR spectroscopy. PseB homologs from *C. jejuni* and *H. pylori* catalyze formation of UDP-2-acetamido-2,6-dideoxy-β-L-*arabino*-4-hexulose (**2**) and UDP-2-acetamido-2,6-dideoxy-α-D-*xylo*-4-hexulose (**5**) in a sequential manner, and both compounds exist in equilibrium with their hydrated forms (**2´** and **5´**; Figure S3.2A). The NMR chemical shifts of **2**, **2´**, **5**, and **5´** we observe (Supplementary Table 3.1) are in good agreement with those reported previously (Schoenhofen et al., 2006, McNally David et al., 2006). Eleven minutes after addition of PseB, a decrease in signal intensity of **1** was observable and the signals for **2**, **2´** and **5´** appeared simultaneously in the ratio 1:3:0.05 (Figure S3.2B). At the 27 hour time point, 89% of UDP-GlcNAc substrate was consumed, and the ratio of **2**, **2´** and **5´** was 1:3.6:0.4. Compound **5** was not identified due to low signal intensity. Selective 1D TOCSY and NOESY experiments, 2D ^{1}H,^{13}C HSQC and HMBC experiments were used to confirm the structure of **2, 2´** and **5´** (Figures S3.2B, S3.3). The selective TOCSY spectra of anomeric protons H-1 allowed assignment of H-2 and H-3 signals. The HMBC spectrum showed correlations H-6/C-5 and H-6/C-

4, which distinguished unhydrated keto-sugar **2** (C-4 at δ 210.5) and *gem*-diol forms **2′** and **5′** (C-4 at δ 94.5). The β-l-configuration of **2** and **2′** was determined based on $J_{H,P}$ 8.0-8.4 Hz and strong NOE between H-6 and H-3, which lie on the same face of sugar ring. No **5′** H-6/H-3 NOE and $J_{H,P}$ 7.0 Hz were indicative of α-D-configuration of **5′**. These data confirmed that *A. kamchatkensis* PseB is a UDP-GlcNAc 4,6-dehydratase 5-epimerase.

3.2.3 Analysis of PseC activity and product

A coupled assay containing UDP-GlcNAc, PseB and PseC, together with an amino group donor (L-glutamate) and the PseC cofactor (pyridoxal 5-phosphate; PLP), generated a new product, peak **3**. LC-MS of the reaction mixture revealed an $[M–H]^-$ ion peak at *m/z* 589.09, consistent with the anticipated conversion of the keto-group to an amino group. In a control reaction lacking L-glutamate, no conversion of **2** to **3** was observed. However, adding PLP was not necessary, suggesting that PseC was purified with endogenous cofactor.

Based on ^{1}H, ^{13}C and ^{31}P NMR data, the PseC product was identified as UDP-4-amino-4,6-dideoxy-β-L-Alt*p*NAc. The signals within the sugar spin system were unambiguously assigned by 2D COSY and TOCSY spectra (Table S3.1, Figure 3.2A, 3.2C). The COSY spectrum facilitated the tracing of connectivities between all neighboring protons, starting from anomeric proton H-1 at δ 5.61. The resonances of the carbon signals were assigned based on HSQC and HMBC experiments (Figure 3.2C). The position of the C-6/H-6 signal at δ_C/δ_H 19.3/1.35 is characteristic of the methyl group of a 6-deoxysugar. The signals for C-2 and C-4 at δ 53.9 and 53.3, respectively, indicate that these carbons are linked to nitrogen atoms. The signals at δ_C 23.2 and 175.8 demonstrated the presence of one N-acetyl (NAc) group; thus one of the two amino groups in the sugar moiety is N-acetylated. The location of the NAc group at position 2 was established by an HMBC experiment, which demonstrated correlation between the NAc CO and H-2 at δ 175.8/4.18. Based on ^{1}H and ^{13}C chemical shifts and the proton-coupling constants (Table S1), as well as comparison with NMR data reported for UDP-4-amino-4,6-dideoxy-L-AltNAc (Schoenhofen et al., 2006), the sugar residue has an *altro* configuration. A one-bond $^1J_{C\text{-}1,H\text{-}1}$ coupling constant of 170 Hz indicates a β-anomeric configuration for L-sugars (Duus et al., 2000), while values for $^3J_{H\text{-}1,H\text{-}2}$ of 2.3 Hz and $^3J_{H\text{-}1,P}$ of 8.4 Hz are consistent with a β-L-configuration. If PseC is a monofunctional aminotransferase as expected,

the inversion of the absolute configuration of the initial substrate (UDP-α-D-GlcNAc) has to occur prior to PseC catalysis. This is consistent with the predicted C-5 inverting activity of PseB.

3.2.4 Analysis of PseFT activity and product

Finally, in a one-pot reaction containing PseB, PseC, PseFT and a formyl donor (N^{10}-fTHF), UDP-GlcNAc was almost completely converted into a new product **4**. The observed $[M–H]^-$ ion peak at *m/z* 617.09 indicated a gain of 28 atomic mass units, which is expected for addition of a formyl group. To fully characterize products **3** and **4**, they were purified from scaled-up reactions by semi-preparative reverse-phase HPLC and analyzed in detailed by NMR spectroscopy (described below) and mass spectrometry. MS/MS analyses of both products **3** and **4** showed characteristic fragment ions corresponding to $[UDP-H]^-$ at *m/z* 402.99, $[UDP-H_2O]^-$ at *m/z* 384.98 and $[UMP-H]^-$ at *m/z* 323.03, thus confirming the identity of purified compounds as UDP-sugars.

NMR analysis of the purified PseFT product confirmed its identity as UDP-4,6-dideoxy-4-formamido-β-L-Alt*p*NAc. The HSQC spectrum (Figure 3.2B) revealed characteristic signals for the N-formyl group at δ_H/δ_C 8.13/165.4 (*Z*-isomer) and 8.06/168.8 (*E*-isomer), with *Z*–*E* ratio of 7.4: 1 according to the ^{1}H NMR data. For H/C-4, correlations in the HMBC spectrum enabled identification of C-4 for both *Z* and *E*-isomers, by assigning the protons of methyl group H-6^Z and H-6^E using H-6/C-4 correlations. The remaining spin system was assigned using COSY and TOCSY experiments, and ^{13}C NMR chemical shifts were determined by HSQC (Table S3.1, Figure 3.2C). N-acetylation at position 2 was confirmed by HMBC data. As expected, the largest difference between chemical shifts of *Z* and *E*-isomers ($\Delta\delta_H$ 0.40 ppm, $\Delta\delta_C$ 5.0 ppm) was observed for the C/H pair at the position of attachment of the formyl group (position 4). A similarly large effect on ^{13}C NMR signals of an N-formylated sugar was reported previously for 3,6-dideoxy-4-formamido-D-glucose (Qui3NFo) (Kondakova et al., 2012), although in other cases the difference between ^{13}C NMR data of the two stereoisomers can be less pronounced (Katzenellenbogen et al., 1995).

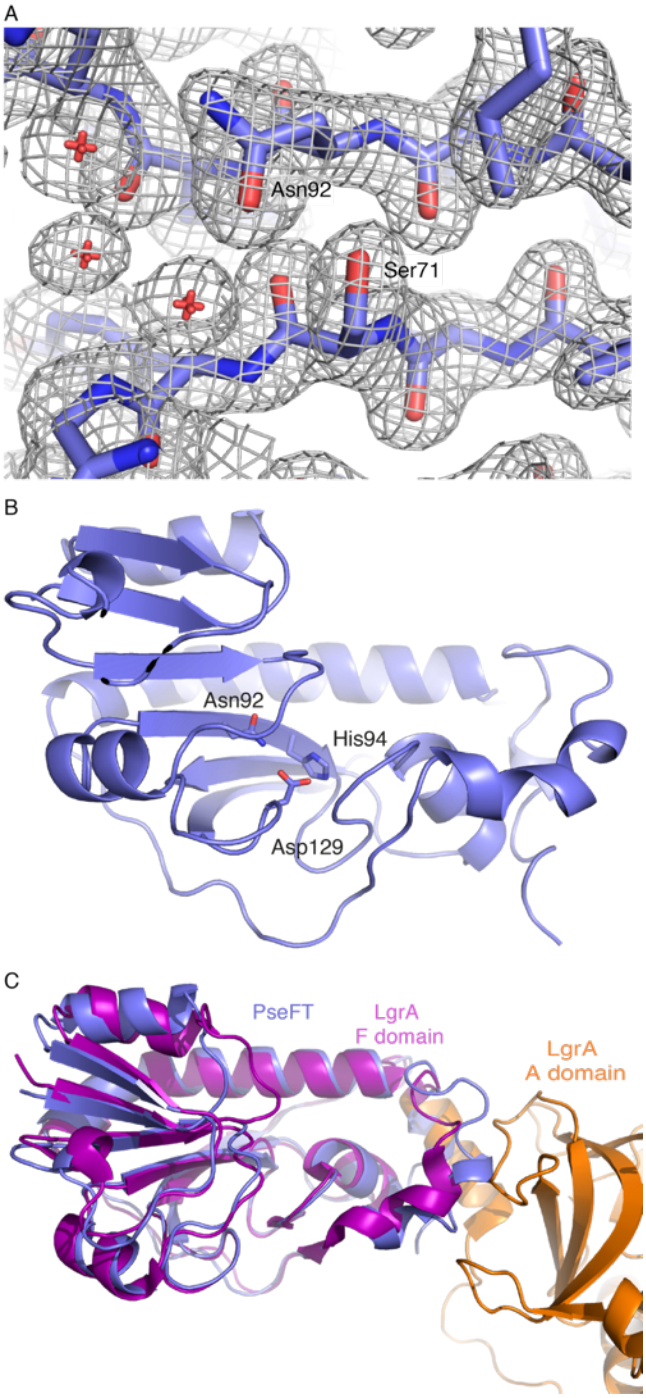


Figure 3.3 | Structure of PseFT.
A, Composite omit map of PseFT contoured to 2 σ. **B**, Structure of PseFT with the catalytic triad indicated. **C**, The LgrA F domain (Reimer et al., 2016a) is structurally homologous to PseFT.

3.2.5 Structures of PseFT in absence of ligands and bound to cofactor, substrate or products

To better understand the structural differences between the PseFT with other similar FT proteins or the F domain, we determined the X-ray structure of PseFT to 1.96 Å (Figure 3.3, Table S3.2, Figure S3.5). PseFT contains 193 amino acids and has many hallmarks of sugar FTs (Figures S3.6). The structure contains a Rossman fold common to N^{10}-fTHF binding enzymes, with the catalytic triad of Asn92, His94 and Asp129 at the center of the active site. The HxSLLPKxxG motif characteristic of N^{10}-fTHF utilizing enzymes (Gatzeva-Topalova et al., 2005) is altered to HxSYLPWNKG with Pro99 adopting the typical *cis* conformation within the motif. PseFT exhibits high structural similarity to the LgrA F domain (Figure 3.3C), with a backbone RMSD of 1.3 Å over 171 residues. Structural differences between F domain and PseFT are subtle: PseFT helix α1 is rotated outwards slightly, and the loop connecting helix α1 with strand β1 is 3 residues shorter. The last 20 residues of PseFT have a higher deviation from the F domain, but have the same fold.

We have also determined three ligand-bound structures of PseFT by soaking experiments. Structures with cofactor N^{10}-fTHF, with substrate UDP-4-amino-4,6-dideoxy-L-AltNAc, and with product UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF, were determined to 1.8 Å, 1.8 Å and 2.0 Å resolution, respectively (Figure 3.4, Table S3.2, Figure S3.4). In the co-complex structure with

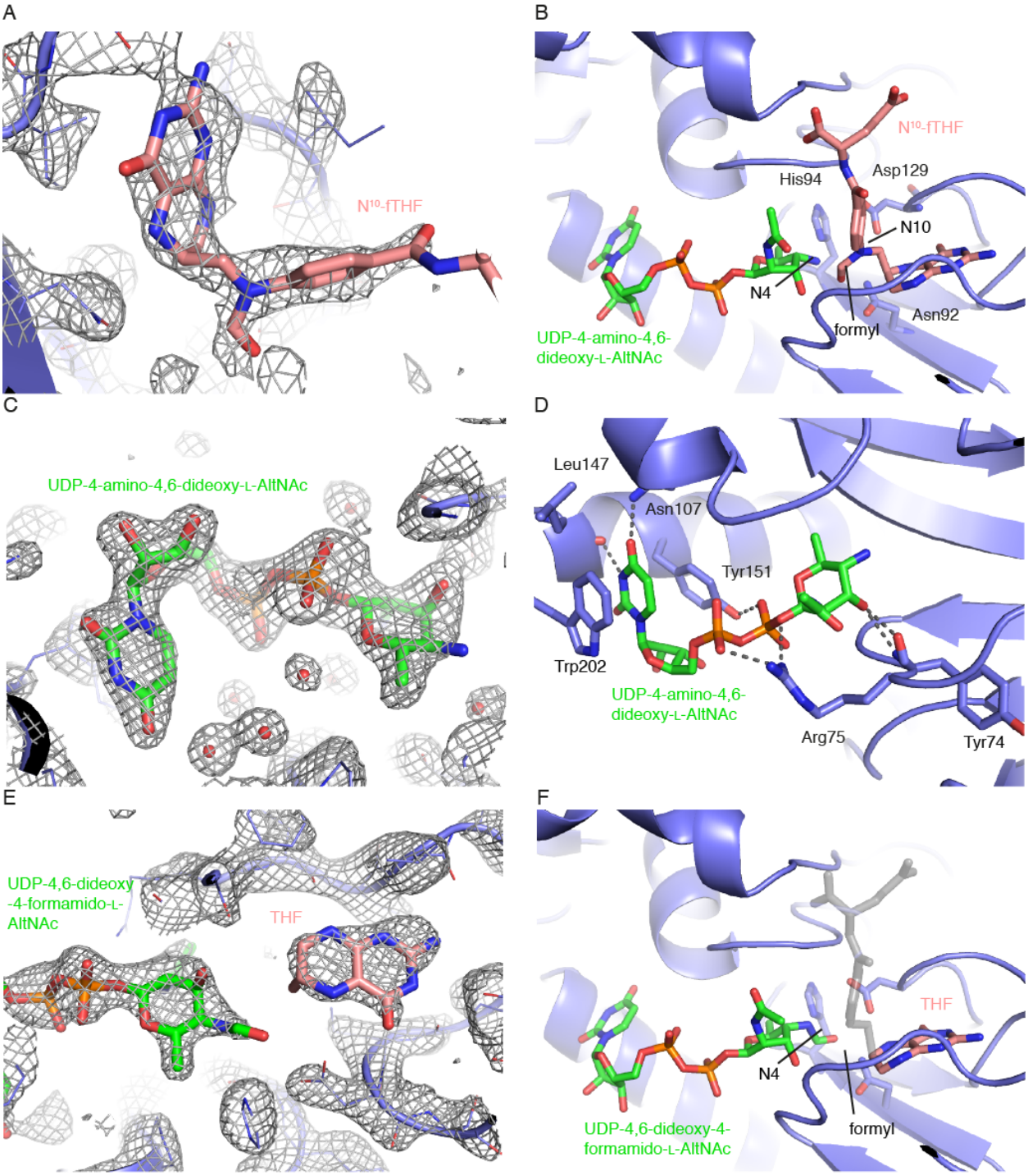


Figure 3.4, caption on next page.

Figure 3.4 | Structure of PseFT bound to substrate, cofactor and products.
A, Simulated annealing omit map of PseFT with its cofactor, N^{10}-formyltetrahydrofolate (N^{10}-fTHF). **B,** The N4 of UDP-4-amino-4,6-dideoxy-L-AltNAc is positioned for transfer of the formyl group. **C,** Simulated annealing omit map of PseFT with its substrate, UDP-4-amino-4,6-dideoxy-l-AltNAc. **D,** PseFT substrate UDP-4-amino-4,6-dideoxy-L-AltNAc binds by hydrogen bonding and π stacking. **E,** Simulated annealing omit map of PseFT with its product, UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF. **F,** Little structural change occurs following formylation, as seen by this structure of UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF. The pterin ring of THF is visible in maps, but the rest of that molecule is disordered (as indicated by transparent grey depiction). See also Table S2, Figure S4. All simulated annealing omit maps are contoured to 1.5 σ.

N^{10}-fTHF, the stereoisomer (6S)N^{10}-fTHF is selectively bound from the racemic mixture (Figure 3.4A). The (6R)N^{10}-fTHF enantiomer is accepted as the natural compound, but all FT:(f)THF co-complex structures feature the same C6 chirality observed here ((6S)N^{10}-fTHF, (6R)N^{5}-fTHF, (6R)THF) (Williams et al., 2005, Woodford et al., 2015, Thoden et al., 2013, Genthe et al., 2015, Woodford et al., 2017, Riegert et al., 2017, Dunsirn et al., 2017, Almassy et al., 1992). N^{10}-fTHF binds in a pocket next to the catalytic triad in a manner similar to that seen before, positioning the formyl group in the center of the catalytic triad (Figure 3.4B, Figure 3.5A) (Thoden et al., 2013, Woodford et al., 2017). The substrate co-complex structure of PseFT shows UDP-4-amino-4,6-dideoxy-L-AltNAc bound to the sugar-nucleotide pocket (Figure 3.4D), but with substantially different binding interactions from those seen before (Figure 3.5C, D; see also Discussion section). In PseFT, the sugar-nucleotide's uracil ring stacks between Tyr151 and Trp202, and its Watson-Crick face makes hydrogen bonds with the sidechain amine of Asn107 and the backbone carbonyl oxygen of Leu147. Trp202 is rotated ~55° degrees from its position in apo PseFT to form this stack, and Tyr151 also uses its hydroxyl to hydrogen bond with the β phosphate (Figure 3.4D). The sidechain of Arg75 becomes ordered upon substrate binding and hydrogen bonds with both the α phosphate and the Alt 1-oxygen, while also stacking with the 2-acetamido group. There are also several water-mediated protein-substrate interactions, especially with the phosphates which are exposed in this rather open binding site. A hydrogen bond between the carbonyl of Tyr74 and the 3-hydroxyl likely helps fine positioning of the acceptor 4-amine. Superimposition of the cofactor and substrate-bound structures show the 4-amine 2.5 Å from the N^{10}-formyl group and 3.7 Å from the catalytic base, His94 (Figure 3.4B), in good position to start general base-catalyzed formylation.

Transfer of the formyl group from the cofactor onto the 4-amino group produces UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF. The presence of the formyl group at position 4 is visible in electron density maps (Figure 3.4E), and only the pterin ring of the THF is ordered, otherwise, there is almost no difference between substrate and product bound structures (Figure 3.4F).

3.3 Discussion

Horizontal gene transfer enables biosynthetic pathways to co-opt and integrate new enzymes and impart beneficial functionalities to their products. NRPSs have acquired N-formylation functionality through several independent fusion events. Interestingly, the type of FT source differs among NRPSs: Unlike LgrA, kolossin A synthetase (Bode et al., 2015), anabaenopeptilide synthetases (Rouhiainen et al., 2000) and some other NRPSs appear to have fused the gene for a methionyl-tRNA formyltransferase (FMT) into their initiation modules to give an altered modular architecture (FMT-FMT_{CTD}-A-PCP) (Reimer et al., 2016a). This parallel adoption of the formyltransferase tailoring into NRPSs hints at a general utility of the N-terminal formyl group in secondary metabolites. Notably, formylation is also needed for full activity of many of the ribosomally synthesized, "unmodified, leaderless" class of bacteriocins (Hanchi et al., 2016).

The six-step biosynthetic pathway of CMP-Pse5Ac7Ac has been extensively studied in Gram-negative pathogens *H. pylori* and *C. jejuni* (reviewed in (Salah Ud-Din and Roujeinikova, 2017)). In both species, Pse5Ac7Ac is used to decorate flagella via O-glycosylation, which is essential for assembly of functional flagella. Sugar nucleotide biosynthesis genes are often, but not always, clustered by pathway. CMP-Pse5Ac7Ac biosynthetic genes in *C. jejuni* are found in two fairly nearby loci, but in *H. pylori* they are not at all clustered, except that *pseF* and *pseG* are fused into one open reading frame. Gram-positive *Bacillus thuringiensis* converts UDP-GlcNAc to CMP-Pse5Ac7Ac in a modified, seven-step pathway, for which all the enzymes are encoded in one operon (Li et al., 2015, Delvaux et al., 2017, De Maayer and Cowan, 2016). In contrast to the *H. pylori* pathway, two *B. thuringiensis* enzymes (Pen and Pal) are required to generate UDP-2-acetamido-2,6-dideoxy-β-L-*arabino*-4-hexulose (the PseC substrate), after which the remainder

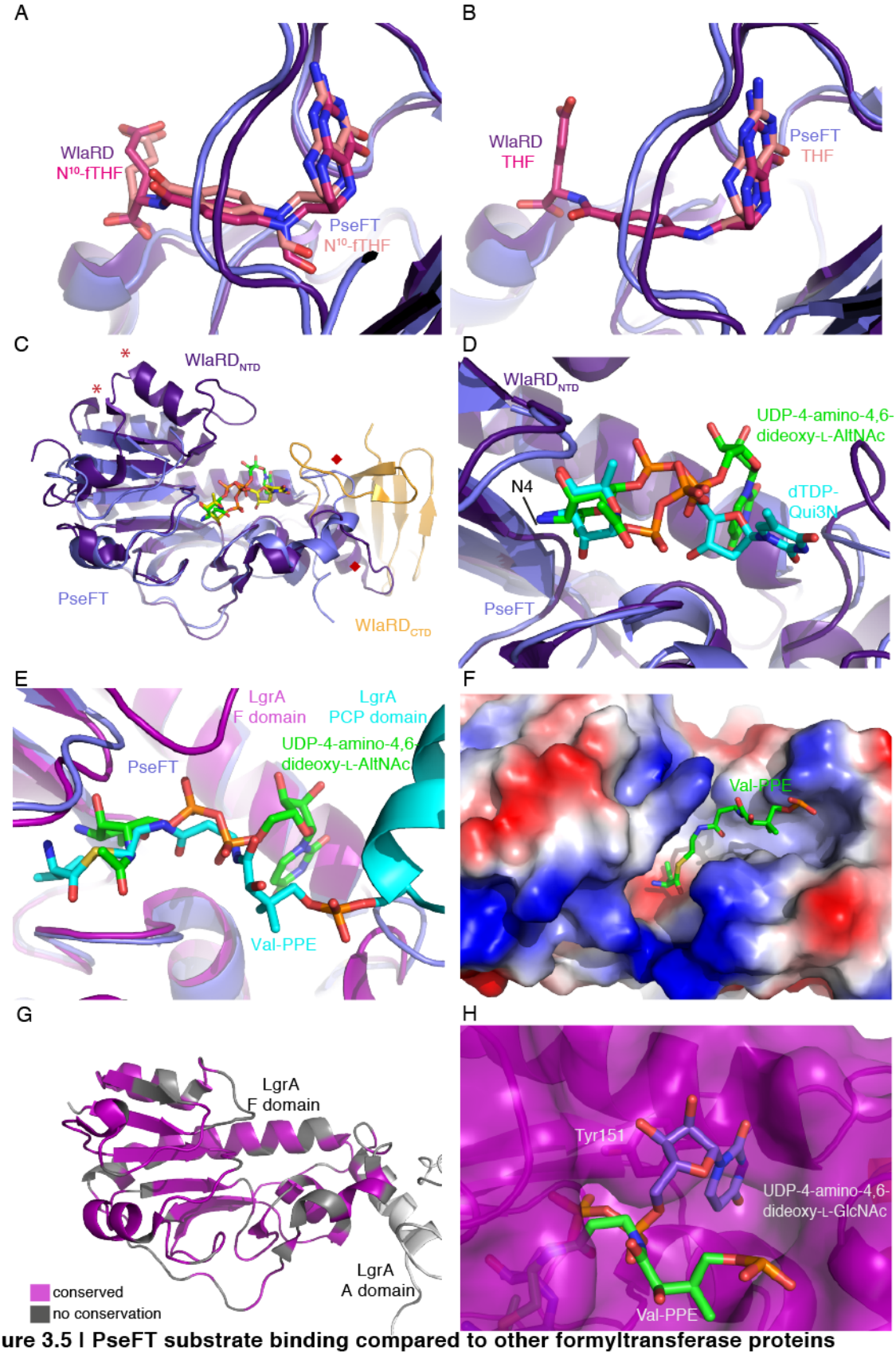


Figure 3.5 I PseFT substrate binding compared to other formyltransferase proteins
Caption on next page.

Figure 3.5 | PseFT substrate binding compared to other formyltransferase proteins.
A, B, WlaRD (Thoden et al., 2013) and other formyltransferases bind the cofactors N^{10}-fTHF and THF in similar manners. **C, D,** PseFT has an altered binding mode for its substrate, UDP-4-amino-4,6-dideoxy-L-AltNAc, compared to *O*-antigen related sugar FTs (protein WlaRD, PDB 4LY3 (Thoden et al., 2013)). Asterisks mark elements that differ or are lost in PseFT. Diamonds mark the substrate binding loop and helix α8. **E**, LgrA F domain binds Val-PPE, tethered to the PCP domain (PDB 5ES9; most of Val-PPE is modeled due to low resolution of that structure (Reimer et al., 2016a)). **F**, In the apo PseFT structure, Trp202 is rotated and occupies the nucleotide binding pocket, providing an excellent binding surface for a substrate like Val-PPE **G,** The LgrA F domain had to undergo minimal changes to switch from a stand-alone sugar FT to a functional F domain with an NRPS. Grey residues have no conservation with PseFT. See also Figure S5. **H**, The LgrA F domain has a remnant nucleotide binding pocket where Tyr151 could participate in π stacking interactions with the nucleobase, as illustrated by the superimposed PseFT substrate.

of the *H. pylori* and *B. thuringiensis* pathways are equivalent. The CMP-Pse5Ac7Fo biosynthetic gene cluster in Gram-positive *A. kamchatkensis* described here encodes six proteins, unequivocally demonstrating that six-step pathways are not confined to Gram-negative bacteria. As we demonstrate in this study, five out of six *A. kamchatkensis* enzymes share amino acid similarities with corresponding enzymes from *H. pylori* and/or *C. jejuni*, and the sixth enzyme appears to be switched from N-acetyltransferase PseH to N-formyltransferase PseFT. Interestingly, one protein returned by the BLAST search of the F domain, that from *Candidatus altiarchaeum* (GenBank accession PIX48949), is a PseH-PseFT fusion, perhaps illustrating the intermediate step of acquiring an alternative N-acyltransferase. Inactivation of *C. jejuni pseH* gene by mutation renders this organism non-motile (Guerry et al., 2006), demonstrating the biological importance of the 7-N-acyl group. Whether *C. altiarchaeum* can produce both 7-N-formamido and 7-N-acetamido derivatives of pseudaminic acid is unknown.

We biochemically characterized *A. kamchatkensis* PseB, PseC and PseFT enzymes and confirmed the role of PseFT in generating of UDP-4,6-dideoxy-4-formamido-L-AltNAc. Reconstituting the last three steps of the pathway to confirm the structure of CMP-Pse5Ac7Fo product was not central for this study. Currently, there are six stereoisomers of 5,7-diamino-3,5,7,9-tetradeoxy-non-2-ulosonates found in nature, including pseudaminic acid, legionaminic acid (Leg), C4- and C8-epimers of legionaminic acid (4eLeg and 8eLeg, respectively), acinetaminic acid (Aci) and C8-epimer of acinetaminic acid (8eAci) (Kenyon et al., 2017). The biosynthesis of Pse and Leg is known, whereas the synthetic pathways for 8eLeg, Aci and 8eAci were proposed based

on bioinformatic analyses (Kenyon et al., 2015, Kenyon et al., 2017). The 9-carbon skeleton of these nonulosonic acids is synthesized through the condensation of phosphoenolpyruvate (PEP) with hexose, thus the stereochemistry that the hexose possesses determines the configuration of C5-C8 chiral centers of the generated nonulosonic acid. Based on the defined stereochemistry of PseFT product and the naturally occurring isomers of nonulosonic acids, and together with the absence of any additional genes with epimerase, dehydrogenase or reductase activities in the *A. kamchatkensis* gene cluster, the final product is described as CMP-Pse5Ac7Fo.

PseFT is the first sugar FT involved in biosynthesis of nonulosonic acids to be assayed or structurally characterized, and has distinct structural features compared to known sugar FT structures (Thoden et al., 2013, Zimmer et al., 2014, Williams et al., 2005, Riegert et al., 2017, Genthe et al., 2015, Woodford et al., 2015, Woodford et al., 2017). PseFT is a single domain protein with only the FT catalytic domain. This is not common in sugar FTS, but is a typical feature in FTs involved in *de novo* purine synthesis. The overall size of PseFT is also smaller than the FT catalytic domain of sugar FTs because it is missing helix α2 and strand β3 (Figure 3.5C, Figure S3.5). These structural elements, also absent in the LgrA F domain (Reimer et al., 2016a), do not contain any catalytic or substrate-binding residues. The final secondary structure element of PseFT is also distinctly different from all other characterized sugar FTs. In all seven structurally characterized sugar FTs, helix α8 starts at the domain periphery and extends back toward the domain center, packing against α5 and the end of long helix α6 (Figure S3.5). After helix α8, there is a loop that forms part of the uracil binding site, followed by the C-terminal domain. In PseFT, the order and directionality of these two elements are reversed (Figure S3.5): The loop before α8 is the uracil binding loop, and helix α8 (short in PseFT) packs against α5 and α6 in the orientation opposite to α8 of other sugar FTs. This alters the mode of binding to the nucleotide (see below) and reverses the direction of the backbone of α8. Remarkably, in LgrA, the analogous position of α8 is occupied by the first helix of the A domain, and it has the same directionality as PseFT α8. (i.e., LgrA F domain α8 and A domain α1 are one and the same element) (Figure S3.5).

The C-terminal domain (CTD) attached to the FT catalytic domain is varied in function, size and structure in different sugar FTs (Figure S3.6). PseFT and similar FTs found in sequence databases do not have a CTD at all. WbtJ has a CTD consisting of a 55 residue β-hairpin motif that

facilitates dimerization (Zimmer et al., 2014); QdtF and FtdF contain a linker middle domain that connects an unexpected ankyrin repeat, which binds an additional substrate molecule (Woodford et al., 2015, Woodford et al., 2017); WlaRD has a 70 residue, four stranded β-sheet of unknown function (Thoden et al., 2013); ArnA exists as a dual function protein and, adjacent to the FT catalytic domain, has a middle domain that bridges to a C-terminal decarboxylase domain (Breazeale et al., 2005). By analogy, the A domain of LgrA could be considered a more C-terminal catalytic domain. The CTDs in sugar FTs often mediate dimerization (Woodford et al., 2017), so it is not surprising that PseFT is a monomer in solution, not having a C domain. The monomer nature of PseFT is another characteristic that may have facilitated its adoption into LgrA, as NRPSs are also monomeric.

The mode by which characterized sugar FTs and PseFT bind cofactor is very similar, but there are substantial differences in substrate binding (Figure 3.5A-D).The pterin and aminobenzoic rings of N^{10}-fTHF are in the equivalent positions in PseFT and WlaRD, and the formyl moiety is presented to the substrate in a highly similar manner (Thoden et al., 2013). In contrast, the binding of the sugar nucleotide substrates is markedly different. In PseFT, the uracil moiety is sandwiched into a π stacking pocket formed by Trp202 and Tyr151. The ribose, base and α phosphate are shifted ~90° relative to the analogous position in FTs WlaRD (Thoden et al., 2013), WbtJ (Zimmer et al., 2014), VioF (Genthe et al., 2015), QdtF (Woodford et al., 2015), FdtF (Woodford et al., 2017) and Rv3404C (Dunsirn et al., 2017) where the uracil extends toward the enzyme surface (Figure 3.5D), and is stabilized by stacking with a single aromatic residue (WlaRD Trp222) from the loop after helix α8. The α phosphorous atoms are in equivalent positions, but the β phosphates splay out in different directions and the attached sugars enter their binding pockets at very different angles, stabilized by different interactions. The difference in binding of the nucleotide moiety is caused by the reversed relative orientation of helix α8, necessitating non-analogous loops to contribute nucleotide-binding interactions. In contrast, analogous protein segments form the sugar-binding pockets in all the various enzymes, but the residue identities are not conserved in PseFT, so different side chains interact with the sugar moiety. Despite these differing modes of binding of both nucleotide and sugar, the amino group to be formylated is presented in the precisely equivalent position (Figure 3.5D).

Conversely, PseFT and the LgrA F domain are very structurally similar (Figure 3.3C, Figure S3.5), suggesting that relatively minor alterations would have been required after the gene fusion event to evolve a bona fide F domain. Because of their similar orientations of the helices, a fusion point between helix α8 of (Pse)FT and helix α8 of (pre)LgrA allowed the full folds of both FT and A domains to be preserved. The only substantial difference in backbone position between PseFT and F domain is in loops (between β1 and α2, and between α7 and α8). The residues that are part of or are near the core Rossman fold show high conservation between PseFT and F domain, while periphery residues and the substrate binding site exhibit higher variation (Figure 3.5G). In the F domain, with its new configuration and sequence, the loop connecting helices α7 and α8 functions acts as a mini-docking platform for the substrate-bearing PCP domain, instead of direct substrate binding. Despite sequence differences at the substrate binding site, an overlay of the LgrA Val-PPE substrate with apo PseFT shows that Val-PPE has an excellent steric fit and good complementation of electrostatics with the PseFT surface (Figure 3.5E, F; Note that Val-PPE is modelled because the crystal structure of the formylation state of LgrA (PDB 5ES9) has too low resolution to resolve most of the PPE moiety (Reimer et al., 2016a).) Furthermore, the F domain appears to retain a remnant of the uracil binding pocket, albeit without the second aromatic residue (PseFT Trp202) needed to bind the nucleotide (Figure 3.5H). Val-PPE is less bulky than a sugar-nucleotide, and the F domain has a larger loop between β1 and α2 which it can use to contact its substrate, whereas in PseFT, Arg83 extends to coordinate the nucleotide phosphates.

It would be desirable to further expand the known set of NRPS tailoring domains by genome mining, rational bioengineering or selection experiments. That the co-opting of a FT to make an NRPS F domain seems to have been relatively straightforward highlights the tantalizing possibility of bioengineering experiments which target additional enzymes for incorporation into NRPSs to artificially expand their repertoire of tailoring domains and the chemistry of their products.

3.4 Significance

Nonribosomal peptide synthetases are fascinating macromolecular machines that produce many diverse small molecules. NRPSs have expanded the chemical space which their products can

occupy by acquiring tailoring domains through fusion events with genes for enzymes involved in unrelated cellular processes. The linear gramicidin synthetase contains as its first domain a vital formylation tailoring domain that we identified as having sugar formyltransferase ancestry. Here, we investigate a sugar formyltransferase, PseFT, found in *Anoxybacillus kamchatkensis,* which is representative of the native FT prior to gene fusion with the NRPS. We show that PseFT participates in a newly described pathway responsible for the synthesis of CMP-7-formamidopseudaminic acid used for glycosylation of flagellin. We also present four X-ray crystal structures of PseFT alone, with cofactor, substrate, and products. These structures, the first of a sugar FT involved in biosynthesis of a nonulosonic acid, reveal substantial contrasts to other studied sugar FTs in substrate binding and architecture. Moreover, the structures reveal key adaptions that were needed to co-opt and evolve a sugar FT into a functional and useful NRPS domain. This narrative of the F domain's origins may inform future bioengineering experiments to create novel tailoring domains and thus designer nonribosomal peptides.

3.5 Acknowledgements

We are grateful to members of the T.M.S. and C.W. laboratories for advice and discussion; Shaun Labiuk and Pawel Grochulski (Canada Light Source), as well as Frank Murphy (Argonne National Laboratory) for help with diffraction data collection; Dyanne Brewer and Armen Charcholyan for technical assistance with mass spectrometry; Sameer Al-Abdul-Wahid and Andy Lo for technical support with NMR spectroscopy; and Albert Berghuis for reading of the manuscript. This work was funded by the Canadian Institutes of Health Research (FDN-148472) (T.M.S.), Canada Research Chairs (C.W. and T.M.S.) and the Natural Sciences and Engineering Research Council of Canada (C.W.)

3.6 Author Contributions

Conceptualization, J.M.R., T.M.S.; Investigation, I.H., J.M.R., O.G.O., J.J.; Formal Analysis, O.G.O.; Writing – Original Draft, J.M.R., T.M.S., I.H., O.G.O.; Writing – Review & Editing, J.M.R., I.H., O.G.O., C.W., T.M.S.; Visualization, J.M.R., I.H., O.G.O.; Funding Acquisition, C.W. and T.M.S.; Supervision, C.W. and T.M.S.

3.7 Methods

STAR★Methods

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to, and will be fulfilled by, the lead contact, T. Martin Schmeing (martin.schmeing@mcgill.ca)

3.7.1 Experimental model and subject detail

Escherichia coli BL21 (DE3) cells were used for the recombinant expression of PseFT and PseB. *Escherichia coli* Rosetta 2 (DE3) cells were used for the recombinant expression of PseC. All cells were cultured in LB media at 37 °C.

3.7.2 Method Details

3.7.2.1 Cloning of PseB, PseC and PseFT

Genes *pseB* and *pseFT* were amplified from *Anoxybacillus kamchatkensis* G10 (Genbank ALJT01000001.1) using primers PseB_fwd and PseB_rev and PseFT_fwd and PseFT_rev, respectively (Table S3.3). PCR products were ligated into a pET21-derived vector containing an N-terminal octa-histidine tag and tobacco etch virus (TEV) protease cleavage site. The *pseC* gene was synthesized by GenScript USA Inc. and cloned in a pET21-derived vector containing an N-terminal octa-histidine tag and TEV protease cleavage site.

3.7.2.2 Expression and purification of PseB, PseC and PseFT

PseB was heterologously expressed in *Escherichia coli* BL21 (DE3) cells. A 10 mL overnight culture was used to inoculate 1 L of lysogeny broth (LB) media supplemented with 35 μg mL^{-1} kanamycin. The culture was grown at 37 °C to an optical density at 600 nm (OD_{600}) of 0.6 before induction of protein expression using 0.5 mM isopropyl-ß-D-1-thiogalactopyranoside (IPTG) and incubation at 16 °C for 18 hr. Cells were harvested by centrifugation at 4500×*g* for 20 minutes at 4 °C, and pellets were resuspended in immobilized metal affinity chromatography (IMAC) buffer (25 mM Tris-HCl pH 7.5, 250 mM NaCl, 2 mM imidazole pH 8.0, 2 mM β-mercaptoethanol (βMe) and 0.1 mM phenylmethanesulfonyl fluoride (PMSF)) supplemented with deoxyribonuclease I

(BioShop). Cells were lysed by sonication and clarified by centrifugation for 30 minutes at 19000×*g* at 4 °C. Clarified lysate was loaded onto a 5 mL HiTrap IMAC FF column (GE Healthcare) charged with Ni^{2+} and equilibrated in IMAC buffer. After loading the clarified lysate, the column was washed using IMAC buffer supplemented with 75 mM imidazole, and PseB was eluted using IMAC buffer supplemented with 250 mM imidazole. Fractions containing PseB were identified using SDS-PAGE, pooled and concentrated using a 3K MWCO Amicon Ultra-15 centrifugal filtration unit (EMD Millipore). PseB was subjected to size exclusion chromatography (SEC) using a HiLoad 16/60 Superdex 200 column (GE Healthcare) in 25 mM Tris-HCl pH 7.5, 250 mM NaCl and 2 mM βMe. PseB was concentrated to 5 mg/ml, flash frozen with liquid nitrogen and stored at -80 °C.

PseFT was expressed using the same protocol as PseB, except that protein expression was induced with 1 mM IPTG. Cells were harvested and lysed using the same protocol. The clarified lysate was loaded onto a 5mL HiTrap IMAC FF column equilibrated in IMAC buffer. The column was washed with IMAC buffer containing 25 mM imidazole and PseFT was eluted using a 60 mL gradient from 25 mM to 250 mM imidazole in IMAC buffer. Fractions containing PseFT were pooled and diluted with 25 mM Tris-HCl pH 7.5 and 2 mM βMe to reduce the NaCl concentration to 25 mM. PseFT was applied to a 5 mL HiTrap Q HP (GE Healthcare) column and eluted using a 60 mL gradient of 0 – 200 mM NaCl. PseFT was concentrated using a 3K MWCO Amicon Ultra-15 filtration unit and applied to a HiLoad 16/600 Superdex 75 column (GE Healthcare) equilibrated in 25 mM Tris pH 7.5, 150 mM NaCl and 2 mM βMe. Pure PseFT was concentrated to 5 mg/ml, flash frozen with liquid nitrogen and stored at -80 °C.

PseC was expressed in *Escherichia coli* Rosetta 2 (DE3) cells (EMD Millipore) grown in LB media supplemented with 34 μg mL^{-1} chloramphenicol and 100 μg mL^{-1} ampicillin. Protein expression was induced with 0.1 mM IPTG and grown at 16 °C for 18 hr. Cells were harvested and lysed as described above. Clarified lysate was loaded onto a 5 mL HiTrap IMAC FF column, the column was washed with IMAC buffer containing 5 mM imidazole, and PseC was eluted using a 60 mL gradient of 5 – 250 mM imidazole in IMAC buffer. PseC was then subjected to SEC using a HiLoad 16/60 Superdex 200 (GE Healthcare) column equilibrated in 25 mM Tris pH 7.5, 150 mM NaCl and 2 mM βMe. Purified PseC was concentrated to 5 mg/ml, flash frozen with liquid nitrogen and stored at -80°C.

3.7.2.3 Synthesis of 5,10-methenyl-THF

5,10-methenyl-THF was synthesized from N^5-formyl-THF (Sigma-Aldrich) as described previously (Breazeale et al., 2002). Briefly, 15.8 μmol N^5-formyl-THF was dissolved in 1.5 ml 1% (v/v) aqueous β-mercaptoethanol, the pH was adjusted to 1.9 by adding 0.1 M HCl, after which the volume was brought to 2.2 mL with water, and the sample was incubated at room temperature. Formation of 5,10-methenyl-THF was monitored by following increase in absorbance at 355 nm. Upon reaction completion (~ 20 min), the mixture was stored at –20 °C.

3.7.2.4 PseB, PseC and PseFT activity assays

The activities of PseB, PseC and PseFT were assayed by reverse phase-HPLC-UV and LC-ESI-MS. Reactions were carried out in 60 μl of 50 mM Tris-HCl pH 7.4. To test PseB activity, the reaction mixture contained 0.5 mM UDP-GlcNAc and 0.42 μM PseB. PseC activity was assayed in a coupled reaction containing 0.5 mM UDP-GlcNAc, 15 mM L-glutamic acid monosodium salt (L-Glu), 1 mM pyridoxal 5′-phosphate (PLP) (Sigma-Aldrich), 0.42 μM PseB and 0.7 μM PseC. PseFT activity was tested in a one-pot reaction containing all components for the PseB and PseC reactions, 2 mM 5,10-methenyl-THF and 3.91 μM PseFT. The reaction mixture containing all components except enzymes was preincubated for 10 minutes at 25 °C to convert 5,10-methenyl-THF to N^{10}-fTHF, after which the enzymes were added. All reactions were incubated for 16 h at 25 °C in a thermal cycler, the lid of which was heated to 50 °C to prevent condensation. The reactions were stopped by addition of 60 μl chloroform, followed by centrifugation at 13 000×*g* for 5 minutes. The aqueous layer (2 μl) was analysed by LC-ESI-MS. For HPLC-UV analysis, 40 μl of the aqueous layer was mixed with 80 μl water, and 20 μl of this mixture was injected into the HPLC system (Beckman Coulter). A Synergi 4 μm Fusion-RP 80 Å LC column (250 × 4.6 mm) was used for chromatographic separation with the following solvents: 50 mM triethylammonium acetate pH 6.8 (solvent A) and acetonitrile (solvent B). The mobile phase gradient was as follows: 1% B isocratic gradient for 1 min; linear increase to 5% B over 30 min; linear increase to 10% B over 10 min; return to 1% B over 5 minutes, which was held for 10 minutes for re-equilibration. The flow rate was maintained at 1 mL min^{-1}, and elution was monitored by UV detection at 254 nm.

3.7.2.5 LC-ESI-MS

LC-ESI-MS analyses were performed on an Agilent 1260 HPLC liquid chromatograph interfaced with an Agilent UHD 6540 Q-TOF mass spectrometer at the Mass Spectrometry Facility of the Advanced Analysis Centre, University of Guelph. Chromatographic separation was performed with an Agilent Poroshell 120 PFP column (50 × 4.6 mm, 2.7 µm) using water with 0.1% formic acid (solvent A) and acetonitrile with 0.1% formic acid (solvent B). The mobile phase gradient was as follows: initial 5% B isocratic gradient for 1 minute; increase to 100% B over 15 minutes; column wash at 100% B for 1 minute; column re-equilibration for 10 minutes. The flow rate was maintained at 0.4 mL min^{-1}. The mass spectrometer electrospray capillary voltage was maintained at 4.0 kV and the drying gas temperature at 250 °C, with a flow rate of 8 mL min^{-1}. The nebulizer pressure was 30 psi and the fragmenter was set to 160. Nitrogen was used as both nebulizing and drying gas. The mass-to-charge ratio was scanned across the *m/z* range of 50–1,500 in 4 GHz (extended dynamic range) negative-ion mode. The acquisition rate was set at two spectra per second. The instrument was externally calibrated with the ESI Tune Mix (Agilent). The data was analyzed using MassHunter Workstation Software (Agilent).

*3.7.2.6 Enzymatic synthesis of UDP-4-amino-4,6-dideoxy-L-AltNAc (**3**) for NMR spectroscopy*

Synthesis of UDP-4-amino-4,6-dideoxy-L-AltNAc was performed in a 12 mL reaction mixture containing 50 mM Tris-HCl pH 7.4, 0.5 mM UDP-GlcNAc, 15 mM L-Glu, 1 mM PLP), 0.42 µM PseB and 0.7 µM PseC in a 25 °C water bath for 16 hours. The reaction was lyophilized and the dry residue dissolved in 2 mL water, centrifuged at 4700×*g* for 5 minutes and passed through a Microcon-10kDa centrifugal filter to remove protein. The reaction product was purified by HPLC using a semi-preparative Synergi 4 µm Fusion-RP 80 Å LC column (250 × 10 mm) at a flow rate of 5 mL min^{-1}, using the same mobile phase gradient described above for the analytical column. The desired fraction was collected manually and lyophilized. Purified product was dissolved in water and lyophilized several times to remove residual triethylammonium acetate.

*3.7.2.7 Enzymatic synthesis of UDP-4,6-dideoxy-4-formamido-l-AltNAc (**4**) for NMR spectroscopy*

For synthesis of UDP-4,6-dideoxy-4-formamido-L-AltNAc, fresh N^{10}-fTHF was prepared by dissolving 12.3 mg N^5-fTHF in 3.6 ml water and 1% (v/v) 2-mercaptoethanol. The pH was adjusted to 1.9 with 1 M HCl and the reaction mixture was incubated at room temperature until the absorbance at 355 nm ceased increasing. Next UDP-GlcNAc, L-Glu and PLP were added, and the solution incubated for 10 minutes at 25 °C to allow conversion of 5,10-methenyl-THF to N^{10}-fTHF, before enzyme addition. The full reaction, consisting of 12 mL of 50 mM Tris-HCl pH 7.4, 0.5 mM UDP-GlcNAc, 15 mM L-Glu, 1 mM PLP, 2 mM N^{10}-fTHF, 0.42 µM PseB, 0.70 µM PseC and 3.91 µM PseFT was incubated at 25 °C for 16 h. The UDP-4,6-dideoxy-4-formamido-L-AltNAc product was purified by semi-preparative HPLC as described above.

3.7.2.8 NMR spectroscopy

^{1}H and ^{13}C NMR spectra were obtained using a Bruker Avance III 600 MHz spectrometer equipped with a 5 mm TCI cryoprobe, and ^{31}P spectra were collected using a Bruker 400 MHz Avance III spectrometer equipped with a 5 mm broadband Prodigy cryoprobe, both located in the NMR Centre of the Advanced Analysis Centre at the University of Guelph. The sample temperature was regulated to 295 ± 1 K for purified compounds and 298 ± 1 K for reaction monitoring experiments. Samples typically included sodium 3-trimethylsilylpropanoate-2,2,3,3-d4 (TSP) as a chemical shift reference in the ^{1}H and ^{13}C dimensions (δ_H = 0 ppm, δ_C = –1.6 ppm). Referencing of ^{31}P spectra was performed by substitution with a solution of 85% phosphoric acid (δ_P = 0 ppm).

1D selective TOCSY and NOESY experiments were collected using presaturation of the water signal during the relaxation delay period; the TOCSY employed 100 ms of DIPSI2 mixing while the NOESY employed an 800 ms mixing time. 2D spectra (COSY, TOCSY, HSQC, and HMBC) were collected using standard pulse sequences and parameters with the following exceptions: the TOCSY mixing time was 100 ms, and the optimal HMBC coupling constant was set to 8 Hz.

To characterize PseB reaction products, the reaction was carried out in a 5 mm NMR tube. PseB was exchanged into deuterated 25 mM phosphate buffer (pD 7.6) by centrifugal ultrafiltration at 4 °C. Before addition of enzyme, ^{1}H NMR and HSQC spectra of substrate solution

(10 mM UDP-GlcNAc, 1 mM TSP, 25 mM phosphate buffer pD 7.6, D_2O, 550 µL total volume) were recorded. The reaction was initiated by addition of 50 µL (315 µg) PseB to give a 600 µL solution containing 9.2 mM UDP-GlcNAc and 13.3 µM PseB. ^1H NMR spectra (8 scans) were recorded every 10 min during first 1.5 h and then after longer time intervals up to 27 h.

To characterize **3** and **4**, purified enzymatic reaction products were deuterium-exchanged by freeze-drying twice from 99.9 % D_2O. The samples were suspended in 250 µL 99.99 % D_2O and placed in 5 mm Shigemi NMR microtube.

3.7.2.9 Crystallography and diffraction data collection

PseFT was assayed for crystallization using a sitting drop, vapor diffusion method against commercially-available screening solutions Classics I, Classics II, JCSG+, PACT and PEGS (Qiagen). Two hundred nanoliters of PseFT at 2.5 mg min^{-1} or 5.1 mg min^{-1} concentration was mixed with 200 nL reservoir solution and equilibrated against 50 µl reservoir solution in 96-well sitting drop crystallization trays, at 22 °C. Crystallization occurred using a crystallization solution of 0.2 M $(NH_4)_2SO_4$, 0.1 M 2-(N-morpholino)ethanesulfonic acid (MES) pH 6.5, and 30% w/v polyethylene glycol monomethyl ether 5000 (PEG5000 MME). Iterative optimization of crystallization conditions led to a final protocol for crystallization of PseFT in 24-well sitting drop crystallization trays whereby 2 µL PseFT (5 mg min^{-1}) was mixed with 2 µL of a crystallization solution with fine variations of pH and PEG5000 MME concentrations. The crystallization solution for unliganded crystals of PseFT used for diffraction experiments was 0.2 M (NH4)2SO4, 0.1 M MES pH 6.6 and 30.67% PEG5000 MME. Crystals were cryo-protected by transfer into crystallization solution supplemented with 15% glycerol, looped into CryoLoops (Hampton Research) and flashed cooled in liquid nitrogen.

For crystallography of PseFT bound to UDP-4-amino-4,6-dideoxy-L-AltNAc, the crystallization solution consisted of 0.2 M $(NH_4)_2SO_4$, 0.1 M MES pH 6.5 and 28.4% PEG5000 MME. After crystal growth, the crystals were transferred into a solution of 5 mM UDP-4-amino-4,6-dideoxy-L-AltNAc, 0.2 M $(NH_4)_2SO_4$, 0.1 MES pH 6.5, 25 mM Tris pH 7.5, 150 mM NaCl and 31.24% PEG5000 MME, and incubated for 15-30 minutes. The crystals were then transferred for

cryoprotection into a solution containing the above-listed components and supplemented with 20% glycerol, looped and flash cooled in liquid nitrogen.

For crystallography of PseFT bound with products, the crystallization solution consisted of 0.2 M $(NH_4)_2SO_4$, 0.1 M MES pH 6.6, 28.16% PEG5000 MME. After crystal growth, the crystals were transferred into a solution of 5 mM of UDP-4,6-dideoxy-4-formamido-L-AltNAc, 2.5 mM THF, 0.2 M $(NH_4)_2SO_4$, 0.1 M MES pH 6.6, 28.16% PEG5000 MME, 10% ethylene glycol, incubated for 4 hours, looped, and flash cooled in liquid nitrogen.

For crystallography of PseFT bound with N^{10}-fTHF, crystallization solution consisted of 0.2 M$(NH_4)_2SO_4$, 0.1M MES pH 6.5, and 27.3% PEG5000MME. N^{10}-fTHF was generated using the procedure described above. The crystals were transferred into a solution of 6.25 mM N^{10}-fTHF, 0.2 M$(NH_4)_2SO_4$, 0.1M MES pH 6.5, and 27.3% PEG5000MME, 10% ethylene glycol, incubated for 4 hours, looped and flash cooled in liquid nitrogen.

Diffraction data sets for crystals of unliganded PseFT and of PseFT bound to UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF were collected using beamline 08ID-1 of the Canadian Light Source (Saskatoon, Saskatchewan, Canada) with a wavelength of 0.979 Å and at temperature of 100 K. Diffraction data sets for crystals of PseFT bound to UDP-4-amino-4,6-dideoxy-L-AltNAc and of PseFT bound to N^{10}-fTHF were collected using beamline 24-ID-C at the Advanced Photon Source (Argonne, Illinois, USA) with a wavelength of 0.979 Å and at a temperature of 100 K.

Diffraction data from unliganded PseFT crystals was indexed and integrated using iMosflm (Leslie and Powell, 2007) and scaled using Aimless (Evans and Murshudov, 2013). Structure determination of PseFT in the C1 2 1 space group was performed using molecular replacement with the F domain from of LgrA F-A (PDB 5ES6) (Reimer et al., 2016a) as a search model using the program Phaser (McCoy et al., 2007). The programs COOT (Emsley et al., 2010) and PHENIX (Adams et al., 2010) were used for iterative building and refinement to produce the final PseFT structure (Table S3.2).

Diffraction data from the crystals of the complex of PseFT with UDP-4-amino-4,6-dideoxy-L-AltNAc was indexed, integrated and scaled using HKL2000 (Otwinowski and Minor, 1997). Data from crystals of PseFT with UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF was indexed and integrated using DIALS (Winter et al., 2018) and scaled using Aimless (Evans and Murshudov,

2013). Data from crystals of PseFT with N^{10}-fTHF was indexed and integrated using iMosflm (Leslie and Powell, 2007) and scaled using Aimless (Evans and Murshudov, 2013). In each case, structures were determined by isomorphous replacement. The program eLBOW (Moriarty et al., 2009) was used to generate starting models and geometric restraints for the ligands. Iterative modeling in the program COOT (Emsley et al., 2010) and refinement in the program PHENIX (Adams et al., 2010) were used to produce the final structures (Table S3.2).

3.7.3 Data availability

Atomic coordinates and structure factors for PseFT, PseFT bound to N^{10}-THF, PseFT bound to UDP-4-amino-4,6-dideoxy-L-AltNAc, and PseFT bound to UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF have been deposited in the Protein Data Bank under accession codes 6CI2, 6CI3, 6CI4 and 6CI5, respectively (will be released upon publication).

3.7.4 Key Resources Table

Table 3.2 | Key Resources Table.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
Escherichia coli BL21 (DE3)	New England Biolabs	C2527I
Escherichia coli Rosetta 2 (DE3)	Novagen	71400
Anoxybacillus kamchatkensis G10	Lee et al., 2012	ALJT01000001.1
Brevibacillus parabrevis	ATCC	ATCC 8185
Chemicals, Peptides, and Recombinant Proteins		
PseB	This study	N/A
PseC	This study	N/A
PseFT	This study	N/A
UDP-GlcNAc	Sigma-Aldrich	Cat#U4375
UDP-4-amino-4,6-dideoxy-L-AltNAc	This study	N/A
UDP-4-formamido-4,6-dideoxy-L-AltNAc	This study	N/A
NcoI	New England Biolabs	Cat#R0193S
NotI	New England Biolabs	Cat#R0189S
XhoI	New England Biolabs	Cat#R0146S
IPTG	Fischer Scientific	Cat#FLBP1755100
PMSF	Bioshop	Cat#PMSF123
DNase I	Bioshop	Cat#DRB001
Kanamycin	BioBasic	Cat#KB0286
Chloramphenicol	Bioshop	Cat#CLR201.1
Ampicillin	Bioshop	Cat#AMP201
Ammonium sulfate	Bioshop	Cat#AMP303
2-(N-morpholino)ethanesulfonic acid	Bioshop	Cat#MES503
PEG5000 MME	Sigma-Aldrich	Cat#81323
Folinic acid calcium salt	Sigma-Aldrich	Cat#F7878
Pyridoxal 5'-phosphate	Sigma-Aldrich	Cat#P9255
ESI TuneMix	Agilent	Cat#G1969-85000
Deposited Data		
PseFT	This paper	PDB: 6CI2
PseFT complexed with N^{10}-formyltetrahydrofolate	This paper	PDB: 6CI3
PseFT complexed with UDP-4-amino-4,6-dideoxy-L-AltNAc	This paper	PDB: 6CI4
PseFT complexed with UDP-4-formamido-4,6-dideoxy-L-AltNAc and tetrahydrofolate	This paper	PDB: 6CI5
Oligonucleotides		
Primers for cloning and mutagenesis, see Table S3		
Recombinant DNA		
Plasmid: PseB	This paper	N/A
Plasmid: PseFT	This paper	N/A
Plasmid: PseC	GenScript	Cat#SC1691
Software and Algorithms		
T-Coffee	Notredame et al., 2000	http://tcoffee.crg.cat/apps/tcoffee/do:regular

Clustal Omega	Sievers et al., 2011	http://www.ebi.ac.uk/Tools/msa/clustalo/
Ident and Sim	Stothard, 2000	http://www.bioinformatics.org/sms2/ident_sim.html
Coot	Emsley et al., 2010	http://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
PHENIX	Adams et al., 2010	https://www.phenix-online.org/
Phaser	Adams et al., 2010	https://www.phenix-online.org/
Aimless	Evans and Murshudov, 2013	http://www.ccp4.ac.uk/html/aimless.html
DIALS	(Winter et al., 2018)	https://dials.github.io/
Pymol	Schrödinger	http://www.pymol.org
eLBOW	Moriarty et al., 2009	https://www.phenix-online.org/
wwPDB Validation server		http://wwpdb-validation.wwpdb.org
MassHunter Workstation Software	Agilent	https://www.agilent.com/en/products/software-informatics/masshunter-suite/masshunter/mashunter-software
Other		
HiTrap IMAC FF	GE Healthcare	Cat#17092104
HiTrap Q HP	GE Healthcare	Cat#17115301
HiLoad 16/60 Superdex 200 pg	GE Healthcare	Cat#28989335
HiLoad 16/60 Superdex 75 pg	GE Healthcare	Cat#28989333
3K MWCO Amicon Ultra-15	EMD Millipore	Cat#UFC900308
Microcon-10kDa Centrifugal Filter	EMD Millipore	Cat#MRCPRT010
Poroshell 120 PFP	Agilent	Cat#699975-408T
Synergi 4 µm Hydro-RP 80 Å, LC column (250 x 10 mm)	Phenomenex	Cat#00G-4375-N0

3.8 Supplemental Information

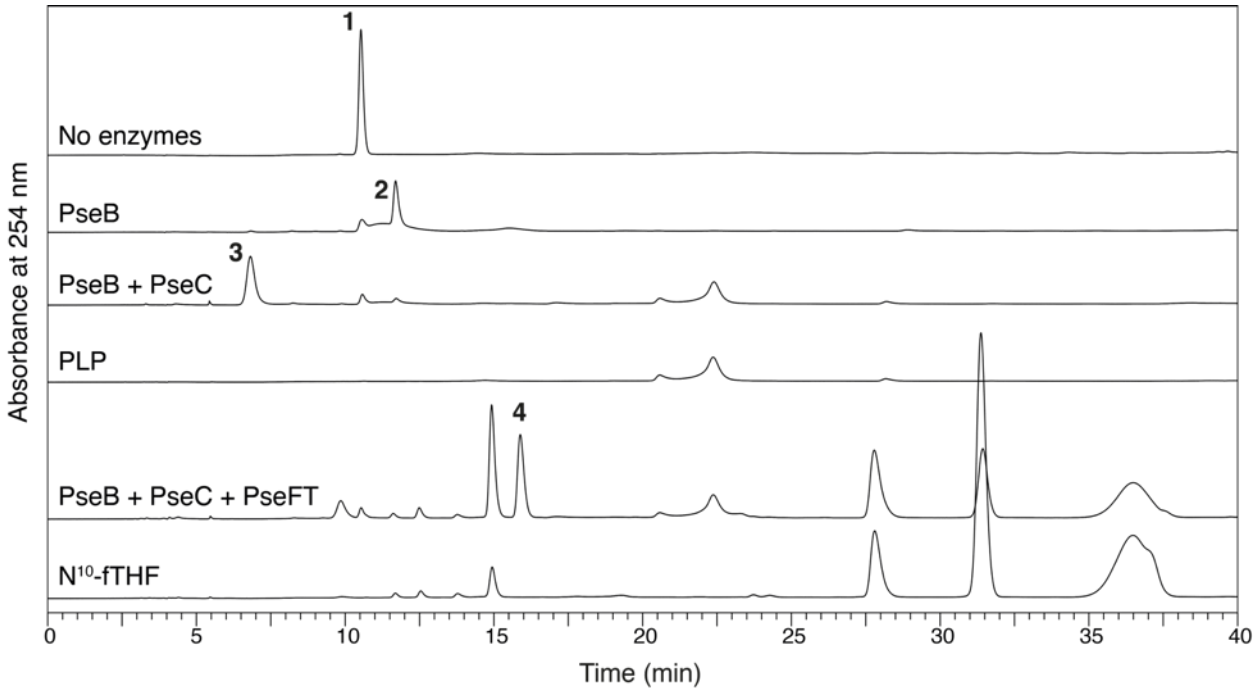


Figure S3.1 | In vitro activity assay of PseB, PseC and PseFT.
Related to Figures 1 and 2.
HPLC analysis of initial substrate UDP-GlcNAc, products of enzymatic reactions, PLP (PseC cofactor) and formyl donor N^{10}-fTHF. **1** - UDP-GlcNAc; **2** - UDP-2-acetamido-2,6-dideoxy-β-L-*arabino*-4-hexulose; **3** - UDP-4-amino-4,6-dideoxy-L-AltNAc; **4** - UDP-4,6-dideoxy-4-formamido-L-AltNAc.

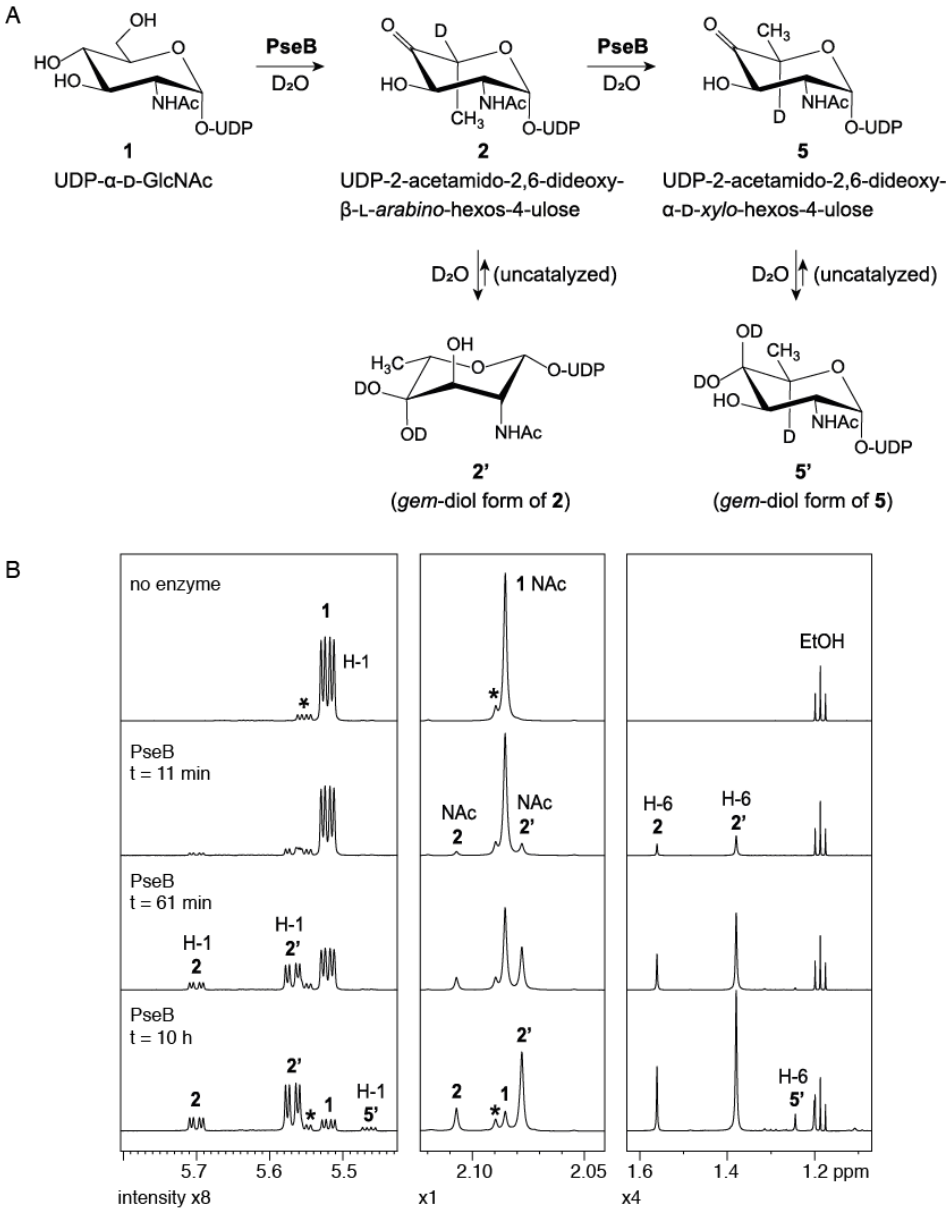


Figure S3.2 |PseB reaction monitored by NMR spectroscopy. *Related to Figures 1.*
A, The first step of Pse biosynthesis. PseB-catalyzed C-4, C-6 dehydration/C-5 epimerization of UDP-GlcNAc in D_2O yields keto-sugar **2** with solvent-derived deuterium incorporated at C-5. Further PseB-catalyzed C-5 epimerization of **2** gives keto-sugar **5**. Both **2** and **5** exist in equilibrium with their *gem*-diol (hydrated) forms. **B**, Monitoring PseB reaction by ^{1}H NMR (600 MHz) at 25°C. The reaction mixture contained 9.2 mM UDP-GlcNAc, 13.3 µM PseB, 25 mM phosphate buffer in D_2O, pD 7.6. The anomeric region and the region of resonances of methyl groups ($NCOCH_3$ and H-6) are shown. Due to incorporation of deuterium at C-5 the signals for methyl groups H-6 appear as singlets. The asterisk indicates the signals for contaminating UDP-GalNAc (present in the commercial substrate), which is not a substrate for PseB.

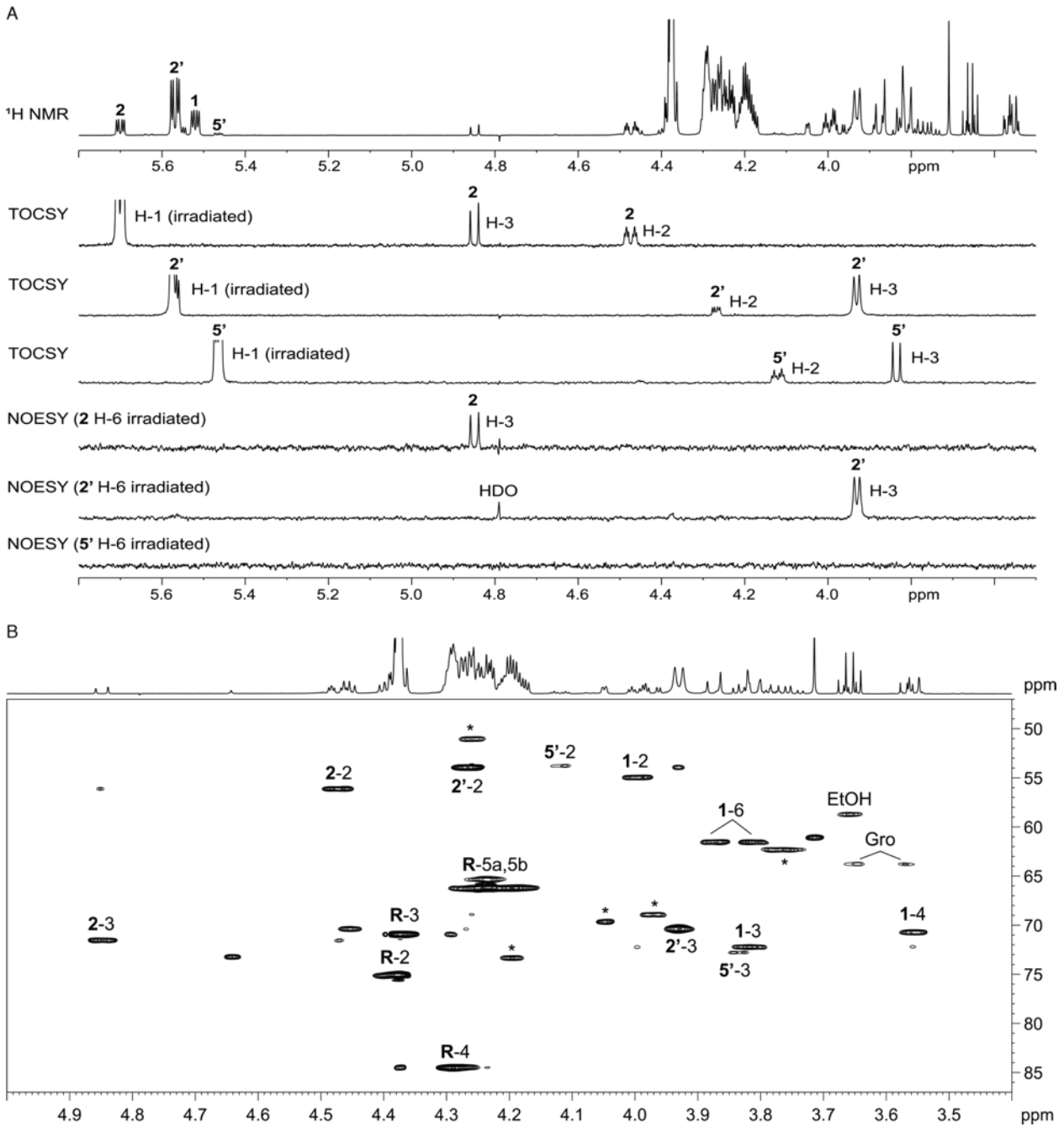


Figure S3.3 | NMR spectra of PseB reaction. *Related to Figure 1.*
A, ^{1}H NMR spectrum, 1D selective TOCSY (90 ms) and 1D selective NOESY (800 ms). For selective 1D experiments, irradiated signals are indicated. **B**, Part of ^{1}H,^{13}C HSQC spectrum, recorded after 10 h. The asterisk indicates the signals for contaminating UDP-GalNAc (present in the commercial substrate), which is not a substrate for PseB.

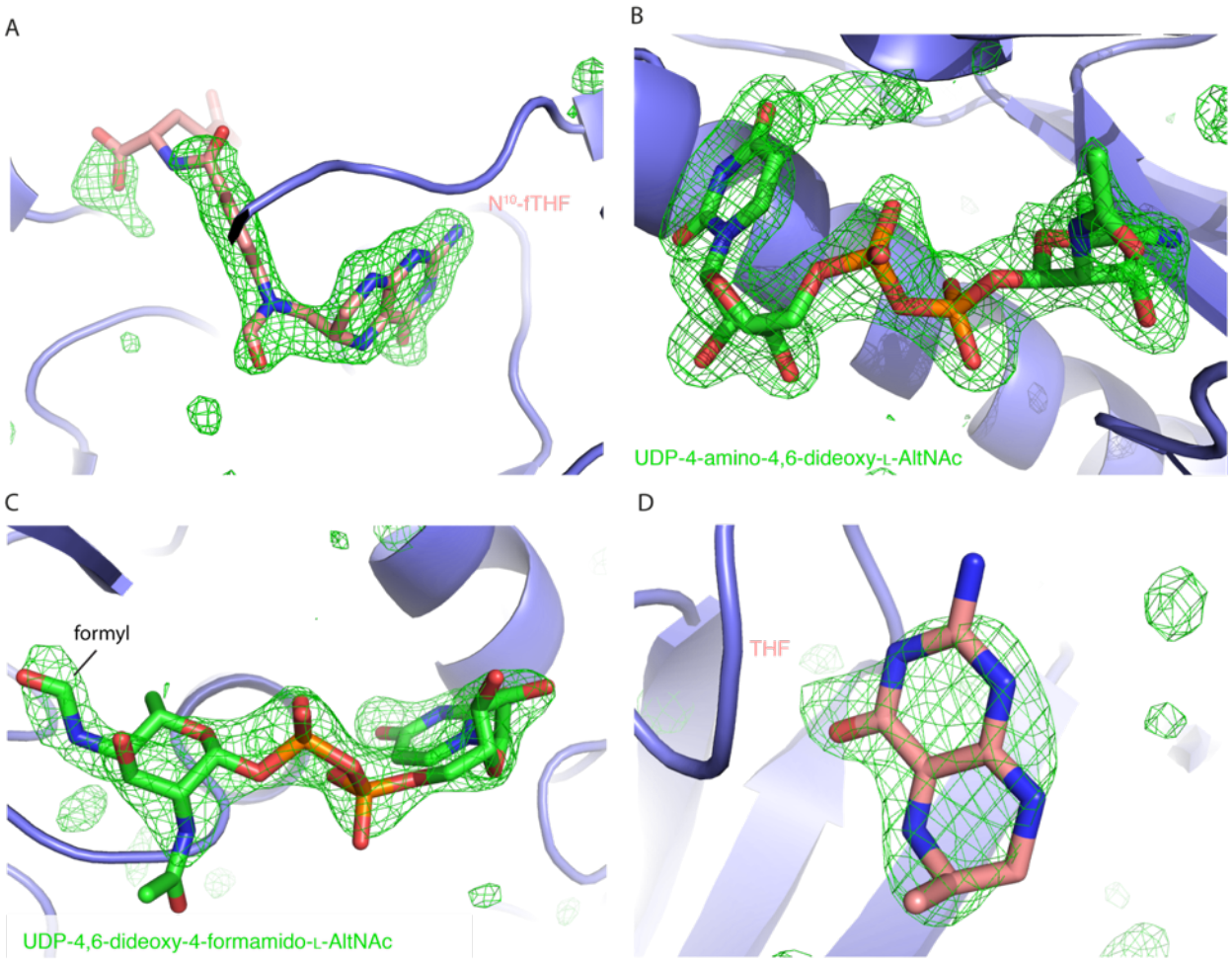


Figure S3.4 | Electron density for PseFT structures. *Related to Figures 3 and 4.*
A-D, Unbiased Fo-Fc electron maps contoured at 3σ for crystal soaking experiments (A) with N^{10}-fTHF co-factor, (**B**) with PseFT substrate UDP-4-amino-4,6-dideoxy-L-AltNAc, and with double-soak of (**C**) product UDP-4,6-dideoxy-4-formamido-L-AltNAc and (**D**) THF.

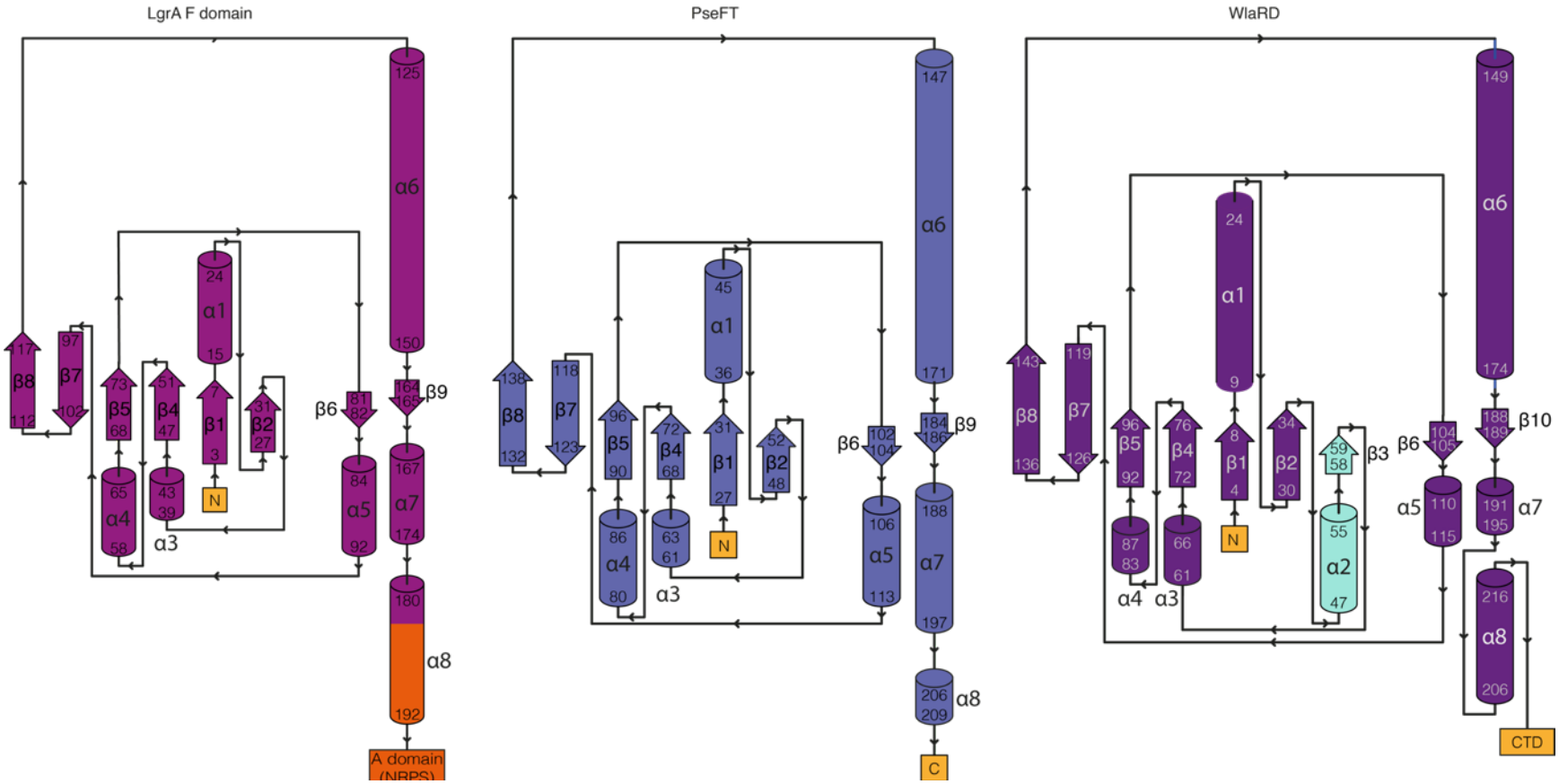


Figure S3.5 | Topology diagrams of formyltransferases. *Related to Figures 3, 4 and 6.*
Topology diagrams for the LgrA F domain (Reimer et al., 2016a), PseFT and sugar FT WlaRD (Thoden et al., 2013). The F domain and PseFT have lost strand β3 and helix α2 (colored teal in WlaRD), and show different connectivity to helix α8 that affects the mode of substrate binding. Preliminary diagrams were generated using PDBsum (Laskowski, 2009).

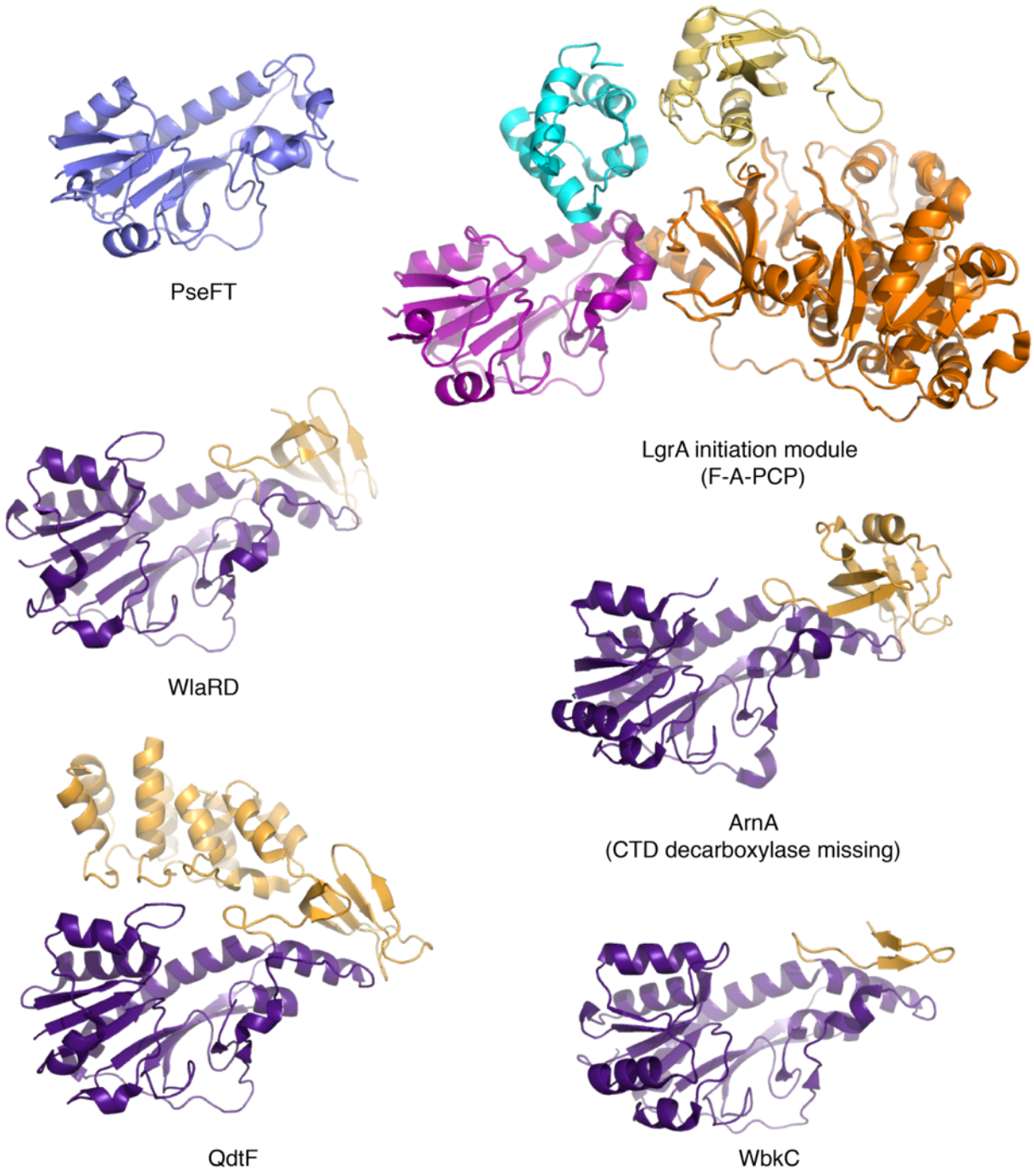


Figure S3.6 | Sugar formyltransferase C-terminal domains. *Related to Figure 5.*
The C-terminal domains of sugar formyltransferases are extremely varied in both structure and function. WbkC, PDB 5VYR (Riegert et al., 2017); WlaRD, PDB 4LY3 (Thoden et al., 2013); ArnA, PDB 2BLN (Williams et al., 2005); QdtF, PDB 4XD0 (Woodford et al., 2015); LgrA F–A–PCP, PDB 5ES9 (Reimer et al., 2016a).

Table S3.1 | NMR chemical shifts (δ, ppm) and coupling constants (J, Hz) for UDP-GlcNAc and PseB, PseC and PseFT products.

Moiety		*H-1* C-1 $J_{(1,2)}$	*H-2* C-2 $J_{(2,3)}$	*H-3* C-3 $J_{(3,4)}$	*H-4* C-4 $J_{(4,5)}$	*H-5* C-5 $J_{(5,6)}$	*H-6* C-6	NAc[a] CH_3	Fo
α-D-Glc*p*NAc (**1**)	δ_H	*5.52*	*4.00*	*3.81*	*3.56*	*3.93*	*3.81; 3.88*	*2.09*	
	δ_C	95.8	55.0	72.2	70.8	74.3	61.6	23.4	
	$^3J_{H,P}$	7.3							
2-acetamido-2,6-dideoxy-β-L-*arabino*-hexose-4-ulose (**2**)	δ_H	*5.70*	*4.47*	*4.85*			*1.56*	*2.11*	
	δ_C	96.3	56.1	71.5	210.5	78.9	21.5	23.4	
	$^3J_{H,H}$	3.1	11.7						
	$^3J_{H,P}$	8.0							
2-acetamido-2,6-dideoxy-β-L-*arabino*-hexose-4-ulose (*gem*-diol form) (**2'**)	δ_H	*5.57*	*4.27*	*3.93*			*1.38*	*2.08*	
	δ_C	95.7	53.9	70.4	94.5	76.7	16.9	23.4	
	$^3J_{H,H}$	3.1	7.7						
	$^3J_{H,P}$	8.4							
2-acetamido-2,6-dideoxy-α-D-*xylo*-hexose-4-ulose (*gem*-diol form) (**5'**)	δ_H	*5.46*	*4.12*	*3.83*			*1.24*		
	δ_C	95.6	53.8	72.8	94.5	70.9	12.5		
	$^3J_{H,H}$	3.5							
	$^3J_{H,P}$	7.0							
4-amino-4,6-dideoxy-β-L-Alt*p*NAc[b] (**3**)	δ_H	*5.61*	*4.18*	*4.07*	*2.90*	*3.95*	*1.35*	*2.05*	
	δ_C	94.6	53.9	68.8	53.3	73.7	19.3	23.2	
	$^3J_{H,H}$	2.3	5	3	8.1	6.5			
	$^3J_{H,P}$	8.4							
4,6-dideoxy-4-formamido-β-L-Alt*p*NAc (major *Z* isomer) (**4**)	δ_H	*5.65*	*4.17*	*4.04*	*3.98*	*4.02*	*1.28*	*2.06*	*8.13*
	δ_C	94.4	54.3	68.4	50.8	72.3	19.0	23.2	165.4
	$^3J_{H,H}$	2.2							
	$^3J_{H,P}$	8.5							
4,6-dideoxy-4-formamido-β-L-Alt*p*NAc (minor *E* isomer) (**4**)	δ_H	*5.64*	*4.17*	*4.11*	*3.58*	*4.08*	*1.33*	*2.06*	*8.06*
	δ_C	94.5	54.3	68.9	55.8	72.4	19.1	23.2	168.8
Ribose (R)	δ_H	*5.99*	*4.36*	*4.36*	*4.27*	*4.17; 4.23*			
	δ_C	89.6	75.1	70.8	84.4	66.1			
Uridine (U)	δ_H					*5.97*	*7.97*		
	δ_C		153.6		168.3	103.9	142.8		

[a]The signals for the N-acetyl group (CO) are at δ_C 175.7–175.9.
[b]The signals for diphosphate group are at δ_P –10.9 and –13.0.

Table S3.2 | Crystallographic statistics for data collection and processing.

Related to Figures 3 and 4.

Values in parenthesis denote the highest-resolution shell.

	PseFT	PseFT with N^{10}-fTHF	PseFT with UDP-4-amino-4,6-dideoxy-L-AltNAc	PseFT with UDP-4,6-dideoxy-4-formamido-L-AltNAc and THF
Diffraction Data				
Space group	C 1 2 1	C 1 2 1	C 1 2 1	C 1 2 1
Unit-cell parameter (Å, °)	95.63 77.51 41.73 90.00 103.84 90.00	97.20 77.30 40.46 90.00 103.9 90.00	96.23 73.78 41.80 90.00 103.47 90.00	96.42 73.78 41.80 90.00 103.47 90.00
Resolution (Å)	59.50-1.96 (2.00-1.96)	47.18 – 1.80 (1.84-1.80)	46.89-1.82 (1.85-1.82)	47.23-1.73 (1.73-1.70)
I/σ	16.6 (5.9)	12.2 (2.0)	12.1 (0.957)	11.4 (1.2)
Temperature (K)	100	100	100	100
Measured reflections	79 498 (4851)	157 911 (9645)	67 938 (2696)	204 645 (9811)
Unique reflections	21 304 (1482)	26 389 (1627)	23 891 (1025)	32 206 (1720)
Completeness (%)	99.4 (98.3)	97.9 (99.1)	93.9 (82.0)	100.0 (100.0)
Multiplicity	3.7 (3.3)	6.0 (5.9)	2.8 (2.6)	6.5 (5.7)
Rmerge (%)	0.046 (0.124)	0.064 (0.763)	0.080 (0.593)	0.042 (0.984)
Estimates of resolution limits (Å):				
From half-dataset correlation CC(1/2) > 0.50	1.96	1.83	1.82	1.95
From Mn(I/sd) > 2.00	1.96	1.90	1.96	2.0
Refinement statistics				
Resolution range (Å)	31.33 – 1.96	47.17 – 1.8	46.89 – 1.82	46.88 – 2.0
R factor/R_{free} (%)	18.0 / 22.9	17.5 / 20.7	18.4 / 20.9	20.6 / 22.7
Average B factor (Å^2)	34.0	46.0	47.0	72
R.M.S.D in bond lengths (Å)	0.017	0.017	0.011	0.004
R.M.S.D in bond angles (°)	1.330	1.500	1.217	0.693
Ramachandra plot (%)				
Favored	99.47	98.43	97.87	98.40
Allowed	0.53	1.57	2.13	1.60
Outliers	0.0	0.0	0.0	0.0
PDB ID	6CI2	6CI3	6CI4	6CI5

Table S3.3 | Primers used in this study. *Related to STAR Methods.*

Oligonucleotide name	Sequence (5' → 3')
Primers for PseB and PseFT	
PseB_fwd	ACGTTACCATGGAAATGTTTGAAAATCAAGTCGTCCTT
PseB_rev	GTACTAGCGGCCGCTTACATTCCATCTACCAACTCTCG
PseFT_fwd	TGACTACCATGGGGAAAATTTTATTGTTAGGTCCT
PseFT_rev	CGATCGGCGGCCGCTCATTCGTTATT

3.9 Segue to Chapter 4

The last step of the initiation cycle in LgrA is donating formyl-valine to the adjacent elongation module for formation of the first peptide bond of linear gramicidin. It was speculated how a donor PCP domain would interact with the C domain based on biochemical data and structures of PCP domains interacting with C domain homologues. However, a bonafide PCP-C domain structure depicting the substrate donation state had not been reported at the time. Indeed, very little was known about the overall architecture and organization of multi-modular NRPSs. My goal going into Chapter 4 was to answer how formyl-valine is handed off to the elongation module using a crystallographic approach combined with chemical biology tricks. Three different constructs of LgrA, each including the entire initiation module and 1 to 3 domains of the elongation module, were put into crystallographic trials. The resulting structures reveal the elegance and ingenuity guiding nonribosomal peptide synthesis. The work presented in Chapter 4 is currently being prepared for publication and will be submitted for peer review in the near future.

CHAPTER 4 | STRUCTURES OF A DIMODULAR NONRIBOSOMAL PEPTIDE SYNTHETASE PROTEIN

Janice M. Reimer, Ingrid Harb, Maximilian Eiavaskhani, and T. Martin Schmeing. Structures of a dimodular nonribosomal peptide synthetase protein. *Manuscript in preparation.*

4.1 Introduction

Nonribosomal peptide synthetases (NRPSs) are intricate macromolecular machines capable of manufacturing small molecules with incredible chemical diversity and functionality (Walsh, 2004). Product synthesis is guided by modular assembly-line logic involving large domain movements and a complex network of active sites (Weissman, 2015). A canonical module contains three domains to successfully add one building block to the final peptide: a peptidyl carrier protein (PCP) domain that is modified with a prosthetic phosphopantetheinyl (PPE) arm, an adenylation (A) domain that activates amino acids by adenylation and then covalently attaches the activated amino acid onto the PPE arm, and a condensation (C) domain that covalently links substrates together through peptide bonds (Hur et al., 2012). The A domain itself is comprised of two subdomains: an A_{core} domain that contains the active site, and a smaller A_{sub} domain that provides catalytic residues to the adenylation reaction. Additionally, the A_{sub} domain undergoes substantial conformational changes to facilitate both parts of the A domain reaction and help the PCP domain travel between active sites (Conti et al., 1997, Reger et al., 2008, Yonus et al., 2008, Gulick, 2009, Mitchell et al., 2012). Modules can be expanded by integrating tailoring domains to provide additional chemical modifications to the product, which are often necessary for the bioactivity of the molecule (Walsh et al., 2001). The linear gramicidin synthetase contains a tailoring formylation (F) domain at the N-terminus of its first dimodular subunit, LgrA (Kessler et al., 2004). The F domain catalyzes N-formylation of valinyl-PPE to generate formyl-valinyl-PPE prior to substrate donation to the second module (Schoenafinger et al., 2006, Reimer et al., 2016a).

An excellent structural understanding of a module's synthetic cycle has been gained over the past decade using structures containing individual domains (Weissman, 2015) and up to entire initiation (Reimer et al., 2016a) and termination modules (Tanovic et al., 2008, Drake et al., 2016, Miller et al., 2016). Together, these structures illustrate the large intra-modular domain rearrangements needed for the PCP domain to transport substrates between active sites over 50 Å apart. However, much less is known about how modules work together in the context of the larger NRPS. There are only two reported crystal structures that contain domains from two adjacent modules. The TycC PCP_5-C_6 didomain shows the PCP domain in an unproductive

conformation (Samel et al., 2007), and DhbF A_1-$PCP_{1\text{-gly-AVS}}$-C_2:MLP[3] (MLP, MbtH like protein) (Tarry et al., 2017), in which the sole inter-module contact would have to be broken in the course of peptide synthesis. The only 3D data for a multi-modular NRPS are low-resolution negative stain electron microscopy (EM) reconstructions (26-29 Å) of dimodular DhbF (C_1-A_1-PCP_1-C_2-A_2-PCP_2:MLP; MLP, MbtH-like protein) (Tarry et al., 2017). These reconstructions showed heterogeneity in the module:module conformation, despite the protein being stalled in the thiolation state by mechanism-based inhibitors (Tarry et al., 2017, Mitchell et al., 2012, Sundlov et al., 2012, Sundlov and Gulick, 2013, Drake et al., 2016, Miller et al., 2016). In absence of definitive data, several hypothetical models of multi-modular NRPSs have been constructed by consecutively overlapping multi-domain structures from different synthetases (Marahiel, 2016, Reimer et al., 2018). The different models did not reveal any consistent modular organization, and without structural data to validate these *in silico* models, fundamental questions regarding NRPS synthesis remain. Higher-quality data is needed to fully understand NRPS organization and the architecture throughout a synthetic cycle of an NRPS.

We have determined seven structures of large constructs of LgrA to resolutions between 2.2 and 6.7 Å (Figure 4.1, Supplemental Figure 4.1). The crystallized constructs include the full first (initiation) module and between one and all three canonical domains from the second (elongation) module, and have domain organizations F_1-A_1-PCP_1-C_2, F_1-A_1-PCP_1-C_2-A_2 and F_1-A_1-PCP_1-C_2-A_2-PCP_2. Each structure is in a catalytically-relevant state and some reveal previously unobserved catalytic states such as a full condensation conformation, where both donor and acceptor PCP domains are bound to the C_2 domain. The structures demonstrate that a multi-modular NRPS stalled at the same catalytic state can have markedly different overall conformations, and that conformational changes on the scale of hundreds of ångströms are likely to occur during nonribosomal peptide synthesis.

[3] Subscripts after a PCP_n domain demark the loaded state of the domain with *n* corresponding to the module number of the synthetase. $PCP_{n\text{-SH}}$, holo PCP_n. $PCP_{n\text{-amino acid-AVS}}$, PCP_n modified with an adenosine-vinylsulfonamide (AVS) inhibitor. $PCP_{n\text{-SH-fVal}}$ or $PCP_{n\text{-NH-fVal}}$, PCP_n domains modified with a CoA or amino-CoA analogue, such as formyl-valine-NH-CoA, using Sfp.

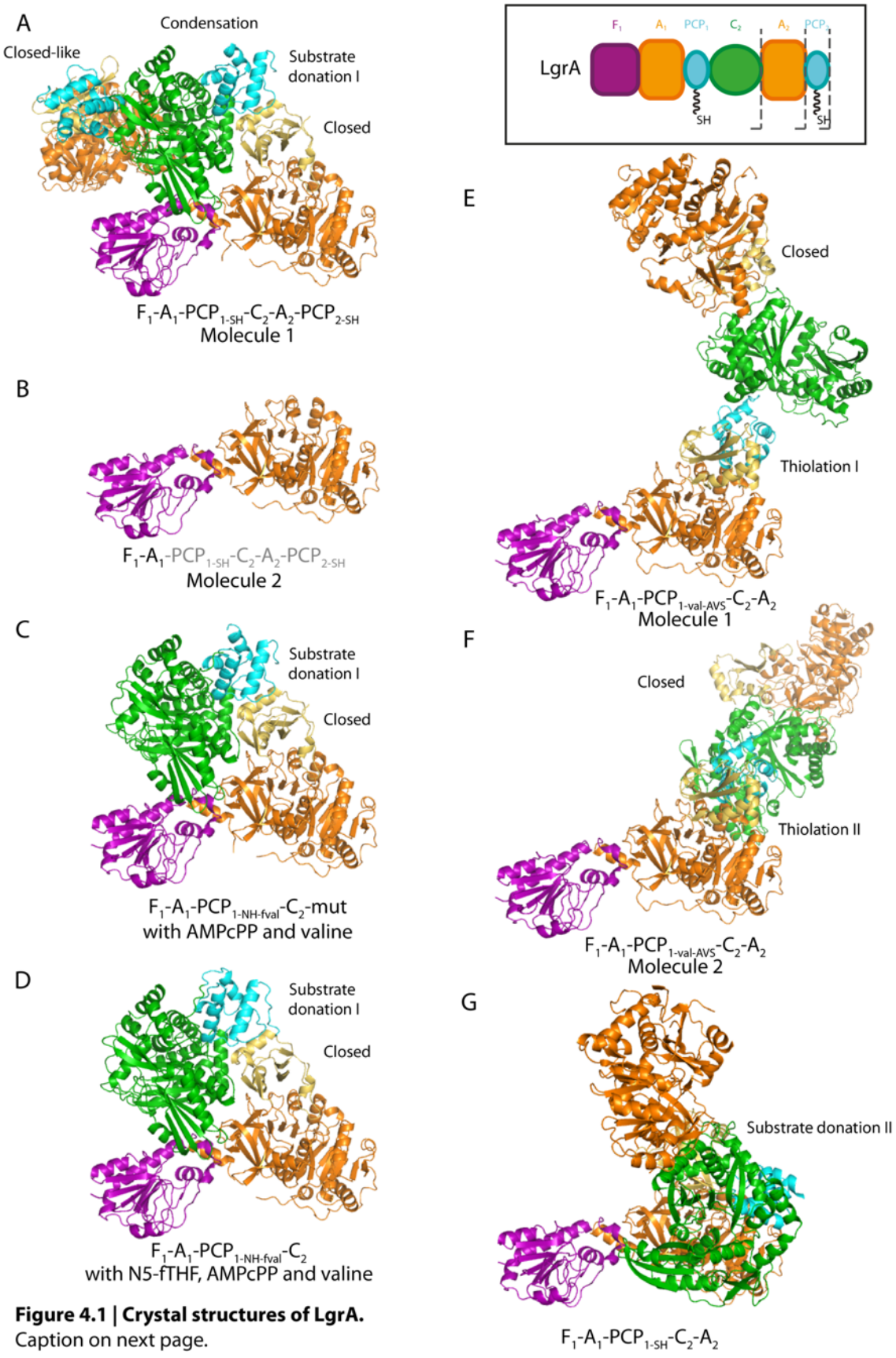


Figure 4.1 | Crystal structures of LgrA.
Caption on next page.

Figure 4.1 | Crystal structures of LgrA.
Models of (**A, B**) F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ (PCP_1-C_2-A_2-PCP_2 disordered in Molecule 2), (**C**) F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut with AMPcPP and valine, (**D**), F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ with N5-formytetrahydrofolate (N5-fTHF), AMPcPP and valine, (**E, F**) F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ and (**G**) F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 (Parts of A_{2core} and all of A_{2sub} are disordered). Domain colours: F, purple; A_{core}, orange; A_{sub}, yellow; PCP, cyan; C, green. Inset: LgrA domain organization. Dashed lines show different crystallization constructs. Colour code is maintained throughout text.

4.2 Results

4.2.1 LgrA crystallography

The excised initiation module of LgrA is a well behaved protein and proved highly amenable to crystallization efforts (Reimer et al., 2016a, Reimer et al., 2016b). We wanted to take advantage of LgrA's robust nature to address questions relating to the modular organization of NRPSs. Full length LgrA, F_1-A_1-PCP_1-C_2-A_2-PCP_2-E_2, contains two modules with an inactive epimerization (E) domain at its C-terminus. Six constructs were designed and used in crystallization trials: F_1-A_1-PCP_1-C_2, F_1-A_1-PCP_1-C_2-mut with C31S, C191S, C318S, R792C mutations, F_1-A_1-PCP_1-C_2-A_2, F_1-A_1-PCP_1-C_2-A_2-$PCP_{2(short)}$, F_1-A_1-PCP_1-C_2-A_2-PCP_2 and F_1-A_1-PCP_1-C_2-A_2-PCP_2-E_2. The use of chemical biology strategies were essential for obtaining crystals of LgrA. We targeted specific catalytic states by modifying PCP_1 and/or PCP_2 with the following small molecules: coenzyme A (CoA), amino-CoA (NH-CoA) (Liu and Bruner, 2007, Reimer et al., 2016a), (formyl)-valinyl-SH-CoA, (formyl)-valinyl-NH-CoA, glycinyl-NH-CoA, and valinyl-adenosine vinylsulfamonamide (val-AVS) and gly-AVS inhibitors (Qiao et al., 2007b). Crystal optimization was aided by co-crystallization with ligands or ligand analogues. Further details on constructs and crystallization experiments are provided in the Methods and Supplemental Information sections below.

4.2.2 Structures of LgrA – an overview

We determined the structures of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ with valine, AMPcPP and N5-formyltetrahydrofolate (N5-fTHF) in $P2_1\ 2_1\ 2_1$ at 2.5 Å, F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut with valine and AMPcPP in P1 at 2.2 Å, holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 in $P2_1\ 2_1\ 2_1$ at 2.8 Å, F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ in $P2_1\ 2_1\ 2_1$ at 6.4 Å and holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ in C2 2 2_1 at 6.7 Å (Figure 4.1, Supplemental Figures 4.1-3, Supplemental Table 4.1). The structures show that the initiation and elongation

modules do not form a permanent interface with each other and can be in many different orientations relative to each other. Two of the larger construct structures are low resolution, but the complimentary high resolution structures allow us to construct excellent low resolution models that illustrate key domain-domain interactions and overall dimodular architecture. The F_1-A_{1core} didomain of the initiation module and the C_2-A_{2core} didomain of the elongation modules form the structural unit of each respective module. In each structure, the F_1-A_{1core} domains adopt an elongated conformation very similar to what we observed in structures of smaller constructs (Reimer et al., 2016a, Reimer et al., 2016b). In the current conformations, the interface between the two domains buries ~882-893 $Å^2$ of surface area and allows a slight hinging motion between structures that causes the F_1 domain to bend up to ~6° relative to the A_{core} domain and displaces analogous atoms of the F_1 domain up to ~10.6 Å. Likewise, the C_2:A_2 interface of the elongation module shows mild variation between all four dimodular structures, burying between ~606-804 $Å^2$ of surface area and bending up to ~8° relative to the C_2 domain. As with previous structures of modules, these show the didomain core of each module to form a catalytic 'platform' (Tanovic et al., 2008). The initiation module F_1 and A_1 domain active sites and the elongation module acceptor PCP side of the C_2 domain active site and the A_2 domain active site are localized on the same side of their respective module to facilitate substrate transfer. Impressively, while the most pronounced intra-modular conformational changes are localized to the PCP and A_{sub} domains, our series of structures demonstrate that adjacent modules make massive movements relative to each other. The necessity for the PCP domain to travel long distances within its own module led us to ask if the sizeable inter-modular movements we observe in our structures is a consequence of the initiation module's catalytic cycle or simply arbitrary conformations resulting from intrinsic flexibility between the two module or a combination of the two. We will continue to return to this question as the different conformations are analyzed.

4.2.3 LgrA during the initiation module's thiolation state

The crystals of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 contain two molecules in the asymmetric unit with the initiation modules stalled in the thiolation state through the action of the val-AVS inhibitor.

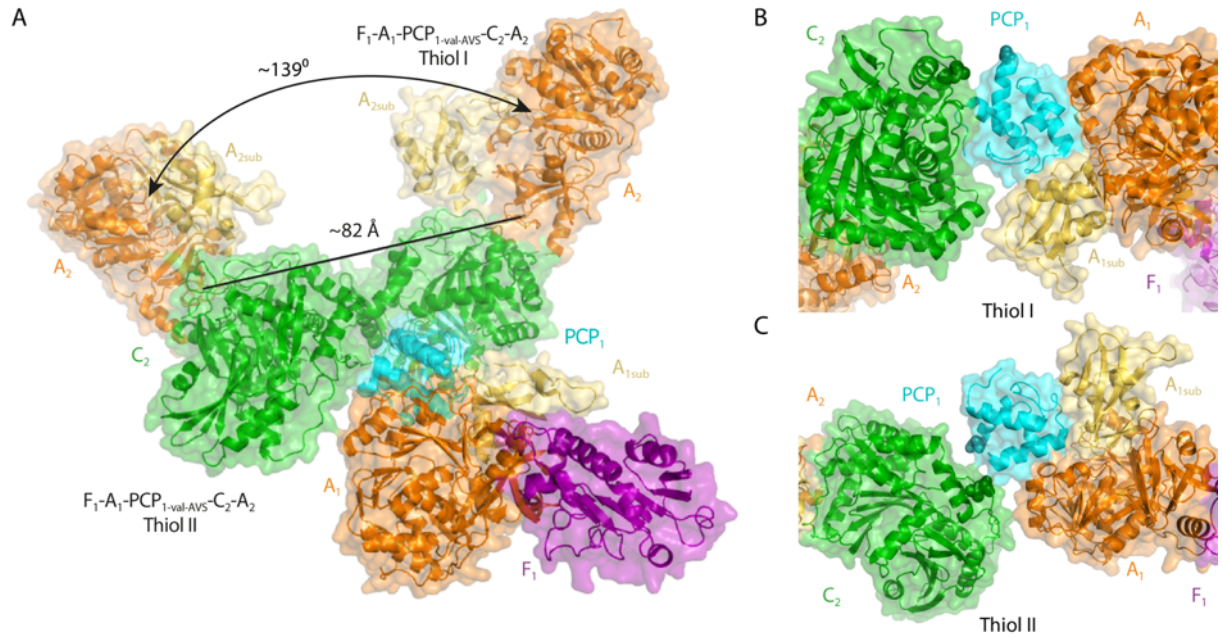


Figure 4.2 | Initiation thiolation in LgrA.
A, An overlay of the two molecules found in the F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 structure illustrate the elongation modules are in two very different locations during the thiolation state of the initiation module. Close up of the (**B**) Thiol I and (**C**) Thiol II inter-module region. The last ordered residue of the PCP_1 domain and the first ordered residue of the C_2 domain are shown with spheres. The linker connecting the PCP_1 and C_2 domains is predominantly disordered.

The conformation of the initiation modules in both structures resemble the thiolation structure of F_1-A_1-$PCP_{1\text{-NH-val}}$ (Reimer et al., 2016a). However, the elongation modules have remarkably different orientations relative to the initiation module. The distance between the C_2-A_{2core} domain center of masses is ~82 Å and they are rotated ~139° relative to each other (Figure 4.2A). These large differences are possible because the initiation and elongation modules do not form substantial interactions with each other, consistent with the lack of contacts between A_1 and C_2 domains in the structure of DhbF A_1-$PCP_{1\text{-gly-AVS}}$-C_2:MLP (Tarry et al., 2017). In the conformation of one molecule of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 (herein called thiolation conformation I; Thiol I), the back "face" of the PCP_1 domain is positioned towards the donor PCP binding site of the C_2 domain (Figure 4.2B). The A_1-PCP_1 linker is partially disordered, but the C_2 domain C-lobe does make some contacts with a portion of the linker closest to the PCP_1 domain. In the conformation of the other molecule, Thiol II, only a small portion of the C_2 domain floor loop contacts the PCP_1-C_2 linker. The

PCP_1-C_2 linker is predominantly disordered and acts as a flexible tether between the two modules. The A_{2sub} domain is in the "closed" adenylation-competent position in both molecules.

A domain-domain interaction between the back face of a PCP domain and a downstream partner, like that between PCP_1 and C_2 in the Thiol I conformation, has been observed in other structures that include a PCP and C or Te domain (Drake et al., 2016, Samel et al., 2007, Tarry et al., 2017). Our lab and Gulick's lab previously suggested that this non-catalytic interaction may possibly help the C_{n+1} domain (or Te_n) to remain close to the PCP_n domain, to promote more efficient transfer of substrate to the downstream module or domain (Tarry et al., 2017). This interaction must be transient since it must be broken for the PCP to deliver its PPE-bound intermediate. In the Thiol II conformation, the back face of the PCP_1 domain is > 10 Å from the C_2 domain (Figure 4.2C), showing the interaction is not obligatory even when the PCP is not delivering substrate downstream.

4.2.4 The LgrA condensation state

The central chemical event of peptide synthesis is condensation, which occurs when donor (here PCP_1) and acceptor PCP (here PCP_2) domains bind simultaneously at the C domain. The structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ represents both the first intact view of an NRPS in this state, and the first detailed 3D view of an intact multi-modular NRPS. (The F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ crystal also contains a second molecule in which PCP_1 and the entire second module are disordered.) Because all the domains other than the ~100 residue A_{2sub} and ~90 residue PCP_2 are present in our other high-resolution structures, we are able to build a high-quality model for the full dimodular protein (Figure 4.1, Supplemental Figure 4.1, Figure 4.3A).

The acceptor PCP binding site on the C_2 domain is located at the junction of the N-lobe and C-lobe. The PCP_2 domain interacts with $C_2\alpha1$ and $C_2\alpha10$ and nearby loops. Electron density for the phosphate of the PPE arm is at the entrance of the active site tunnel (Supplemental Figure 4.3G). The PCP_2 domain is rotated ~22° compared to the acceptor PCP domain in the structure of holo AB3404 (C-A-PCP_{SH}-Te) (Figure 4.3B) (Drake et al., 2016). This suggests that the acceptor PCP can bind the C domain in multiple conformations as long as it competently delivers the aminoacyl-PPE

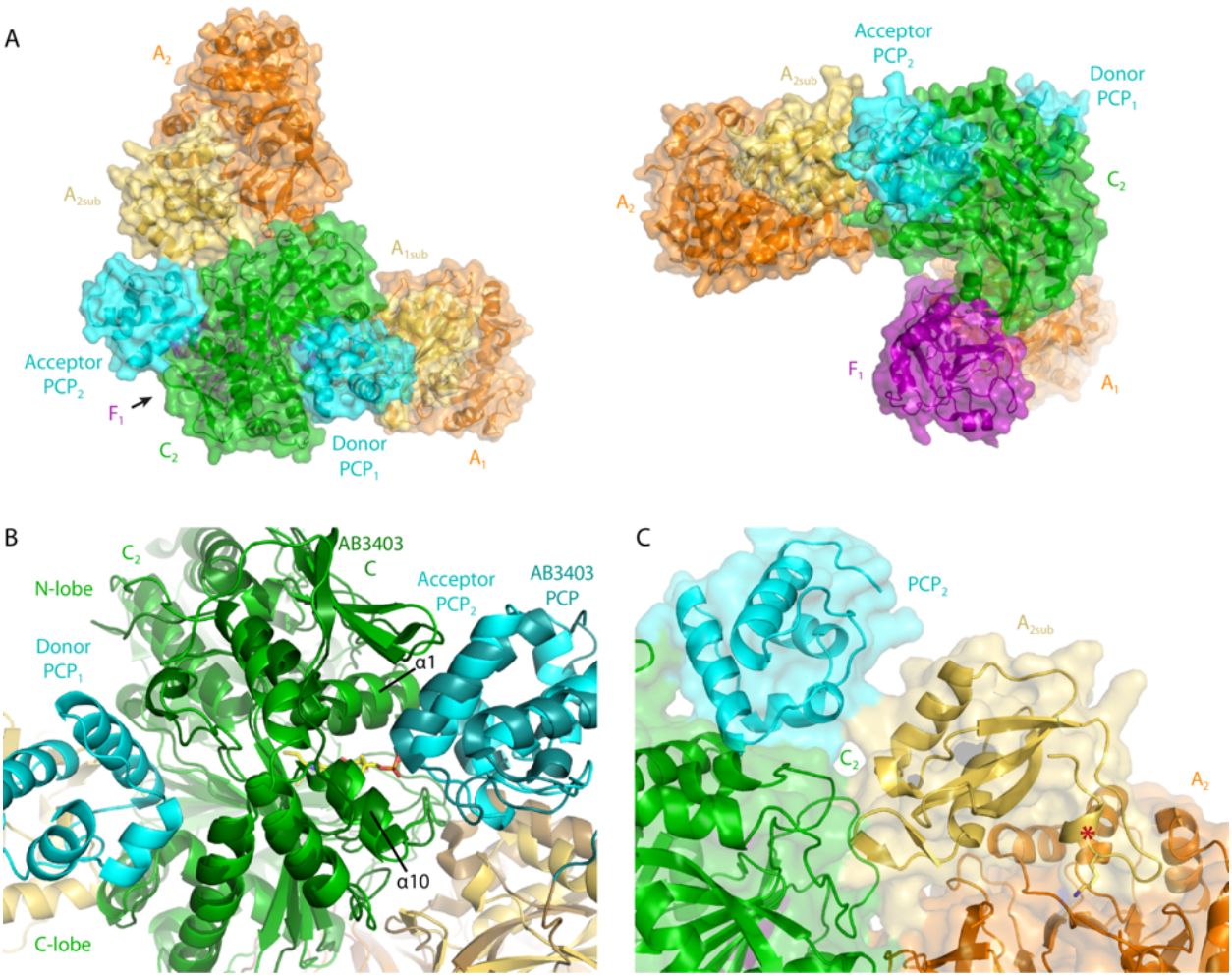


Figure 4.3 | The condensation state in LgrA.
A, Two views of the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$. **B**, Superposition of the holo AB3404 structure onto the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$, aligned to the C domain. The PCP_2 domain is rotated by ~22° compared to AB3404 (Drake et al., 2016). Domain colour code for AB3403: A_{sub}, dark tan; C, dark green; PCP, dark teal. **C**, The A_{2sub} domain is in a pseudo-closed position. Asterisk denotes the catalytic lysine, Lys1704.

substrate to the active site. Unlike in other states, the PCP_2 domain does not contact the A_{2sub} domain and is only positioned by the direct contacts with the acceptor site of C_2. The A_{2sub} is in a pseudo-closed state and would have to rotate by ~22° and shift its center of mass by ~4 Å to bring Lys1704 into a catalytically competent position for adenylation of glycine (Figure 4.3C). In contrast, the structure of holo AB3404 shows how a module can simultaneously be in the condensation state (PCP at acceptor binding site on C domain) and adenylation state with a fully closed A_{sub} domain (Drake et al., 2016).

Four of our structures show PCP_1 bound to the C_2 domain donor site. This is the first observation of a productive interaction between a PCP_n domain and canonical C_{n+1} domain, though analogous interactions have been observed with C domain homologues in the didomain structures of GrsA PCP-E (Chen et al., 2016) and TqaA PCP-C_T (C_T, terminal condensation-like domain) (Zhang et al., 2016). We see two different C_2 domain positions relative to the initiation module: holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$, F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$ and F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut are in substrate donation conformation I (SD I), and holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 is in substrate donation conformation II (SD II) (Figure 4.1, Supplemental Figure 4.1). In both conformations, the C_2 domain is rotated back against the initiation module and makes contacts with either the F_1 domain (in SD I) or A_1 domain (in SD II). Among the three structures in SD I, there are small differences in the angle of the F_1 domain:A_1 domain orientation, there are minor shifts in where the N-lobe of the C_2 domain contacts the F_1 domain, and the C_2 domain center of mass varies by ~3.6 Å, but in each, the N-lobe of the C domains blocks access to the formylation active site (Figure 4.4A). The F_1:C_2 interface varies accordingly and buries ~319-750 $Å^2$ of surface area. Furthermore, we solved several additional structures of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ and F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut (unpublished) in multiple crystal packing environments and all of these display the same general SD I conformation, with minor ~3-5 Å shifts in domain placement. The variability between structures speaks to the transitory nature of the interaction. In the SD II conformation, the location of the elongation module differs dramatically. The entire module is rotated ~114° around the F_1-A_1 didomain, and the C_2-A_{2core} didomain center of mass is ~80 Å from its equivalent position in SD I (Figure 4.4B). This causes the N-lobe of the C_2 domain to rest instead against the A_{core} domain (Figure 4.4C), and buries a modest ~419 $Å^2$ of surface area. Despite this massive rearrangement, the PCP_1:C_2 interaction is similar, as described below.

The A_{1sub} domain adopts very different conformations between the two substrate donation conformations. In SD I, the A_{1sub} domain is in the adenylation conformation (or closed position) and Lys672 coordinates the ATP analogue, AMPcPP (Figure 4.4D). The A_{1sub} domain contacts the PCP_1 domain, seemingly to help stabilize it at the C_2 domain, an interaction reminiscent of that in PCP_1 delivery to the F_1 domain for formylation. The center of mass of the A_{1sub} domain differs by

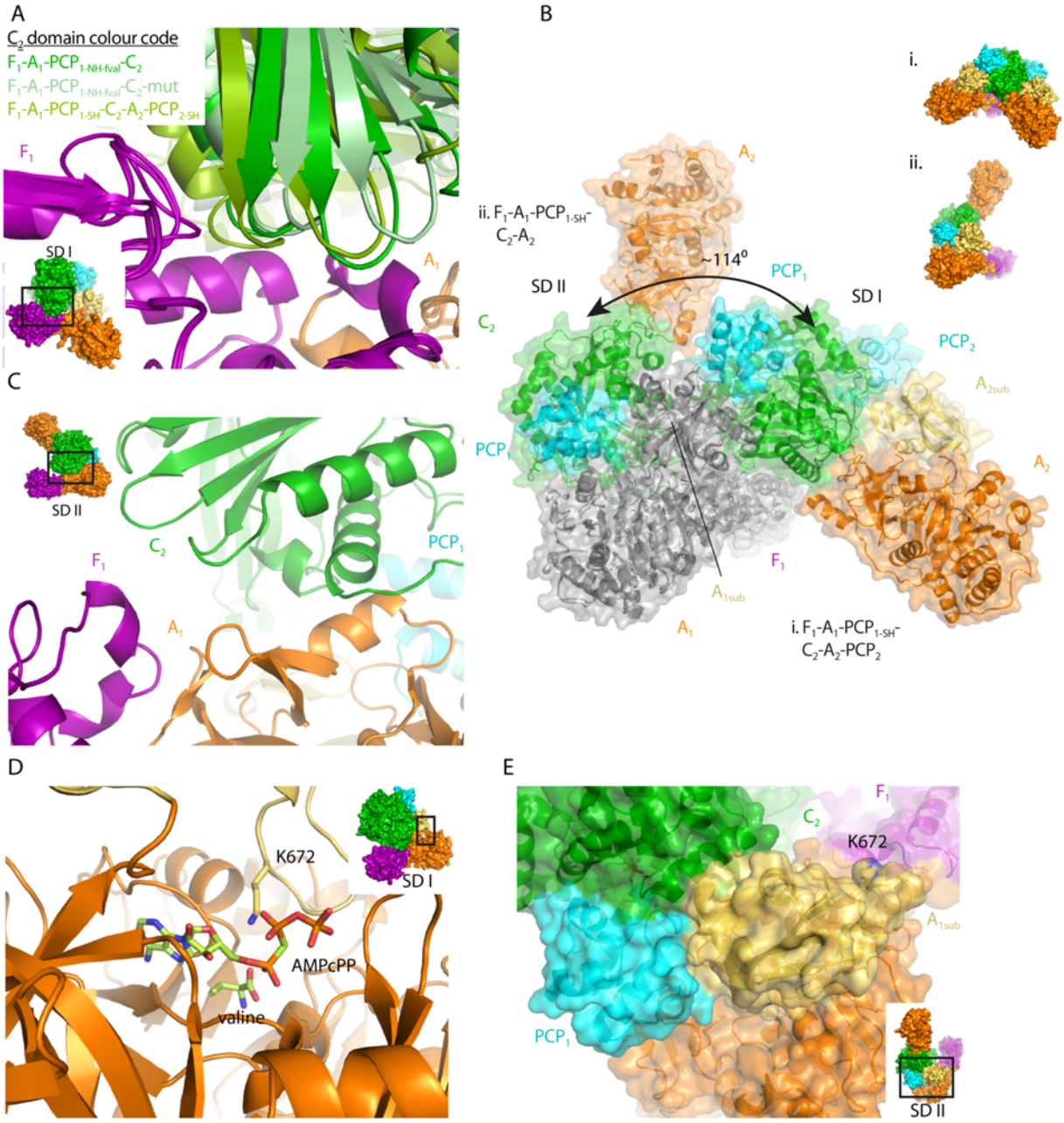


Figure 4.4 | Substrate donation in LgrA.
A, The N-lobe of C_2 contacts the entrance of the F_1 domain at varied positions. Inset view box, F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2-mut. **B**, Overlay of i. F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ and ii. F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2 structures aligned to F_1-A_{1core} emphasizes the different orientations of the elongation module. **C**, The N-lobe of the C_2 domain contacts the A_{1core} domain in F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2. **D**, The A_{1sub} domain is in the closed position in SD I. F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 structure shown. **E**, The A_{1sub} domain is in an extended conformation in SD II.

~5 Å between structures, causing a slight shift in the location of the PCP_1 domain. The A_{1core} and PCP_1 domains do not touch each other. In SD II, the A_{1sub} domain acts as a "linker domain (Gulick,

2016)" between the A_{1core} and PCP_1 domains to span the distance to the C_2 domain active site (Figure 4.4E). The A_{1core} and A_{1sub} domains form a new temporary platform and make extensive contacts with both the PCP_1 and C_2 domains. Accordingly, the PCP_1 domain buries a total of ~1215 $Å^2$ of surface area This conformation resembles an intermediate in the previously-described ~140 degree (Reger et al., 2008) transition of the A_{sub} between adenylation and thiolation states. The electron density maps for the A_2 domain are much weaker than for the rest of the protein, and several loops of the A_{2core}, as well as the entire A_{2sub} domain, could not be modelled. Flexibility in the C_2:A_2 interface is very likely contributing to this plasticity.

C domains have an overall 'V'-like shape where the angle between the N-lobe and C-lobe changes the degree of 'openness' of the domain. The C_2 domain is partially open compared to the CDA C_1 (Bloudoff et al., 2013) and SrfA-C C_7 domains (Tanovic et al., 2008), and does not change significantly between structures. In both SD I and SD II conformations, the PCP_1:C_2 interface is formed using the front face of the PCP_1 domain with the floor loop and surrounding helices of the C_2 domain C-lobe. No crystal contacts interfere with the PCP_1:C_2 interaction, and there is a slight rotation of the PCP_1 throughout the structures that causes a ~3.5 Å shift between equivalent atoms at the distal end of the helical bundle (Figure 4.5A). Accordingly, the amount of surface area buried between the PCP_1 and C_2 domain differs, ranging from ~591-668 $Å^2$. The modified serine (Ser729) is in overlapping positions and directs the PPE arm into the C_2 domain active site.

4.2.5 Substrate donation in the condensation state

Despite extensive biochemical and structural efforts to elucidate the catalytic mechanism of the C domain, the specificity and catalytic determinants of the C domain are under debate. The C domain contains a conserved catalytic motif, HHxxxDG, with condensation activity predominantly relying on the second histidine, henceforth referred to as the catalytic histidine. Originally proposed to act as a general base, it was recently suggested that the catalytic histidine plays a positioning role to mediate condensation. In the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2, strong electron density for the PPE arm is observed extending down the C domain tunnel into the active site (Supplemental Figure 4.3). This outlines the tunnel for the donor aminoacyl-PPE substrate to enter. The tunnel is wedged between strands C2β8 and C2β10 and capped by helix C2α4 (Figure

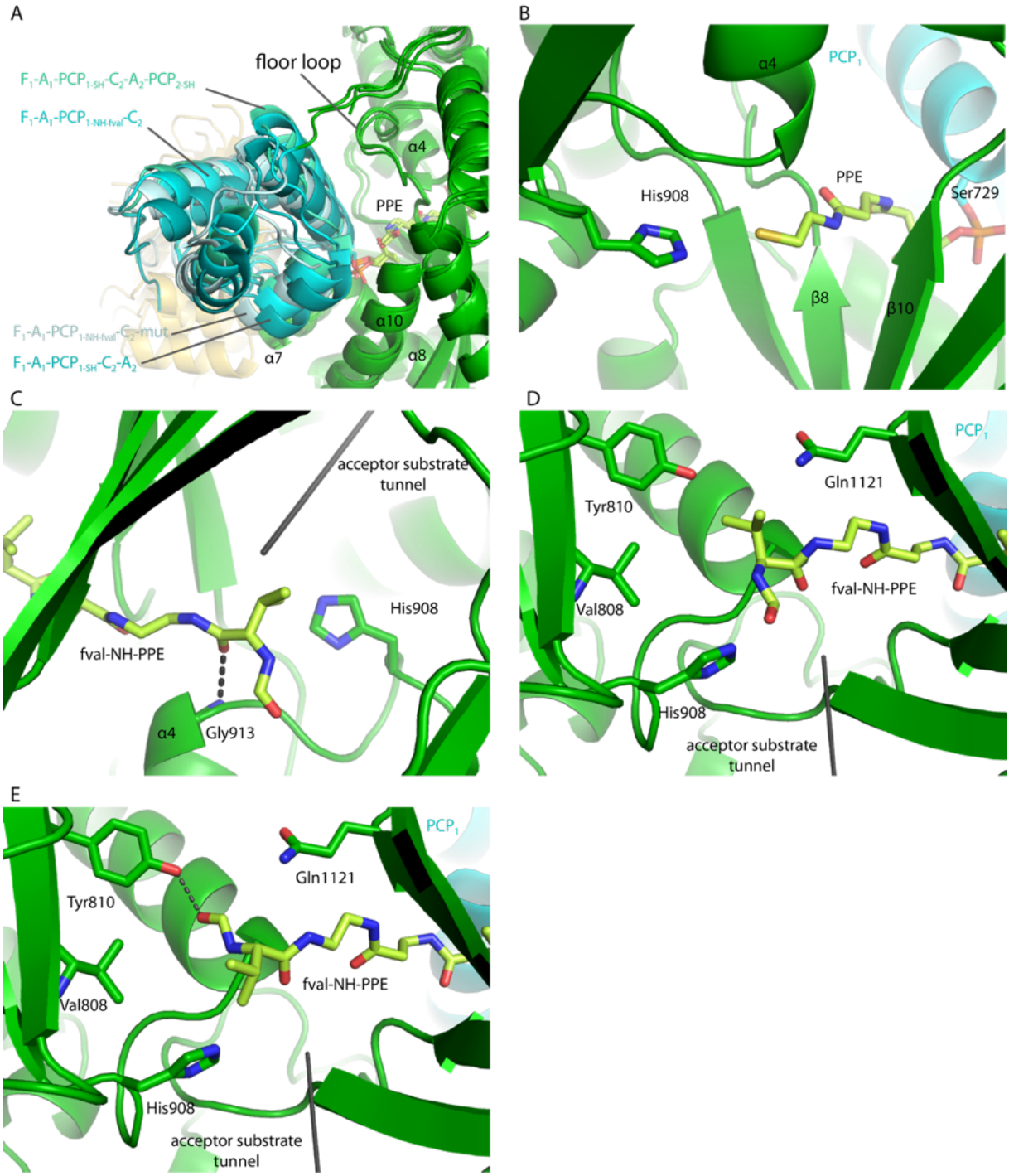


Figure 4.5 | Substrate donation to the condensation reaction.

A, Overlay of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$, F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$, F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut, F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ and F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 structures, and aligned to the C_2 domain. The PCP_1 is slightly rotated between structures, but still places Ser729 in similar positions to orient the PPE arm into the active site. C domain helices are labeled. **B**, The PPE arm extends into the C_2 domain active site between structural elements β8, β10 and α4. **C**, Catalytic motif glycine, Gly913, makes a hydrogen bond with the donor substrate carbonyl. **D**, **E**, Formyl-valine can be modelled in two different orientations. The estimated placement for the acceptor substrate is shown with a line.

4.5B, Supplemental Figure 4.4). The backbone nitrogen and oxygen of Asp1043 and Met1120, respectively, form hydrogen bonds with the PPE arm. The sidechain of Thr1013 hydrogen bonds with the phosphate group and is the only sidechain that contacts the PPE arm. The PPE arm points directly towards the catalytic histidine, His908, with the thiol group ~4 Å from the histidine sidechain. This would place formyl-valine within reaction competent proximity of both His908 and the acceptor substrate.

To visualize formyl-valine-PPE in the C domain active site, we initially synthesized formyl-valine-SH-CoA and modified F_1-A_1-PCP_1-C_2 using Sfp to produce F_1-A_1-$PCP_{1\text{-}SH\text{-}fVal}$-$C_2$. Resulting structures did not show electron density for formyl-valine, likely due to the high hydrolysis rate of the thioester linkage on a crystallographic time scale. Therefore, we synthesized the hydrolytically stable and isosteric analogue, formyl-valine-NH-CoA, and used it with Sfp to generate F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$ and F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$-mut proteins. Both F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$ and F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$-mut structures show strong electron density for the amino-PPE arm entering the C_2 domain donor substrate tunnel with weaker electron density for formyl-valine attached to the end of the amino-PPE arm (Supplemental Figures 4.3C, E). The amino-PPE arm follows the analogous trajectory as the thiol-PPE arm in the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 structure with amide and thiol groups occupying near equivalent positions in the C domain active site. The catalytic motif glycine, Gly913, is located at the proximal end of helix C_2-α4 and hydrogen bonds through its backbone nitrogen with the amide carbonyl tethering valine to the amino-PPE arm (Figure 4.5C). Together with the dipole moment of $C_2\alpha4$, the hydrogen bond helps to correctly orient the electrophilic α-carbon of formyl-valine for nucleophilic attack by the acceptor substrate. The unnatural geometry of the planar peptide bond is suboptimal, but nevertheless, it does approximate the native interaction between the thioester carbonyl and Gly913. This was previously predicted by Samel *et al* after observing a sulfate ion hydrogen bonded to the motif glycine, Gly229, in the presumed position of the donor thioester in the structure of the TycC PCP_5-C_6 didomain (Samel et al., 2007).

The electron density for formyl-valine is too weak to differentiate between the position of the valine sidechain and formyl group, and be modelled in two different conformations. In the first possibility, the valine side chain is directed toward the N-lobe of the C_2 domain (Figure 4.5D). The formyl group extends in front of His908 to partially occupy the acceptor substrate binding area.

This position of the formyl group would block the trajectory of the α-amino group of the acceptor glycyl-PPE, and would need to be rearranged for productive reaction. This is reminiscent of the situation in the large ribosomal subunit, which keeps the peptidyl-tRNA in a non-reactive conformation until the aminoacyl-tRNA binds (Schmeing et al., 2005). In the second, the positions of the valine sidechain and formyl group are switched, and the valine's carbonyl carbon is partially exposed for nucleophilic attack (Figure 4.5E). The formyl group is within hydrogen bonding distance with the sidechain hydroxyl of Tyr810, which could help position formyl-valine for condensation. The lack of clear electron density for the formyl group and valine sidechain indicates that the donor substrate has conformational flexibility within the active site. The presence of the acceptor substrate, glycyl-PPE, would likely help order the active site and cause the donor substrate to assume a reaction competent conformation for condensation.

4.2.7 PCP_1-C_2 linker limits possible elongation module positions

The linker between PCP_1 and C_2 domains covalently connects the elongation module to the initiation module. The length of the linker limits the possible relative positions of the elongation module. The LgrA PCP_1-C_2 domain linker is composed of residues 768-779 (12 residues), as defined by evolutionary covariance analyses using the webserver GREMLIN (Supplemental Figure 4.5A-C) and structure alignments (Morcos et al., 2011, Marks et al., 2011, Ovchinnikov et al., 2015). The distance between adjacent Cα atoms of 3.8 Å (Chakraborty et al., 2013) means the linker could maximally span ~49 Å, though this distance is likely an overestimate of the value. The linker would probably span a shorter distance because it would be entropically unfavourable for the linker to be fully extended. As well, in all of the LgrA C_2 domain-containing structures, regardless of the PCP_1 domain conformation, residues 773-778 are in an analogous (ordered) conformation (R.M.S.D. of 0.64 Å), which likely reduces the size of the practical linker (Supplemental Figure 4.5D). Similarly, the PCP-C linkers in the structures of TycC PCP-C (Samel et al., 2007), DhbF A-PCP-C:MLP (Tarry et al., 2017) and GrsA PCP-E (Chen et al., 2016) display a similar conformation where residues belonging to the PCP-C linker and beginning of the C or E domain wrap around C/Eα5 before inserting between C/Eα5 and C/Eβ3 (Supplemental Figure 4.4, Supplemental Figure 4.5E). This differs from structures that contain an N-terminal C domain (Supplemental Figure 4.5E),

which suggests the presence of the PCP domain causes the PCP-C linker to favour a conformation similar to LgrA and other PCP-C/E structures.

Table 4.1 | Distance between the PCP_1 domain C-terminus (Gln767) and the C_2 domain N-terminus (Glu780).

Structures were aligned to the A_{1core} domain and distances measured between Cα-Cα.

Catalytic conformation Gln767	Structure	Catalytic conformation Glu780	Structure	Distance (Å)
SD I	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	Thiol I	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	89
SD I	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	Thiol II	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	73
Thiol I	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	SD I	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	100
Thiol II	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	SD I	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	101
Thiol I	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	SD II	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	77
Thiol II	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	SD II	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	77
Formylation	F_1-A_1-$PCP_{1\text{-}SH}$	SD I	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	45
Formylation	F_1-A_1-$PCP_{1\text{-}SH}$	SD II	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	83
SD II	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	Thiol I	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	58
SD II	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	Thiol II	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	70

The possible positions of the N-terminal residue of the C domain (Glu780) can be visualized by drawing a sphere of radius 49 Å (maximum PCP_1-C_2 linker length) centered on the alpha carbon of Gln767. As an example, Figure 4.6A shows how Glu780 is within this sphere for the Thiol I and Thiol II conformations. Similarly, by superimposing structures of consecutive catalytic states with corresponding spheres drawn on Gln767, the overlapping area between spheres is representative of the possible positions of the C_2 domain (Glu780) that would allow for successive steps of the initiation catalytic cycle without simultaneous movement of the elongation module. For instance, there is a possible position(s) for Glu780 that would permit both thiolation and formylation states (Figure 4.6B). However, the A_{1sub} domain partially occupies this area and must have sufficient room to rotate from thiolation to formylation conformations, further limiting the available positions. The PCP_1-C_2 linker would also have to be almost fully extended as the overlap occurs at the peripheral of each sphere. In the structure of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut, the PCP_1-C_2 linker is elongated and illustrates a more probable extended length of 38 Å (Figure 4.6C). Using the same approach but with spheres of radius 38 Å, the overlapping area becomes almost negligible (Figure 4.6D).

Although there may be a position that would allow the elongation module to remain in a given location for other consecutive reactions of the initiation synthetic cycle, it is unlikely that there is

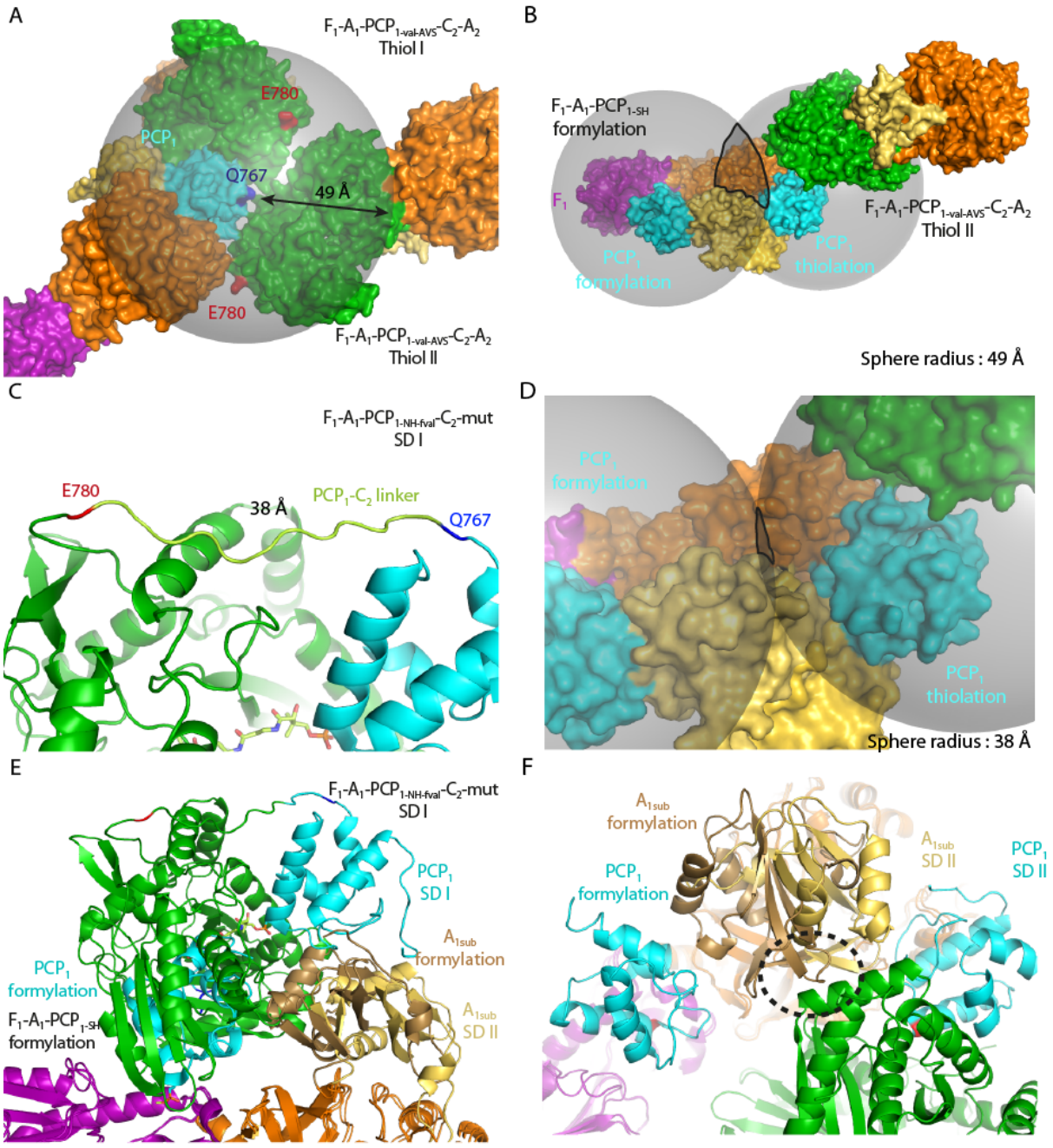


Figure 4.6 | PCP_1-C_2 linker limits possible elongation module positions.
Caption on following page.

Figure 4.6 | PCP-C linker limits possible elongation module positions.
A, Overlay of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 structures with a sphere of radius 49 Å centered on the C-terminus of the PCP_1 domain, Q767 (blue). The N-terminus of the elongation module, Q780 (red), is within the allowed PCP_1-C_2 linker distance. **B**, Overlay of Thiol II conformation with formylation state, aligned to F_1-A_{1core} domains. The overlapping area, outlined in black, is the space in which Q780 could occupy and permit both the thiolation and formylation reactions without movement of the elongation module. **C**, The distance between Q767 and E780 in the structure of F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2-mut represents a reasonable extended linker length. **D**, The area in which the initiation module can undergo both thiolation and formylation reactions without concurrent movement of the elongation module is almost non-existent when the PCP-C linker sphere is redrawn with radius 38 Å. **E**, The position of the PCP_1 during formylation clashes with the C_2 domain in SD I. **F**, The A_{1sub} in the formylation state clashes with the C_2 domain in the SD II conformation.

a single position of the elongation module that would permit both thiolation and formylation states of the initiation module. Despite using several chemical biology approaches involving custom formylation inhibitors, we were unable to obtain a structure of the formylation state of the initiation module in constructs that included domains of the elongation module. There will certainly be multiple positions of the elongation module that are compatible with the formylation state, with Glu780 positioned within a PCP_1-C_2 linker defined area.

The transition from the formylation state and to either SD I or SD II conformations would require nontrivial conformational changes in each module. The PCP_1 domain in the formylation state is completely incompatible with the position of the C_2 domain in SD I (Figure 4.6E). Likewise, while the C_2 domain does not obstruct the F_1 domain in SD II, the A_{1sub} domain in the formylation state clashes with the position of the C_2 domain in SD II (Figure 4.6F). Concerted conformational changes involving large non-linear movements are needed between each module to transition between the formylation state and either substrate donation conformation

To summarize, the conformational changes needed to complete the initiation module synthetic cycle necessitate that the elongation module simultaneously move with the PCP_1 domain because of constraints imposed by the length of the PCP_1-C_2 linker. The elongation module movements observed in the LgrA crystal structures cause equivalent atoms to move up to ~224 Å apart (Figure 4.7). While these are plausible conformations that show relevant catalytic states, the movements are not required to be as substantial as the movements presented here. There may

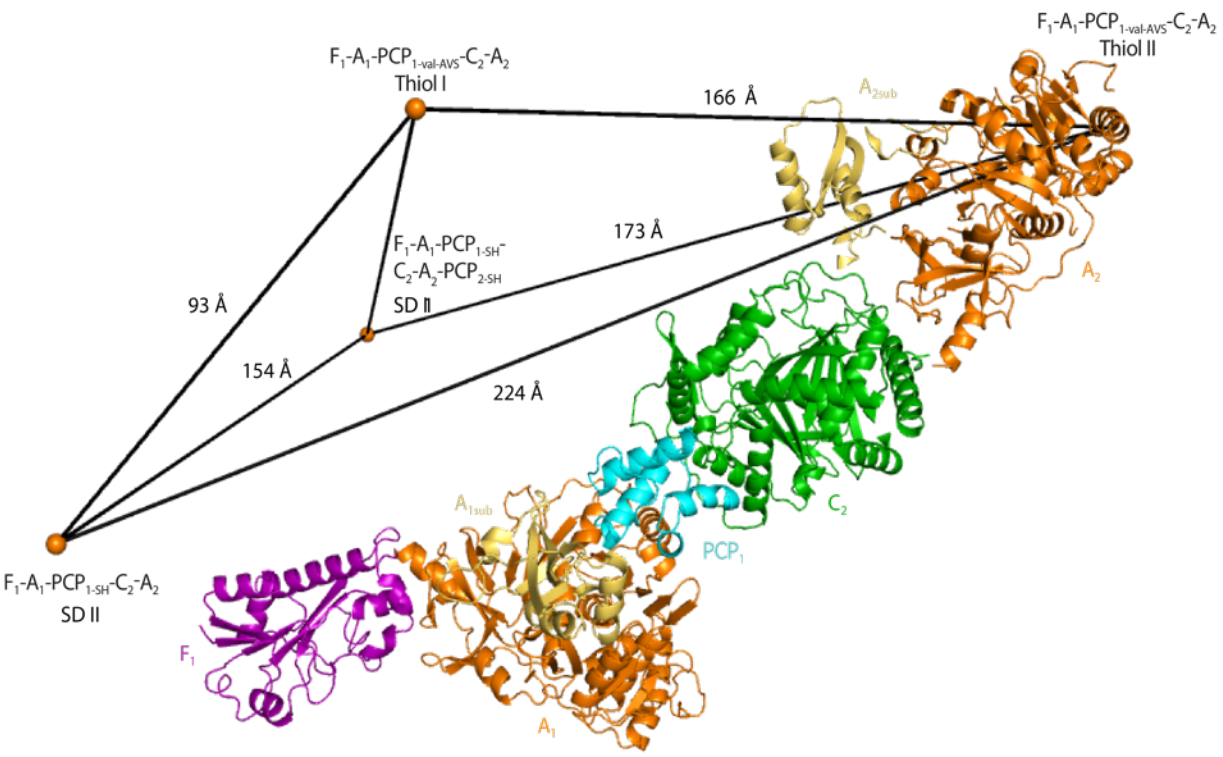


Figure 4.7 | Elongation module movements in LgrA.
Distances between equivalent positions of D1236 in the structures of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2, F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ and F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2.

be orientations of the elongation module that would allow two sequential catalytic states, but there does not appear to be a single conformation that would allow all of the catalytically-relevant states of the LgrA initiation module. Furthermore, the crystal packing of F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ is informative in this context. As mentioned above, there are two molecules in the asymmetric unit of this crystal. The entire protein was modelled in one of the two molecules, but only the initiation module could be modelled in the other. Its lack of electron density indicates that this elongation module is in multiple (and perhaps a continuum of different) positions. Indeed, superimposition of the initiation module of any of the four observed F_1-A_1-PCP_1-C_2-A_2(-PCP_2) structures onto this initiation module cause the elongation module to overlap with symmetry-related proteins, so this partially disordered F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ is certain to be in multiple (unobserved) novel conformations. Hence, the four vastly different orientations of the LgrA elongation module, together with evidence provided by the conformational flexibility exhibited in dimodular DbhF (Tarry et al., 2017) and the Te domain of the EntF termination module (Drake et

al., 2016), strongly supports an NRPS architectural model where module:module interactions are transient and dynamic.

4.3 Discussion

The series of structures we present here add to an impressive body of structures of smaller NRPS fragments (Weissman, 2015, Reimer et al., 2018), and comparison to this body further expands our knowledge of the workings of NRPSs.

4.3.1 Flexibility in the elongation module

This work reinforces the observations on modules from the Marahiel and Gulick labs, that the shape of the C and A domains, and the large C:A interface defines the general shape of an elongation module, but that there is some flexibility about the C:A interface. This flexibility alters the relative orientation between domains, and in some cases can limit the available positions of the A_{sub} domain. In one structure of EntF (C-A-$PCP_{ser\text{-}AVS}$-Te), the C:A interface angle places part of the C in a position which would prevent the A_{sub} domain to adopt its closed position, and so is incompatible with adenylating function (Drake et al., 2016). A subsequent structure of EntF revealed a ~15° shift of the C domain resolves the potential conflict and allows the closed position of the A domain (Miller and Gulick, 2016). Likewise, the positions of the C domains in the SrfA-C, AB3403, EntF and LgrA structures clash with the position of the A_{sub} domain in the open (substrate binding) state (Figure 4.8A). The open conformation has only been observed in the context of the LgrA initiation module (PDB 5ES5, chain A) (Reimer et al., 2016a) and related luciferase structures (Conti et al., 1996) where the motion of the A_{sub} domain is not restricted by a large upstream domain. (The F domain is smaller and in a "lower" relative position to the A domain in the initiation module than a C domain in an elongation module (Reimer et al., 2016a).) It is possible that the C:A interface splays to permit the A_{sub} to fully rotate open, but given the myriad of partially-open A_{sub} domain positions observed in adenylating enzymes (Tanovic et al., 2008, Lee et al., 2010, Hisanaga et al., 2004), it is more likely that the A_{sub} domain is not required to be in one particular, fully-open, conformation to allow binding of ATP and cognate amino acid.

4.3.2 Substrate donation, the functional link between modules

The key catalytic event of peptide synthesis, peptide bond formation, is the only event in nonribosomal peptide synthesis that necessitates coordination between modules. Condensation

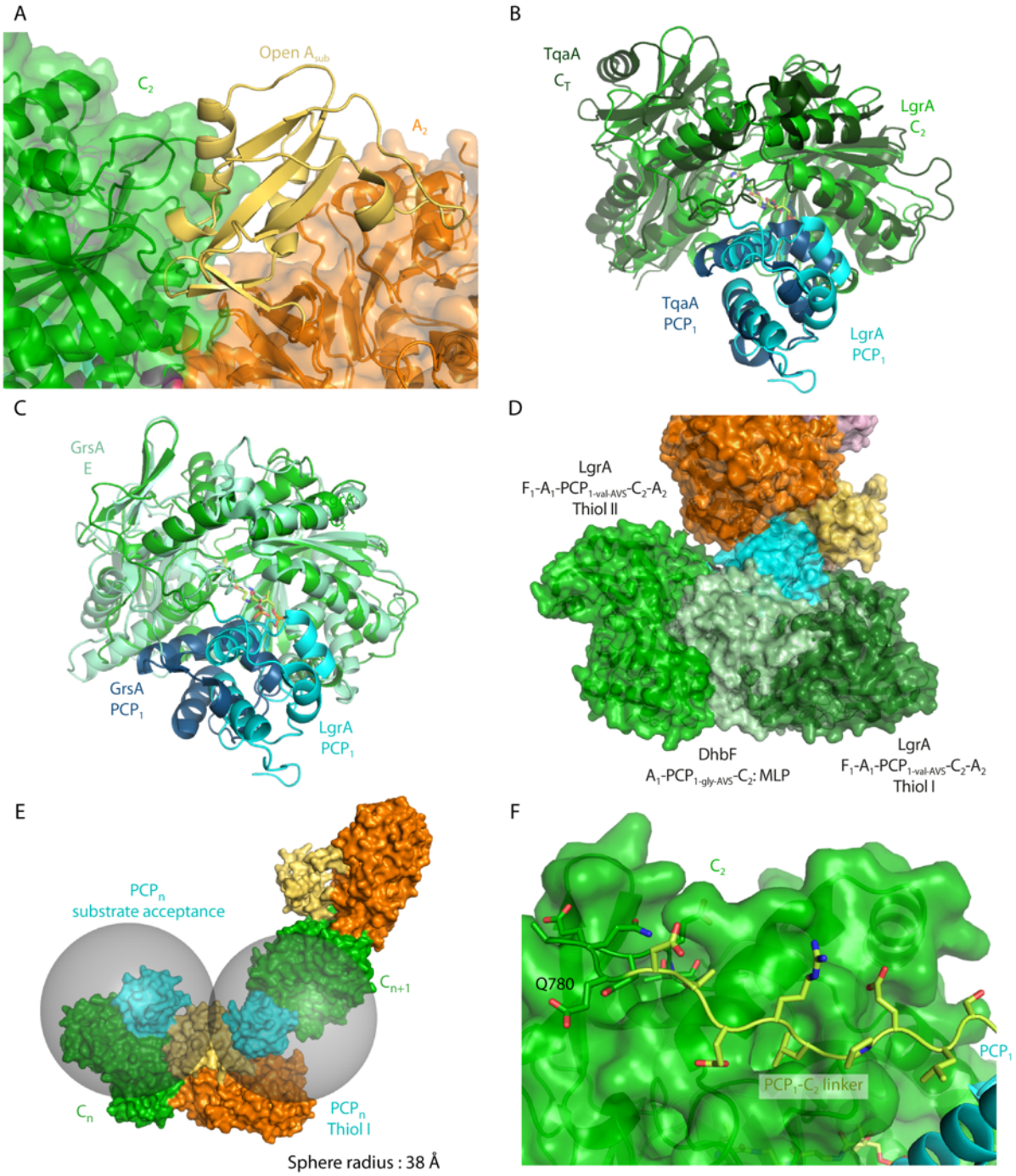


Figure 4.8 | LgrA compared to other NRPSs.
Caption on following page.

Figure 4.8 | LgrA compared to other NRPSs.
A, Overlay of F_1-A_1-PCP_1 (PDB 5ES5 chain A, PCP disordered) (Reimer et al., 2016a) with F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2. The open position of the A_{sub} would clash with the C domain. Overlay of LgrA F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ with (**B**) TqaA PCP-C_T didomain (Zhang et al., 2016) and (**C**) GrsA PCP-E didomain structures (Chen et al., 2016). **D**, The position of the C domain in the structure of DhbF, A_1-$PCP_{1\text{-}gly\text{-}AVS}$-$C_2$:MLP (Tarry et al., 2017), is relatively positioned between the C domains of F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$. **E**, Overlay of the LgrA elongation module from F_1-A_1-$PCP_{1\text{-}}$ -C_2-A_2-$PCP_{2\text{-}SH}$ onto Thiol I conformation. The overlap between PCP-C linker spheres is limited, indicating the elongation module most likely has to move for the PCP to transition from thiolation to substrate acceptance states. **F**, The PCP_1-C_2 domain makes extensive contact with the C_2 domain.

requires PCP_n to make a trans-module interaction with C_{n+1} , here termed a substrate donation conformation, which must be coincident with a C_{n+1}:PCP_{n+1} substrate acceptance conformation. The substrate donation conformation was observed in >4 different crystal packing environments (including unpublished structures), and although the C domain's donor site is rather broad, the PCP domain is observed in quite similar positions related by a modest rotation. The PCP:C interaction has been approximated previously in the structures of TqaA PCP-C_T (Figure 4.8B) (Zhang et al., 2016) and GrsA PCP-E (Chen et al., 2016) (Figure 4.8C). The TqaA C_T domain is in a more open conformation than the LgrA C_2 domain and uses the analogous portion of the donor site to interact with the PCP domain, with a relatively small difference in orientation of the PCP, described by a ~18° rotation and ~4 Å translation. Conversely, the GrsA PCP domain binds to E domain in a very different manner. It sits at the opposite side of the broad binding site, and contacts the opposing lobe (N-lobe) of the E domain, with a position related to that of LgrA PCP_1 by a ~40° rotation and ~11 Å translation. It is possible that this GrsA PCP domain position is caused by the crystal packing (the PCP domain makes several crystal contacts with symmetry-related molecules, and these symmetry related molecules block the position of that PCP domains occupy in LgrA and TqaA), but it is also possible that this is a bonafide PCP:E/C interaction. GrsA, TqaA and LgrA all direct their PPE arms down the donor site tunnel and position the thiol sulphur within ~3 Å of analogous positions of the C or E domain active site. It is likely therefore that there does not exist one single productive donor conformation, but multiple positions of the PCP domain exist at the donor site that can all productively deliver peptidyl-PPE to the C domain. Indeed, in the C domain of SrfA-C, the "bridge loop" section of the donor site is longer than that of LgrA by 2

residues, which would necessitate this loop or the PCP domain to adopt a different position that that seen in LgrA. Although PCP_n:C_{n+1} interactions are trans-modular, the linkers on either side of the PCP domain with the associated great mobility of the PCP domain mean that this interaction does not dictate the overall module:module conformation and super modular architecture.

4.3.3 Flexibility between modules

The present LgrA structure are perhaps most informative about larger scale, super modular NRPS architecture. As mentioned, the previous best source of information had been our lab's study on DhbF. In the DhbF A_1-$PCP_{1\text{-val-AVS}}$-C_2:MLP structure, the C domain is ordered by interactions the "back face" of the PCP domain. Its position relative to A_1 is between those observed for the two C_2 domains in F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 (Tarry et al., 2017) (Figure 4.8C). Accompanying low-resolution negative stain EM reconstructions of dimodular DhbF (C_1-A_1-$PCP_{1\text{-gly-AVS}}$-C_2-A_2-$PCP_{2\text{-thr-AVS}}$:MLP) appeared to indicate large conformational flexibility between modules. During canonical elongation, after the thiolation reaction, aa_n-PCP_n (aa, amino acid) binds C_n and accepts the nascent peptide from peptidyl-PCP_{n-1}, and then proceeds to C_{n+1}. As demonstrated above, it is overwhelmingly likely that this is associated with changes in the $module_1$:$module_2$ orientation. It is also very likely in the case of adjacent elongation modules (such as in DhbF and most dimodule segments of NRPSs). As illustrated in Figure 8E with a PCP_n-C_{n+1} linker of typical size, $module_{n+1}$ will likely have to move with the PCP_n domain, and break any transient inter-modular contacts formed in the thiolation state. The lack of conserved $module_1$:$module_2$ contacts in the DhbF study is thus not surprising. In all, the available data strongly suggests there exists a continuum of possible $module_n$:$module_{n+1}$ orientations in NRPSs. Some of these elongation module:elongation module conformations will be the same as we observe in the LgrA initiation module:elongation module case, and some will not. For example, the SD I conformation seen here is not possible for 2 elongation modules (because the F domain and the C domain do not occupy the precisely analogous space in a module, and C_n and C_{n+1} would clash in SD I), whereas SD II is compatible with two or more elongation modules.

The continuum of conformations need not be evenly populated, and NRPSs may have developed ways to favour more productive conformations. For example, the PCP_1-C_2 linker makes

contact with helix $C_2\alpha5$, the bridge loop (residues 1043-1057) and loop $C_2\alpha4$-$C_2\alpha5$ on the N-lobe side of the broad donor site in LgrA (PCP_1 binds the C-lobe) (Figure 8F, Supplemental Figure 5D). This may function to increase the population of conformations that position the C_{n+1} domain's donor site towards the PCP_n. In GrsA, the PCP-E linker is shifted slightly from the position of the PCP_1-C_2 linker in LgrA (Supplemental 5E), but still makes similar contacts with the bridge loop and loop Eα4-Eα5. Mutations in the Eα4-Eα5 linker caused an observable effect on the epimerization ability of GrsA and may have disrupted the interaction between the E domain and PCP-E linker (Chen et al., 2016). In all of the LgrA dimodular structures, and the cross-module structure of DhbF (Tarry et al., 2017), the donor side of the C domain is facing the proceeding module, which could help the NRPS to orient for substrate donation.

4.3.4 Non-canonical NRPSs

NRPSs have proven highly adaptable to deviations within their standard synthetic cycle. Numerous tailoring domains have been co-opted and successfully incorporated into the modular framework without disruption to the synthetic process. Only two structures of *in cis* tailoring domains exist, the F domain of LgrA (Reimer et al., 2016a) and a methyltransferase that interrupts the A_{sub} domain in TioS (Mori et al., 2018). LgrA and TioS use different strategies for integrating their respective tailoring domains, and additional creative tactics will certainly be used by other NRPSs to adapt additional tailoring functions in their own synthetic cycles. NRPSs have also been found with altered module compositions, such as the heterobactin HtbG synthetase where the second module has the domain order, C-PCP-A (Bosello et al., 2013), or the beauvericin fungal NRPS contains a module with tandem PCP domains, C-A-MT-PCP-PCP (Xu et al., 2008). Further, in hybrid polyketide synthases (PKS)-NRPSs either the NRPS or PKS module or both modules at the transition may require tweaking to allow productive matching of the NRPS and PKS synthetic cycles. The lack of a single, rigid supermodular architecture in NRPSs may facilitate incorporation of these non-canonical domains and catalytic events into the synthetic cycles.

Conversely, NRPS systems may take advantage of their extreme flexibility for unusual synthetic schemes. As an increasing number of biosynthetic clusters are discovered and characterized, we are only starting to understand how versatile, and sometimes bizarre, the NRPS

synthetic cycle can be. For example, the thalassospiramide A synthetase found in α-proteobacterium *Tistrella* sp. is an unorthodox six module hybrid NRPS-PKS (Ross et al., 2013). The synthetase lacks A domains in modules 1 and 5, and the PKS module 4 is missing its substrate activating acyltransferase (AT) domain. The proposed catalytic mechanism has two unique features: The A domain in module 2 acts *in trans* within its own polypeptide to activate and load substrates onto the PCP domains of modules 1, 2 and 5; and PCP_4 is first loaded using the fatty acid AT, FabD, and then reiterates the reactions of modules 2-4 by returning the peptide to module 2 for an additional round of catalysis. *In trans* A domain activation between two separate synthetase subunits has been observed before, such as in yersiniabactin synthetase (Suo et al., 2001), where only productive protein-protein interactions are required between the A domain-containing subunit and PCP domain-containing subunit. However, intra-synthetase *in trans* reactions would require the NRPS have a dynamic modular arrangement for the A_2 domain to productively interact with the PCP domains of module 1 and 5. Likewise, transient non-adjacent modular interactions are needed for the PCP_4 domain to return the peptide to module 2 for iterative synthesis. Another example of an abnormal synthetic cycle that would require super-modular flexibility is the module-skipping observed in myxochromides S1-3 synthetase. Here a proline-activating module between modules 3 and 5 is ignored during synthesis, with PCP_3 donating the peptide directly to module 5 (Wenzel et al., 2005).

While the elongation module movements illustrated in the LgrA crystal structures exceed the minimum requirements, exceptional module movements will certainly be required to facilitate the unusual domain:domain interactions mentioned in the non-canonical NRPSs discussed above. Inter-domain and inter-module linkers will play key roles in enabling these movements. Indeed, the dynamic and flexible architecture exhibited here, and in the study of DhbF (Tarry et al., 2017), enables these elegant megaenzymes to orchestrate synthesis of their incredibly diverse and important products.

4.4 Methods

4.4.1 Cloning of the LgrA constructs

Genomic DNA was isolated from *Brevibacillus parabrevis* ATCC 8185 (Cedarlane Laboratories) using a GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich). All gene constructs were amplified by PCR from the *lgrA* gene using the following primers (PCP domain abbreviated using its alternative abbreviation, T for thiolation domain):

FATC_fwd: 5'-TGACTACCATGGGGAGAATACTATTCCTAACAACATTTATGAGC-3',

FATC_rev: 5'-TCTCTGCGGCCGCTACCAGTTCAAACCTTTTCACC-3',

FATCA_fwd: 5'-TGACTACCATGGGGAGAATACTATTCCTAACAACATTTATGAGC-3',

FATCA_rev: 5'-CGTTGAGCGGCCGCCACACTGCCGTCCGGTTCT-3',

$FATCAT_{(short)}$_fwd: 5'-TGACTACCATGGGGAGAATACTATTCCTAACAACATTTATGAGC-3',

$FATCAT_{(short)}$_rev: 5'-CGTTGAGCGGCCGCGGACGTGACGAATGGGGCAA-3'.

Domain boundaries for each construct were designed using sequence alignments, known structure alignments and the GREMLIN server (Ovchinnikov et al., 2014). PCR products for F_1-A_1-PCP_{1}-C_2, F_1-A_1-PCP_{1}-C_2-A_2 and F_1-A_1-PCP_{1}-C_2-A_2-$PCP_{2(short)}$ were digested using NcoI and NotI (New England Biolabs) and ligated into a pET21-derived vector containing an N-terminal tobacco etch virus (TEV) cleavable octa-histidine tag and a C-terminal TEV cleavable calmodulin binding peptide (CBP) tag. After initial crystallization trials failed, the construct, F_1-A_1-PCP_{1}-C_2-A_2-$PCP_{2(short)}$, was elongated by inserting an additional 6 residues at the C-terminal using site directed mutagenesis and primers: $FATCAT_2$_fwd: 5'-

CCCATTCGTCACGTCCGAGCAGGTCGTCATCGAAGCGGCCGCAGAGA-3',

$FATCAT_2$_rev: 5'-TCTCTGCGGCCGCTTCGATGACGACCTGCTCGGACGTGACGAATGGG -3.'

This construct is denoted as F_1-A_1-PCP_{1}-C_2-A_2-PCP_2, and was successfully crystallized with structure determination. F_1-A_1-PCP_1-C_2-mut was originally constructed for use in alkylation experiments not presented here (See also Supplemental Information). Four mutations, C31S, C191S, C318S and R792C, were introduced into F_1-A_1-PCP_{1}-C_2 by site directed mutagenesis using primers: FATC_R792C_for: 5'-CTTCCGCCGTCGAGAAGTGCATGTACATCATCCAGCAGCAAG-3',

FATC_R792C_rev: 5'-CTTGCTGCTGGATGATGTACATGCACTTCTCGACGGCGGAAG-3',

FATC_C318S_fwd: 5'-AAGCGAACGGTGCCAGCGACATCATCGACG-3',

FATC_C318S_rev: 5'-CGTCGATGATGTCGCTGGCACCGTTCGCTT-3',

FATC_C191S_for: 5'-GCATTAAAGCGGCTTAGTGCAGAGCCAAAAC-3',

FATC_C191S_rev: 5'-GTTTTGGCTCTGCACTAAGCCGCTTTAATGC-3',

FATC_C31S_for: 5'-ATTACACCACGAAGTCGTCATCTCTCAGGAAAAAGTGCACGCG-3',

FATC_C31S_rev: 5'-CGCGTGCACTTTTTCCTGAGAGATGACGACTTCGTGGTGTAAT-3'.

4.4.2 Expression and purification of LgrA proteins

Apo F_1-A_1-PCP_1-C_2, F_1-A_1-PCP_1-C_2-mut, F_1-A_1-PCP_1-C_2-A_2, and F_1-A_1-PCP_1-C_2-A_2-PCP_2 proteins were expressed in *Escherichia coli* BL21 (DE3) *entD*- cells (Chalut et al., 2006). Holo F_1-A_1-PCP_1-C_2-A_2 and holo F_1-A_1-PCP_1-C_2-A_2-PCP_2 were produced in *Escherichia coli* BL21 (DE3) Bap1 cells (Pfeifer et al., 2001). Cultures were induced using a 1:100 dilution of overnight culture in 1 L of lysogeny broth (LB) medium supplemented with 350 ug ml^{-1} kanamycin, and grown at 37 °C to an optical density (OD_{600}) of 0.6. Protein expression was induced at 16 °C using 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and grown for 18 h. Cells were harvested by centrifugation at 4 °C, and the cell pellets were resuspended in buffer A (2 mM imidazole, 150 mM NaCl, 2 mM phenylmethanesulfonyl fluoride (PMSF), 2 mM $CaCl_2$, 2 mM β-mercaptoethanol (β-me), 25 mM Tris-HCl (pH 7.5)). Cells were lysed by sonication, and the lysate was clarified by centrifugation at 20 000x*g* for 30 min at 4 °C. The supernatant was loaded onto a 30 mL calmodulin sepharose 4B column (GE Healthcare) equilibrated in buffer A. The column was washed with buffer A, and protein was eluted with elution buffer (2 mM EGTA, 150 mM NaCl, 2 mM PMSF, 2 mM β-me, 25 mM Tris-HCl (pH 7.5)). The elution peak was collected and loaded onto pre-equilibrated (buffer A) 5 ml HiTrap IMAC FF column (GE Healthcare) charged with Ni^{2+}. The column was washed with 20 mM imidazole and the protein was eluted using 250 mM imidazole. LgrA-containing fractions were pooled. Protein was dialyzed with 1:10 mg protein:TEV overnight in buffer A to remove affinity tags and remove imidazole. Protein was passed back over the IMAC and calmodulin affinity columns to remove uncleaved protein, and the corresponding flow-throughs were collected. Protein was applied to a MonoQ 10/100 (GE Healthcare) equilibrated in buffer Q0 (2 mM β-me, 25 mM Tris-HCl (pH 7.5)), washed with 150 mM NaCl, and eluted using a 150-600 mM NaCl gradient over 80 mL. Protein was pooled, concentrated and loaded onto a final HiLoad 16/60 Superdex S200 column (GE Healthcare) equilibrated in size exclusion (SEC) buffer (150 mM NaCl,

2 mM β-me, 25 mM Tris (pH 7.5)). Purity was assayed by SDS-PAGE and pure fractions were pooled, concentrated to 5 mg ml^{-1} and flash-frozen in liquid nitrogen for storage at -80 °C.

4.4.3 Substrate syntheses

Zamboni Chemical Solutions was commissioned to synthesize both the valine and glycine adenosine vinylsulfomamide (AVS) inhibitors, valine-adenylate analogues (aided in initial crystallization optimizations) and formyl-valine-*N*-hydroxysuccinimide ester (fval-NHS). Amino-coenzyme A (NH-CoA) was prepared as previously published (Reimer et al., 2016a). Formyl-valine-amino-CoA (fval-NH-CoA) was synthesized by coupling 1 molar equivalent of NH-CoA with 8 molar equivalents of fval-NHS in *N,N*-dimethylformamide (DMF, Sigma-Aldrich) with 4 molar equivalents of *N,N*-diisopropylethylamine (DIPEA, Sigma-Aldrich) overnight at 23 °C with stirring. Fval-NH-CoA was purified using the previously described C18 HPLC protocol (Reimer et al., 2016a). Synthesis was monitored by mass spectrometry and NMR.

4.4.4 Charging the PCP domain with phosphopantetheinylates

The PCP domains of apo F_1-A_1-PCP_1-C_2, F_1-A_1-PCP_1-C_2-mut, F_1-A_1-PCP_1-C_2-A_2 and F_1-A_1-PCP_1-C_2-A_2-PCP_2 were loaded with either coenzyme A (CoA) or formyl-valine-NH-CoA by incubating 40 uM apo protein with 10 uM Sfp, 25 mM Tris (pH 7.0), 10 mM $MgCl_2$ and 0.25 mM formyl-valine-NH-CoA or CoA for a minimum of 4 hr at 23 °C. Reaction components were removed by passing the reaction mixture over a Superdex S200 10/300 (GE Healthcare) equilibrated in SEC buffer (150 mM NaCl, 2 mM β-me, 25 mM Tris (pH 7.5)).

4.4.5 Modification with valinyl adenosine vinylsulfonamide inhibitors

Apo F_1-A_1-PCP_1-C_2-A_2 was incubated with 10 uM Sfp, 10 mM Tris (pH 7.0), 10 mM $MgCl_2$, 0.25 mM CoA and 0.25 mM valinyl-AVS for a minimum of 4 hr at 23 °C. Reaction components were removed by passing the reaction mix over a Superdex S200 10/300 (GE Healthcare) equilibrated in SEC buffer (150 mM NaCl, 2 mM β-me, 25 mM Tris (pH 7.5)).

4.4.6 Crystallography

Final crystallization conditions were optimized in 24-well sitting drop plates using a total 4 μL drop made of 2 μL protein solution and 2 μL reservoir solution, and equilibrated against a 500 μL reservoir volume, unless otherwise stated. F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ (2.5 mg ml^{-1}) was co-crystallized with 2.5 mM AMPcPP, 2.5 mM valine and 2 mM N5-fTHF in a 2.5 μL drop containing 2.5 mM $MgCl_2$, 15% PEG 4000, 100 mM Tris pH 7.7, 25 mM Tris pH 7.5 and 100 mM NaCl with a 500 μL reservoir containing 2.5 mM $MgCl_2$, 18% PEG 4000, 0.1 mM Tris pH 7.7 and 150 mM NaCl. Drops were streak seeded using a seed stock of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ crystals and incubated at 22 °C. Crystals grew in space group $P2_1\,2_1\,2_1$. F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut (3.1 mg ml^{-1}) was co-crystallized with 5 mM AMPcPP using a precipitant solution of 26 % PEG 4000 and 0.1 M Tris (pH 8.1) at 22 °C into space group P1. Endogenous valine was found in the resulting crystal structure. F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ (3.0 mg ml^{-1}) was crystallized using a precipitant solution of 0.25 M NaF and 3.1 M sodium formate (pH 6.8) at 4 °C into space group $P2_1\,2_1\,2_1$. Holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 (3.4 mg ml^{-1}) was crystallized using a precipitant solution of 0.2 M Na/K phosphate, 22% PEG 3350 and 0.1 M bisTris propane (pH 7.2) at 4 °C in space group $P2_12_12_1$. Holo F_1-A_1-PCP_1-C_2-A_2-PCP_2 (3.5 mg ml^{-1}) was crystallized using a precipitant solution of 0.25 M Na/K phosphate, 21% PEG 3350 and 0.1 M bisTris propane (pH 7.5) at 4 °C in space group $C2\,2\,2_1$.

Crystals were cryo-protected with either mother liquor supplemented with 10% glycerol or 10% PEG400 and any ligands included in co-crystallization, looped and flash-cooled in liquid nitrogen. Diffraction datasets for all proteins, except F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$, were collected at 200 K using the 08ID-1 beamline of the CMCF at the Canadian Light Source (λ = 0.979 Å) in Saskatoon, Canada. F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ was collected at 200 K using the 24ID-C beamline of the NE-CAT at the Advanced Photo Source in Argonne, Illinois, USA. The program iMOSFLM (Leslie and Powell, 2007) was used for indexing and integrated F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$, F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$, and holo F_1-A_1-PCP_1-C_2-A_2-PCP_2 datasets, while F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut and holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 datasets were indexed and scaled using DIALS (Winter et al., 2018). All datasets were scaled using the program AIMLESS (Evans and Murshudov, 2013).

The structure of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$ was solved using consecutive rounds of molecular replacement (MR) in the program Phaser (McCoy et al., 2007). The A_{1core} domain (residues 200-

585, PDB 5ES5) (Reimer et al., 2016a) from the LgrA initiation module structure was used as an initial search model and yielded a partial model. The A_{1core} solution was fixed and MR was executed using the F domain (residues 1-180, PDB 5ES5 (Reimer et al., 2016a)) as a subsequent search model. Using the resulting solution, numerous MR searches using different C domain structures were attempted without success. A favourable MR solution was found when the N-terminal lobe of TycC C domain (PDB 2JGP) (Samel et al., 2007) was used as a search model. Electron density maps were improved by iterative building in the program COOT (Emsley et al., 2010) and refinement in the program Phenix (Adams et al., 2002). Subsequent maps showed electron density for the C-terminal lobe of the C_2 domain and the PCP_1 domain.

The structure of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 was solved using a similar iterative MR strategy. An initial search with the structure of F-A-PCP (PDB 5ES8) surprisingly did not yield a convincing solution. The C_2 domain from F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 was then used as a search model and resulted in a good solution. The solution containing both C_2 domains was fixed and the initiation module structure was again used as a search model. Next, A_1 domain was sculpted to resemble the A_2 domain. The sculpted A_{2core} was searched for first, and then the sculpted A_{2sub} was manually placed, followed by rigid body refinement.

The structure of holo F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2 was solved using an iterative MR strategy. The didomain, F-A_{1core} (PDB 5ES5) (Reimer et al., 2016a), was used as an initial search model. The resulting solution placed the A_{1core} domain excellently while the F_1 domain required additional rigid body refinement. The C_2 domain from the F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 structure was used as the next search model and the MR search was able to place the entire C domain successfully. The A_{2core} from F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 was manually placed adjacent to the C_2 domain, and rigid body refined. The resulting electron density was of significantly lower quality compared to the rest of the structure. Iterative building in COOT (Emsley et al., 2010) and refinement in both Phenix (Adams et al., 2002) and CNS (Brunger, 2007) facilitated the manual placement of the A_{1sub} and PCP_1 domains

The structure of F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ was solved in $C222_1$ using an iterative MR strategy. Two molecules of F-A_{1core} (PDB 5ES5) (Reimer et al., 2016a) were used a preliminary search model and placed successfully. In the resulting electron density maps, electron density was

observed for the C_2 domain in Molecule 1 in a position similar to that of the F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$ structures. The C_2 domain from the F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$ structure was manually placed and rigid body refined into the structure. Following, the A_{2core} domain from the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 was manually placed with subsequent rigid body refinement. A homology model of PCP_2 was generated using SWISS-MODEL (Biasini et al., 2014) and the PCP_1 domain from F_1-A_1-PCP_1 (PDB 5ES8) (Reimer et al., 2016a) as a template. The PCP_2 domain was manually placed and rigid body refined. Finally, a model of the A_{2sub} domain was constructed in the same manner as the PCP_2 domain, and was manually placed with a final rigid body refinement. No electron density for the A_{1sub}, C_2, A_2 or PCP_2 domains was observed for the second molecule.

All ligands restraints were generated using the program eLBOW (Moriarty et al., 2009).

4.5 Acknowledgements

We thank J. Collucci and K. Guerard at Zamboni Chemical Solutions for help with small molecule chemical synthesis; D. Alonzo for whole protein mass spectrometry and PPE arm modeling assistance; O. Ovchinnikova for N10-fTHF synthesis advice; C. Chalut for the kind gift of BL21 EntD- cells; S. Labuik at the Canadian Light Source (CMCF) for diffraction data collection; F. Murphy at the Argonne Photo Source (NE-CAT) for diffraction data collection; and K. Bloudoff for help in alkylation experiments. We thank the members of the Schmeing lab for helpful discussion and advice. This work was supported by CIHR and a Canada Research Chair in Macromolecular Machines to T.M.S. J.M.R. is supported by an NSERC Alexander Graham Bell studentship, I.H. by a CIHR Canada Graduate Scholarship-Master's (CGS M) and M.E. by a PhD fellowship from the Boehringer Ingelheim Fonds.

4.6 Supplemental Information

4.6.1 LgrA crystallization

Constructs of F_1-A_1-PCP_1-C_2, F_1-A_1-PCP_1-C_2-A_2, , F_1-A_1-PCP_1-C_2-A_2-$PCP_{2(short)}$, F_1-A_1-PCP_1-C_2-A_2-PCP_2 and F_1-A_1-PCP_1-C_2-A_2-PCP_2-E_2 were designed using sequence alignments, secondary structure predictions and coevolution analysis using the GREMLIN server (Ovchinnikov et al., 2014). Apo and holo protein were expressed in BL21 EntD- cells (Chalut et al., 2006) and BL21 Bap1 cells (Pfeifer et al., 2001), respectively, and purified to high-caliber. Initial crystallization trials for each construct were conducted using both apo and holo protein at 4 °C and 22 °C. The inherent flexibility within NRPS modules has led to the development of several chemical biology approaches for limiting conformational heterogeneity, and their use has often facilitated structural determination of NRPS fragments (Tarry et al., 2017, Mitchell et al., 2012, Sundlov et al., 2012, Sundlov and Gulick, 2013, Drake et al., 2016, Miller et al., 2016, Liu et al., 2011). Adapting these chemical biology methods for LgrA, we synthesized six coenzyme A (CoA) analogues (amino-CoA (NH-CoA), (formyl)-valinyl-SH-CoA, (formyl)-valinyl-NH-CoA, glycinyl-NH-CoA) to target specific catalytic states. Apo protein was modified with a CoA analogue using the phosphopantetheinyl transferase, Sfp, and set into crystallization trials. Preliminary holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2 crystal hits were optimized to X-ray diffraction-quality through subsequent modification with formyl-valinyl-NH-CoA and co-crystallizing with either the valine adenylate analogue, 5-*O*-*N*-valylsulfamoyladenosmine, or the nonhydrolyzable ATP analogue, AMPcPP, with valine. Crystals for holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 were obtained after 3 months of growth, and optimized with difficulty due to low reproducibility. Holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ was set and optimized into crystallization trays using the same F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 condition, resulting in 3 different crystal forms in the same tray after 6 months of growth. Crystals were limited to <50 μm in size, with only one crystal form diffracting to 6.8 Å and the other two forms to 25 Å. A valinyl-adenosine vinylsulfamonamide (val-AVS) inhibitor was used to stall the initiation module in the thiolation state, and crystals of F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$, F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$-$PCP_{2\text{-}SH}$ and full length LgrA, F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$-$PCP_{2\text{-}SH}$-$E_2$ were all obtained in the same condition. Of these, only F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ crystals could be optimized to single crystals of a size suitable for X-ray diffraction experiments. Optimization attempts for F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$-$PCP_{2\text{-}SH}$ and F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$-

$PCP_{2\text{-}SH}$-E_2 crystals included modifying PCP_2 with a glycine AVS inhibitor to stall module 2 in the thiolation state, providing ATP, $MgCl_2$ and glycine for self-loading of PCP_2 by the A_2 domain, and co-crystallizing with an nonhydrolyzable glycine adenylate analog to promote the closed state in the A_2 domain. However, crystals resulting from these last experiments were always either too small or grew in needle clusters, and could not be used in X-ray diffraction experiments.

An additional construct of F_1-A_1-PCP_1-C_2, denoted as F_1-A_1-PCP_1-C_2-mut, was generated in attempt to visualize both the donor substrate and an acceptor substrate mimic in the C domain active site using a chemical biology approach developed by Bloudoff *et al* (Bloudoff et al., 2016). We had chemical probes synthesized that imitate the acceptor glycyl-PPE arm and were designed to alkylate an engineered cysteine (R792C) found in the acceptor substrate tunnel to place the acceptor substrate in a reaction competent position in the C domain active site. Surface cysteines C31, C191 and C318 were mutated to serine to prevent undesired alkylation. Alkylation was monitored by native liquid chromatography mass spectrometry (LC-MS) to obtain mono-alkylated F_1-A_1-PCP_1-C_2-mut protein. The resulting electron density maps for alkylated F_1-A_1-PCP_1-C_2-mut crystals did not show electron density for the chemical probe. However, the non-alkylated F_1-A_1-$PCP_{1\_NH\text{-}fval}$-C_2-mut crystals diffracted substantially better than wildtype crystals, and thus, the structure of F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut was included in this study.

4.6.2 Supplemental table

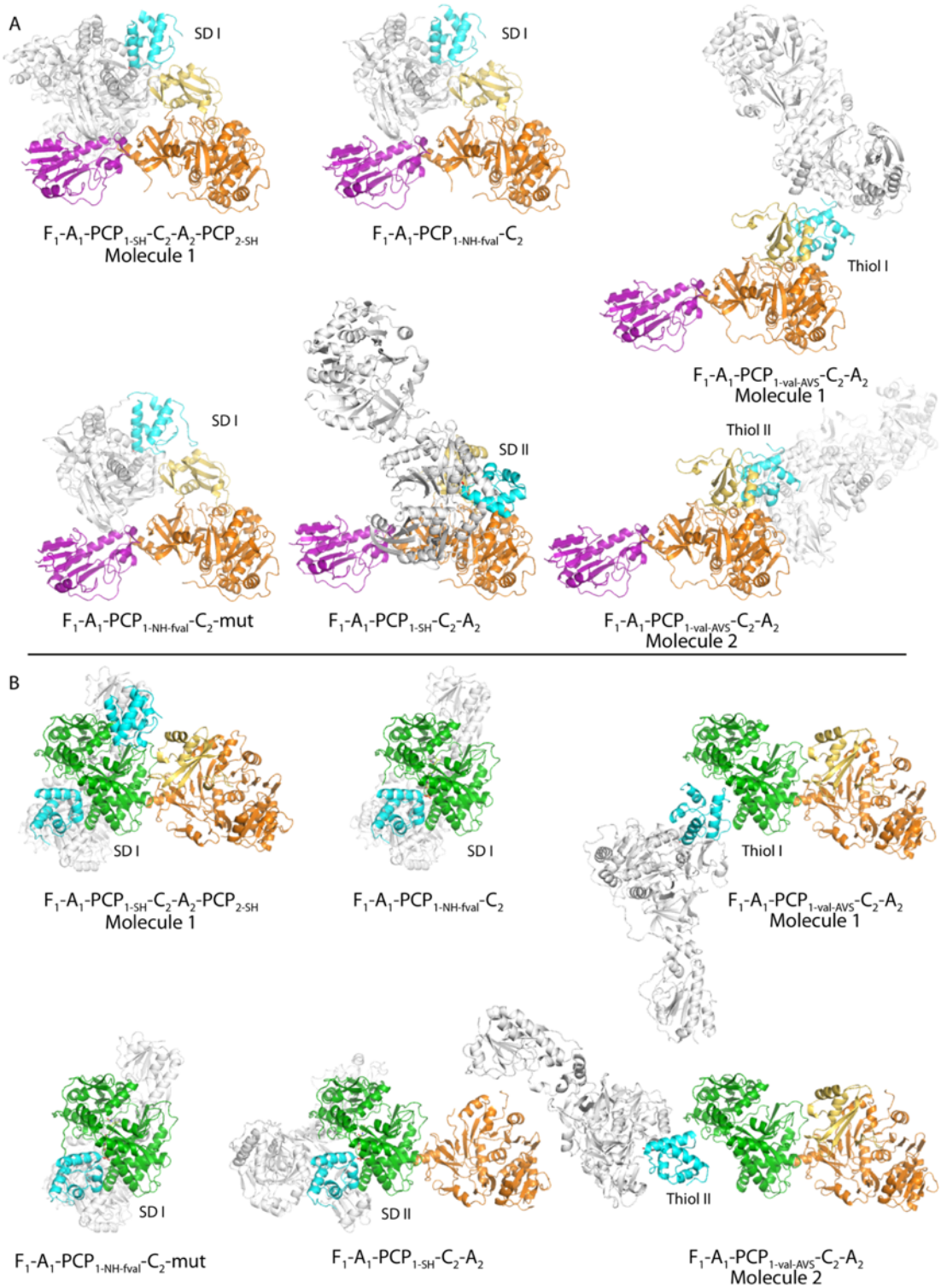
Supplemental Table 4.1 | Preliminary crystallographic statistics.

Values in parenthesis denote the highest-resolution shell.

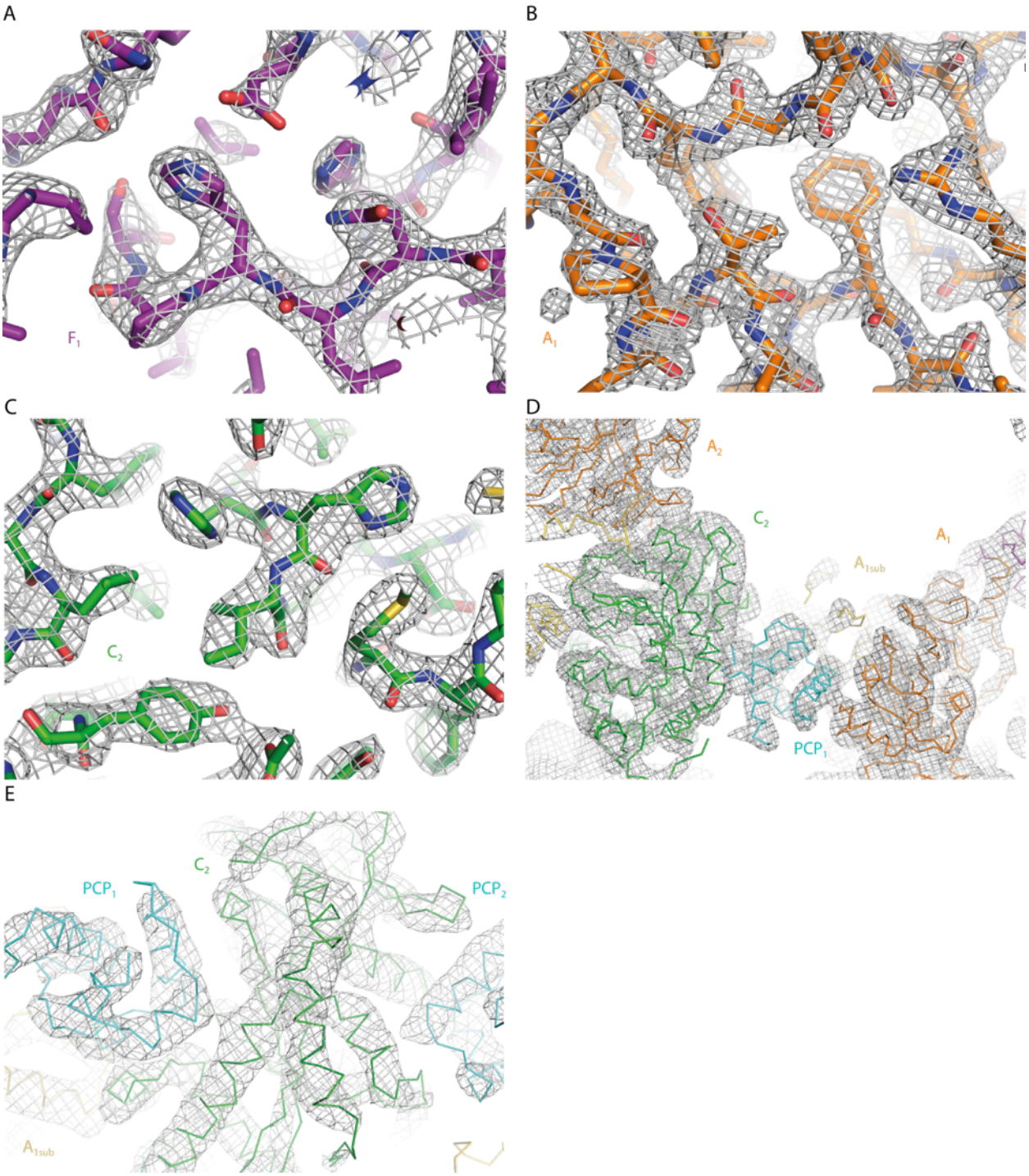
Note: Structures are currently being prepared for deposition, and therefore the refinement statistics represent the current best model.

Protein	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$	F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2	F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$	F_1-A_1-$PCP_{1\text{-}SH}$-C_2-$A_{2\text{-}SH}$
Diffraction Data					
Space group	$P2_1\ 2_1\ 2_1$	P1	$P2_1\ 2_1\ 2_1$	$P2_1\ 2_1\ 2_1$	$C2\ 2\ 2_1$
Unit-cell parameter (Å, °)	66.37 133.87 162.01 90.00 90.00 90.00	67.50 73.99 77.08 94.31 144.86 92.23	89.86 141.11 171.53 90.00 90.00 90.00	161.93 211.89 255.43 90.00 90.00 90.00	211.94 262.40 247.69 90.00 90.00 90.00
Resolution (Å)	50.00-2.35	73.55-2.05	147.53 – 2.50	78.44 – 5.40	99.02 - 6.00
I/σ	19.00 (1.06)	7.2 (1.0)	17.4 (2.0)	6.1 (1.3)	3.7 (1.2)
Temperature (K)	100	100	100	100	100
Measured reflections	57 270 (1068)	255 537 (7896)	997 427	193 202 (28 729)	209 237 (59739)
Unique reflections		78 504 (2727)	80942	30 625 (4409)	17 645 (4943)
Completeness (%)	92.3 (34.8)	93.1 (57.0)	99.7 (95.3)	99.8 (99.9)	100.0 (100.0)
Multiplicity	8.8 (2.4)	3.3 (2.9)	12.3 (9.0)	6.3 (6.5)	11.9 (12.1)
Rmerge (%)	0.141 (0.695)	0.086 (0.743)	0.095 (0.867)	0.166 (1.255)	0.616 (2.454)
Estimates of resolution limits (Å):					
From half-dataset correlation CC(1/2) > 0.30	2.35	2.17	2.50	5.89	6.48
From Mn(I/sd) > 1.50	2.48	2.21	2.50	6.04	6.58
Refinement statistics					
Resolution range (Å)	47.94 – 2.45	73.55 – 2.20	85.77 – 2.80	78.44 – 6.40	99.02 - 6.70
R factor/R_{free} (%)	21.77 / 25.77	20.60 / 24.80	21.88 / 26.56	25.52 / 29.46	24.20 / 26.53
R.M.S.D in bond lengths (Å)	0.002	0.041	0.003	0.16	0.002
R.M.S.D in bond angles (°)	0.48	0.530	0.551	1.178	0.459
Ramachandra plot (%)					
Favored	95.55	98.38	94.56	92.17	92.57
Allowed	3.60	1.36	4.72	5.95	6.22
Outliers	0.85	0.26	0.72	1.88	1.22

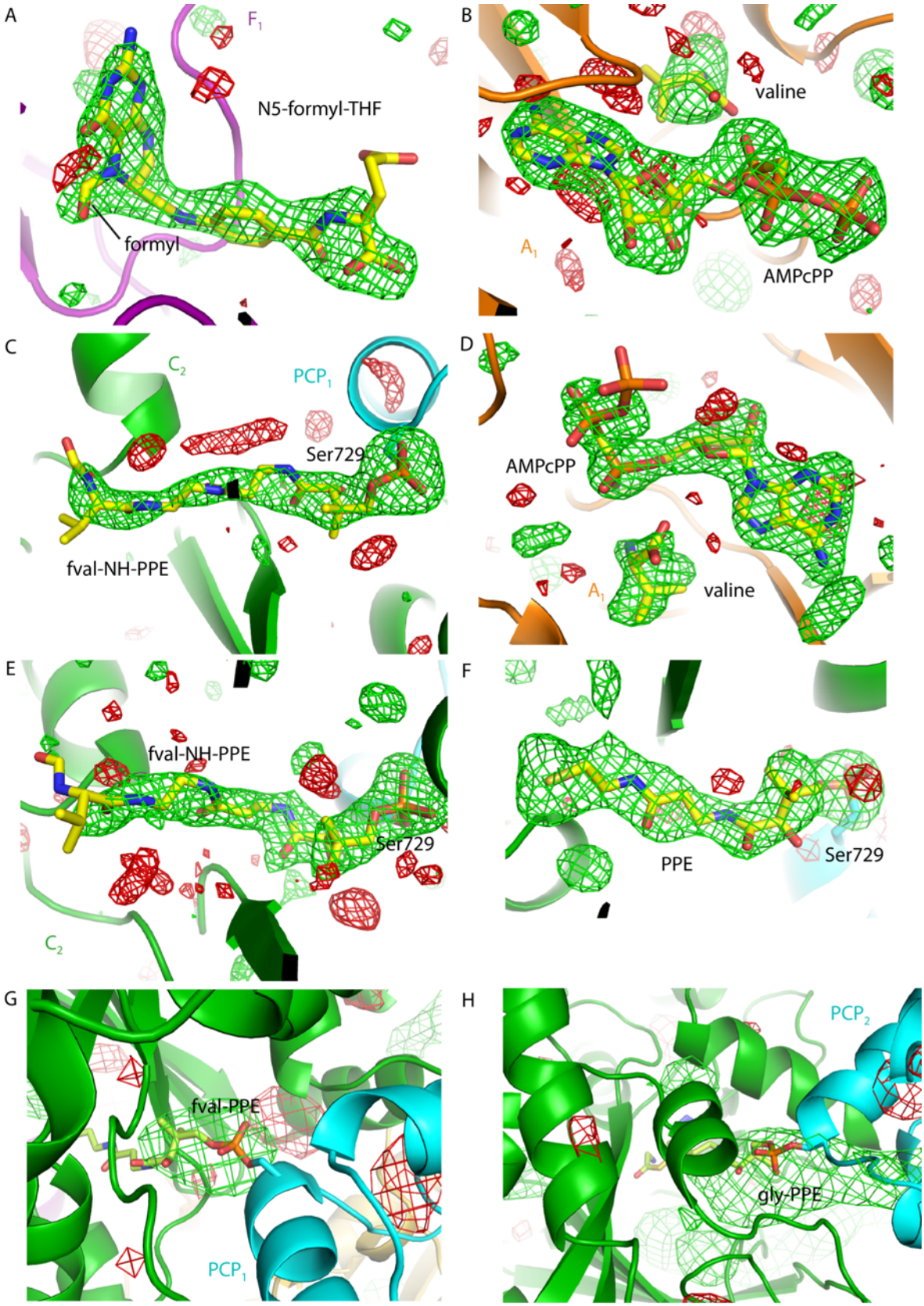
4.6.3 Supplemental Figures



Supplemental Figure 4.1 | Structures of LgrA.
Structures of LgrA aligned to the initiation module (A) and aligned to the elongation module (B). Domain colour code: F, purple; A_{core}, orange; A_{sub}, yellow; PCP, cyan; C, green; unaligned module, grey.

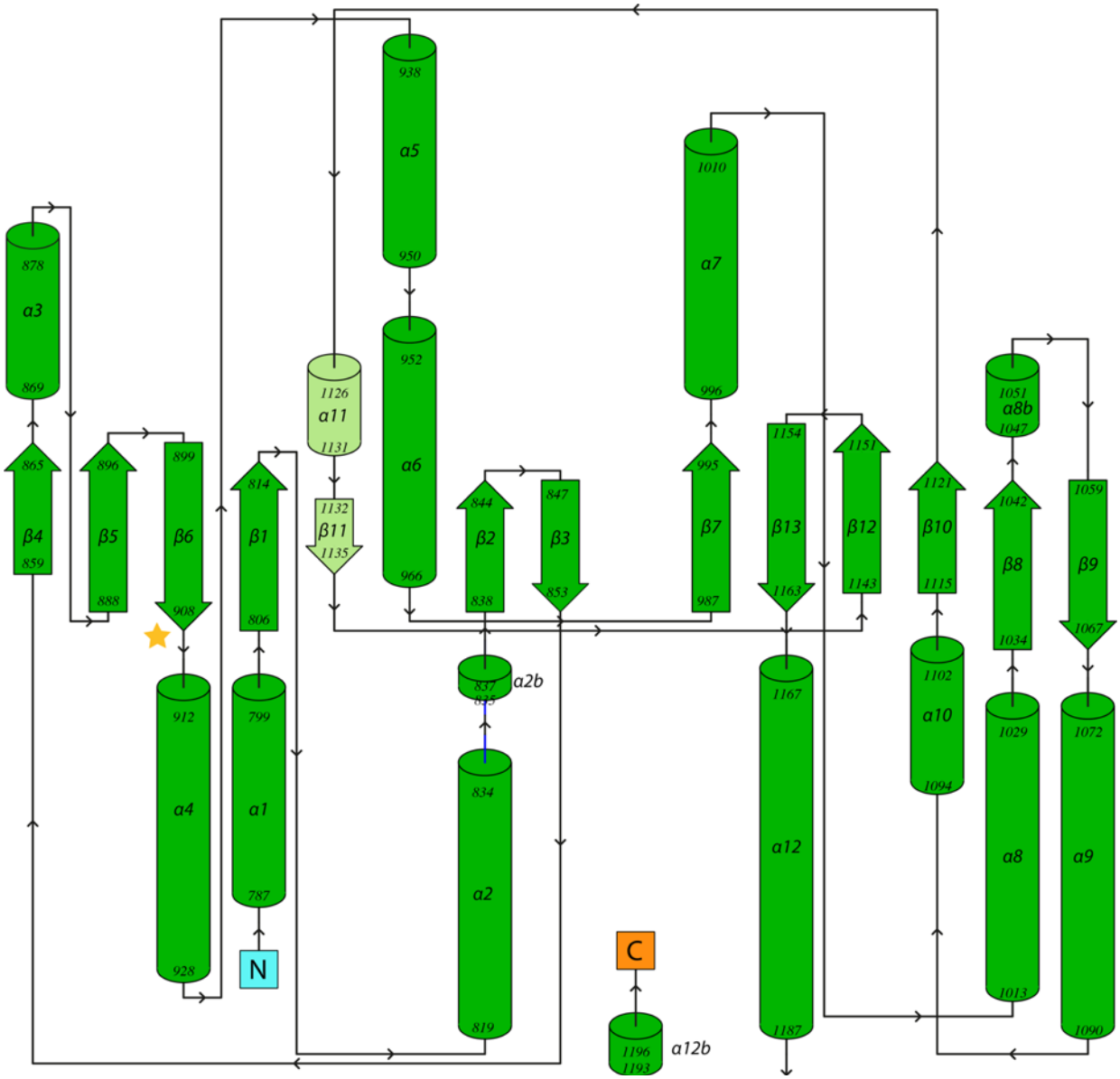


Supplemental Figure 4.2 | 2Fo-Fc electron density maps.
Representative 2Fo-Fc electron density maps for (**A**) F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 contoured to 2σ, (**B**) F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2-mut contoured to 2σ, (**C**) F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2 contoured to 2σ, (**D**) F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2 contoured to 1.5σ, and (**E**) F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ to 1.5σ.

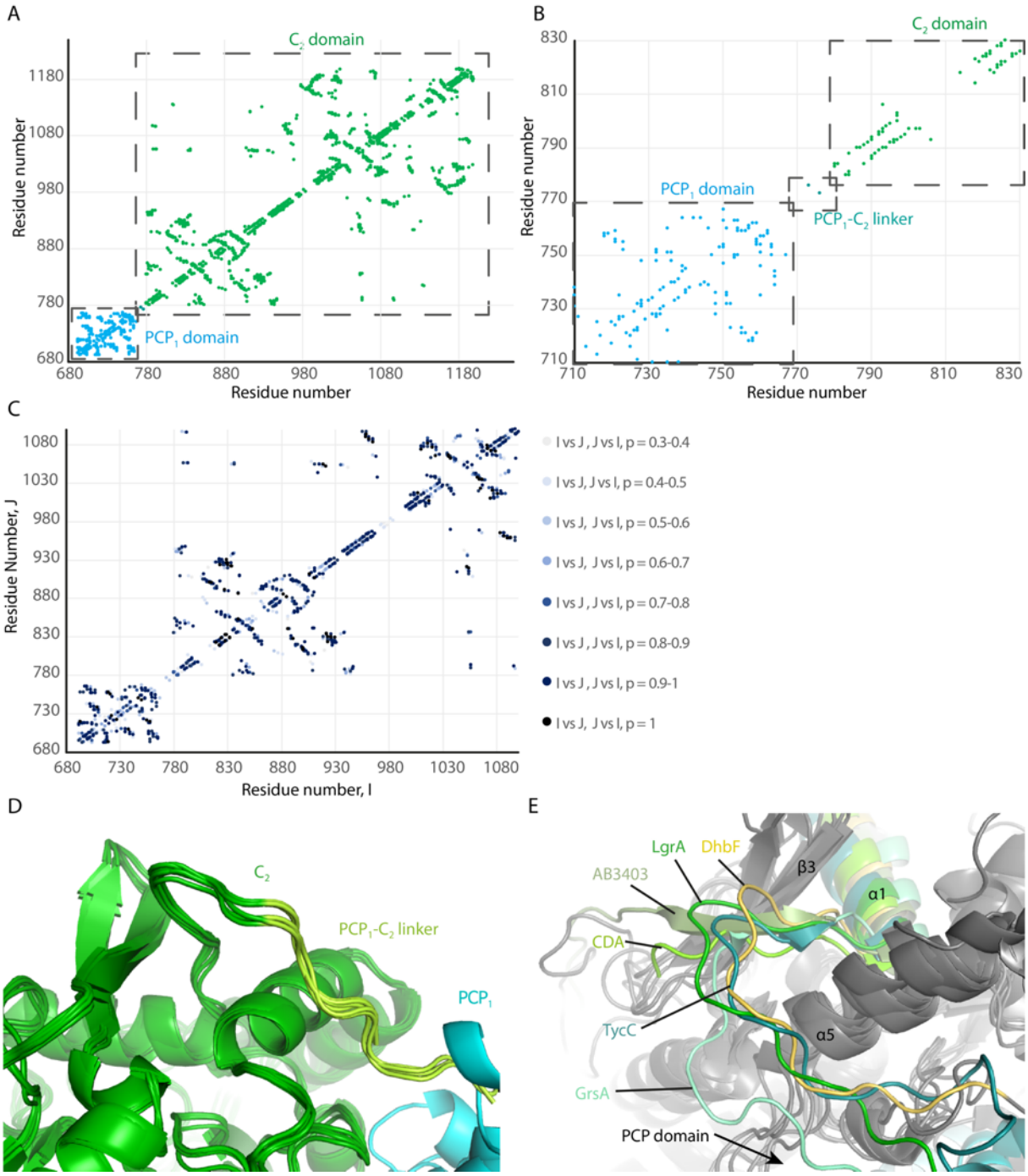


Supplemental Figure 4.3 | Fo-Fc electron density maps.
Caption on following page.

Supplemental Figure 4.3 | Fo-Fc electron density maps for ligands.
Representative Fo-Fc electron density for (**A**) N5-formyl-THF in F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$, (**B**) AMPcPP and valine in F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$, (**C**) formyl-valine-NH-PPE (fval-NH-PPE) in F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$, (**D**) AMPcPP and valine in F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut, (**E**) fval-NH-PPE in F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut, (**F**) PPE arm in F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 and (**G, H**) fval-PPE and gly-PPE at the donor and acceptor sites of the C_2 domain in F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-PCP_2, respectively (Note: fval-PPE and gly-PPE have been modelled.) All electron density maps are contoured to 3σ.



Supplemental Figure 4.4 | C2 domain topology map.
Topology map of the C_2 domain. The catalytic histidine, His908, is marked with a yellow star. The C domain latch is coloured light green. Initial diagrams were generated using PDBsum (Laskowski, 2009).



Supplemental Figure 4.5 | The PCP_1-C_2 linker of LgrA.

A, B, Evolutionary covariance analysis of the PCP_1-C_2 domain linker as determined using the GREMLIN server (Ovchinnikov et al., 2014). **C**, Probability of coevolution with the darker dots corresponding to a higher covariance strength. **D**, Overlay of all LgrA structures. The PCP_1-C_2 linker does not vary significantly between structures. **E**, The conformation of the first residues of the C (or E) domain are influenced by the presence of the PCP domain. Structure domain organization: DhbF, A-PCP-C:MLP (Tarry et al., 2017); AB3404, C-A-PCP-Te (Drake et al., 2016); CDA, C (Bloudoff et al., 2013); TycC, PCP-C (Samel et al., 2007); and GrsA, PCP-E (Chen et al., 2016).

CHAPTER 5 | GENERAL CONCLUSIONS

The first subunit of the linear gramicidin synthetase has served as an superb model for studying the synthetic cycle of these megaenzymes. In the following conclusions chapter, I will elaborate on the themes of tailoring within an NRPS and the conformational changes needed to facilitate the many reactions during synthesis. I will discuss the strategies that enabled LgrA's success as a X-ray crystallography target, as well as the future of NRPS structural biology. Lastly, I will discuss alternative methods that could be used to further explore NRPS dynamics.

5.1 Tailoring within an NRPS

The LgrA initiation module structures were the first visualization of a *cis*-acting tailoring domain embedded into the NRPS architecture. The F domain is recognizable by its homology to free-standing formyl-transferase (FT) proteins. In LgrA, the C-terminus of the F domain is fused to the N-terminus of the A domain, with little disruption of the conformation of either domain. Only ~8 residues are not recognizable as clearly belonging to F or A domain, and they form a small helix-turn, meaning there is no extended linker between F and A domains. Accordingly, a fairly extensive interface is formed the between F and A domains, burying 831 Å^2 of surface area per domain and placing the two domains in an extended conformation. This extended conformation is observed in over 10 different crystal packing environments and is consistent with small angle X-ray scattering models, and places the formylation and adenylation active sites ~50 Å apart from one another. Fortunately, the

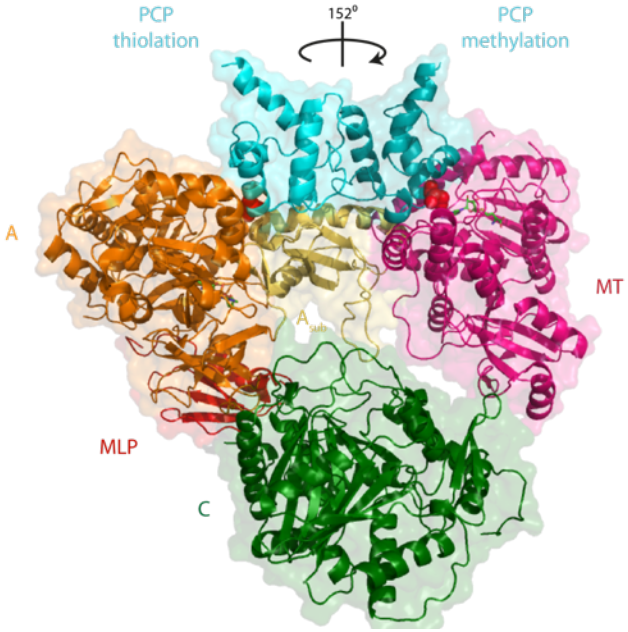


Figure 5.1 Model of TioS module.
Module model constructed by superimposing both the elongation module of LgrA and the PCP domain in the thiolation state onto the A domain of TioS (Mori et al., 2018). PPE attachment sites are shown in red spheres. Colour code: A_{core}, orange; A_{sub}, yelloworange; PCP, cyan/teal; MT, pink; C domain, green; MLP, red.

structures capture each major conformation to show how the PCP domain transports substrates between each active site.

The recent reported structure of TioS, which includes an interrupted A domain with a methyltransferase (MT) domain, is the second and only other structure of an *in cis* tailoring domain (Figure 1.7) (Mori et al., 2018). The MT is inserted between elements a8 and a9 of the A_{sub} domain, and like the LgrA initiation module, relies on the inherent flexibility of the PCP for successful integration into the synthetic cycle of the module. To better understand the conformational changes needed to include a methylation step into the synthetic cycle, Mori *et al* originally constructed a model of the entire elongation module by superimposing the termination module of SrfA-C onto the A domain of TioS (Mori et al., 2018). Similarly, the structures of LgrA can be used to build a model of the TioS module by superimposing the elongation module from LgrA onto the TioS A domain, as well as the PCP domain from LgrA in the thiolation state as the TioS A_{sub} domain is in the thiolation conformation (Figure 5.1). It has been shown that the MT can methylate aminoacyl-PCP (Mori et al., 2018), however, the A and MT domain active sites are ~ 60 Å apart. This distance exceeds the distance between the F and A domain actives sites in LgrA, and must be bridged by the action of the PCP. However, unlike LgrA, no movement of the A_{sub} domain appears to be required for the delivery of substrate to the MT active site. Based on the model, following thioesterification of the substrate onto the PPE arm, a simple ~24 Å translocation accompanied by a ~152° rotation of the PCP would be sufficient to bring the aminoacyl-PPE arm in line with a proposed substrate tunnel that is the approximate length of the PPE arm and leads to the methyl donor, S-adenosyl methionine.

Both LgrA and TioS have integrated their respective tailoring domain (F or MT) with minimal structural changes to the overall architecture of the synthetase, and have taken advantage of the flexible nature of the PCP domain to incorporate the tailoring function into the synthetic cycle. Like the MT in TioS, the insertion of tailoring domains between elements a8 and a9 is a common and clever strategy used by NRPSs. All of the known A_{sub} conformational changes keep the same side of the A_{sub} oriented towards the A_{core}, which causes the loop between a8 and a9 to always be pointed away from the C:A catalytic platform. Accordingly, a tailoring domain inserted between these two elements will project away from the module while remaining within

proximity of the PCP domain, ensuring the catalytic cycle continues undisrupted. Sequence alignments have also found N-MTs inserted between A domain elements a2 and a3, embedding the MT into the A_{core} domain instead of the A_{sub} domain (Al-Mestarihi et al., 2014). Ketoreductase domains have only been found between elements a8 and a9 (Magarvey et al., 2006, Cheng, 2006), while oxidase domains have been found between elements a8 and a9, as well as a4 and a5 (Silakowski et al., 1999, Weinig et al., 2003)}. As more structures of *in cis* tailoring domains are solved within the context of their module, we will certainly see different resourceful strategies for integrating tailoring domains into NRP synthesis.

5.2 The synthetic cycle of LgrA

Our collection of LgrA structures illustrate each major catalytic state of the NRPS synthetic cycle, and together, allow us to build a model of the linear gramicidin synthetase as it proceeds through one possible route in its initiation synthetic cycle to create the first peptide bond of linear gramicidin. Starting with the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ and following events after a previous condensation event (Figure 5.2), the PCP_1 domain returns to the A_1 domain for loading of the next valine molecule onto the PPE arm (Figure 5.2, Step 1A). The A_1 domain is in the closed position of the adenylation reaction, indicating that it has already proceeded with the next round of initiation by adenylating valine. The PCP_1 must rotate by ~157° and translocate its center of mass by ~47 Å along with a simultaneous ~144° rotation and ~17 Å shift of the A_{1sub} domain to assume the thiolation conformation observed in the F_1-A_1-$PCP_{1\text{-}NH\text{-}val}$ (PDB 5ES8) (Reimer et al., 2016a) and F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ structures. The elongation module must concurrently move closer to the C-terminal of the A_{1core} domain as the PCP_1-C_2 domain linker is too short to allow the elongation module to remain in its current conformation. The position of the elongation module is not fixed, and can be anywhere within the limits of the PCP_1-C_2 linker.

The PCP_2 domain must locate the next subunit of the Lgr synthetase, LgrB, to donate formyl-val-gly for continued elongation of the peptide. Sequence alignments and secondary structure predictions indicate that LgrA and LgrB contain docking domains on their C- and N-terminus, respectively. Docking domains are optional small domains of approximately

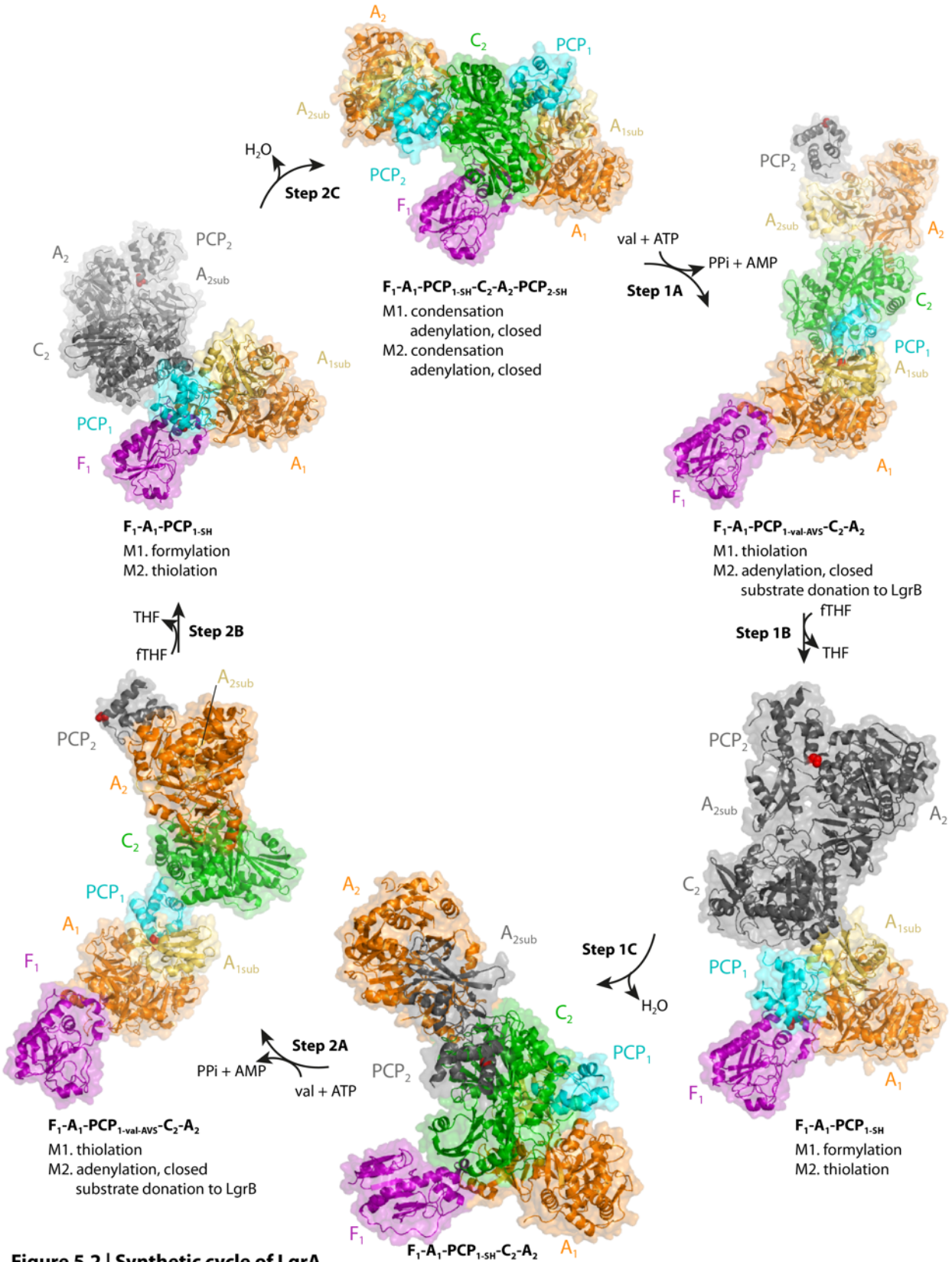


Figure 5.2 | Synthetic cycle of LgrA.
Caption on following page.

Figure 5.2 | Synthetic cycle of LgrA.
Two rounds of the LgrA synthetic cycle are modelled using crystal structures of LgrA. This illustration aims to demonstrate that module:module interactions do not define the catalytic cycle of LgrA, but can adopt multiple conformations to achieve the same catalytic outcome. Domains coloured grey are not part of the crystal structure listed and have been modelled using domains from other LgrA structures to represent a possible configuration of the domain(s). Catalytic states are listed for M1 (module 1, initiation module) and M2 (module 2, elongation module). Note: Only conformations depicting catalytic reactions are illustrated in this model, and thus, the structure of the open state of the initiation module (PDB 5ES5) was omitted. The A_{1sub} will certainly have to rotate open at some point in the catalytic cycle to allow new substrates to enter the A domain active site. PPE arm attachment site shown in red

50 residues that dimerize with a complimentary docking domain to facilitate protein-protein interactions. Dimerization of the LgrA and LgrB docking domains localizes module 3 of the Lgr synthetase to module 2 for substrate donation by PCP_2. Following successful substrate donation at LgrB-C_3, the PCP_2 and A_{2sub} domains must undergo the necessary conformational changes to return to the A_2 domain active site to load the empty PPE arm with glycine (Step 1B). Although the A_{2sub} is in a closed-liked state in the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ structure, it is unclear when adenylation actually occurs. If substrates enter the A_2 domain active site prior to condensation, glycine adenylation could occur during condensation, which is implied by the pseudo-closed state of the A_2 domain in the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 -$PCP_{2\text{-}SH}$ structure. The A_{2sub} domain must then rotate from the closed conformation to thiolation conformation during the period in which the PCP_2 domain transitions from condensation to substrate donation to thiolation states.

Before the $PCP_{1\text{-}SH\text{-}val}$ domain can return to the C_2 domain, the $PCP_{1\text{-}SH\text{-}val}$ domain must transport valine to the F domain for formylation (Step 1B). The F domain active site, which was previously blocked by the C_2 domain in the F_1-A_1-$PCP_{1\text{-}NH\text{-}fVal}$-$C_2$, F_1-A_1-$PCP_{1\text{-}NH\text{-}fval}$-$C_2$-mut and F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ structures, is now unobstructed because of the co-translocation of $PCP_{1\text{-}SH}$ and the elongation module for thiolation of valine. The A_{1sub} and $PCP_{1\text{-}SH\text{-}val}$ undergo large conformational changes, as previously reported (Reimer et al., 2016a), to bring $PCP_{1\text{-}SH\text{-}val}$ to the F domain active site. The location of the elongation module during the formylation state is currently unknown. However, the elongation module cannot remain in either position observed in the F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ structures if the $PCP_{\text{-}SH\text{-}val}$ domain is to make a productive interaction with the F domain as the length of the PCP_1-C_2 linker, once again, requires the elongation module to

concurrently move with the $PCP_{1\text{-}SH\text{-}val}$ domain. The elongation module will likely sample different positions with respect to the initiation module with each round of catalysis, with non-covalent interactions with the initiation module being transient, if at any occur.

Both $PCP_{1\text{-}SH\text{-}fval}$ and $PCP_{2\text{-}SH\text{-}gly}$ are now ready to proceed to the C_2 domain for condensation (Step 1C). As discussed earlier, both A_{sub} domains are in the closed state in the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ structure. A new set of substrates could enter the A domain active sites during the A_{sub} domain transition from thiolation to formylation states or formylation to substrate donation states for the initiation module, and likewise, during the transition from thiolation to substrate acceptance in the elongation module. The transition between the formylation and either substrate donation state requires coordinated movements between the $PCP_{1\text{-}SH\text{-}fval}$ domain and the elongation module to avoid clashing between either the PCP_1 or A_{1sub} domains and the C_2 domain. The initiation cycle repeats after condensation (Steps 2A-2C) with an endless continuum of possible positions the elongation module may adopt relative to the initiation module.

5.3 PCP domain dynamics

It is clear from the synthetic cycle description above that PCP domains must traverse large distances to shuttle the growing peptide between active sites. A common question that arises after seeing the different PCP domain movements is what drives the PCP domain to move. Unlike many other macromolecular machines with moving parts, NRPSs do not have a power stroke: No NTP hydrolysis drives PCP domain movement. Rather, it is likely the PCP domain moves predominantly randomly through tethered Brownian motion. The interdomain linkers are of variable length across NRPSs, but combined with other domain movements, provide enough freedom for the PCP domain to travel the long distances between active sites. Though the PCP domain likely samples binding to all its partners, the acyl moiety on the PPE arm should increase affinity for the appropriate partner. The presence of the various partner domains and the acyl moiety could aid progression to some extent (Goodrich et al., 2015, Linne and Marahiel, 2000), but it is the unidirectionality of the condensation reaction that dictates the direction of synthesis. The lack of power stroke, high degree of flexibility and largely undirected tethered Brownian motion result in a rate of NRPS synthesis of peptides ~3 orders of magnitude slower than that of

the ribosome. Nonetheless, NRPSs appears to be fast enough, and the products important enough, for the host microbes to keep their massive genes in their genome.

5.4 LgrA as a model for NRPS structural biology

5.4.1 LgrA's crystallographic success

LgrA has proven to be an extraordinary protein to use for studying NRPSs using X-ray crystallography. There were several key factors to its success beyond its fortuitous propensity for crystallization. First, chemical biology tools were essential for obtaining and optimizing crystals suitable for X-ray diffraction experiments. The natural flexibility of NRPSs makes them difficult targets for crystallization, and conformational states need to be controlled, or at least limited in some respect, to promote crystallization. The following examples illustrate how the use of small molecules led to structure determination. Preliminary attempts to crystallize apo F_1-A_1-PCP_1 were completely unsuccessful, but following modification with val-NH-PPE using val-NH-CoA and Sfp, over 30 crystal hits were obtained overnight that were optimized to capture the thiolation state of the initiation module. The first crystal hit for F_1-A_1-PCP_1-C_2 was obtained using protein produced in BL21 (DE3) cells. Although the PCP_1 domain wasn't intended to be modified, endogenous PPTase partially modified the PCP_1 domain *in vivo* resulting in a sub-population of holo F_1-A_1-$PCP_{1\text{-}SH}$-C_2 in the sample. This led us to test different PPE arm modifications that resulted in large reproducible crystals, instead of the clustered and bendable crystals found in the original crystal hit. The first datasets of F_1-A_1-$PCP_{1\text{-}NH}$-C_2 allowed us to build all of the F_1 and A_1 domains, but the PCP_1 domain and the C-lobe of the C_2 domain were invisible in the electron density maps. However, the structure revealed the A_1 domain in the closed position, which led to the inclusion of small molecules that would encourage the closed state. Not only did this significantly reduce the crystallization period (1-2 weeks to 1-2 days), but the resulting crystals demonstrated improved X-ray diffraction and allowed building of the missing portions of the protein. While F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ and F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 proteins were only modified with the natural PPE arm, the conformational variability of the proteins may be reflected in the crystallization times, which ranged from 3 weeks to 6 months. Secondly, having high resolution structures of the initiation module and C_2 domain were crucial for solving the structures of the larger constructs, especially

for the low resolution structures. The structure of the initiation module in the thiolation state was initially used as a search model when solving the structure of F_1-A_1-$PCP_{1\text{-val-AVS}}$-C_2-A_2, but to our great surprise, never produced a convincing solution. It was only when the C_2 domain from the structure of F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 was used as a search model that a solution was found and the rest of the domains could be placed. Likewise, the structure of F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ required using the didomain F_1-$A_{1\text{core}}$ as a search model to generate a solution. Additionally, the high resolution structures allowed good and informative models to be constructed from the low resolution data to contain fundamental information on NRPS architecture and synthesis. Altogether, the combination of testing over 30 000 crystal conditions and making informed experimental decisions based on the crystallography and biology of NRPSs, led to the successful determination of over ten crystals structures of LgrA.

5.4.2 Going forward with LgrA X-ray crystallography

The crystals for F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ were only obtained at the beginning of this year and have not gone through rigorous optimization efforts. Three different crystal forms were found in the same condition with the best crystal form diffracting to ~6.8-8 Å. The current model of F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ is in the condensation state, and further optimization of the crystals could facilitate using the structure of F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ to study the condensation reaction in detail. Our group and others have made a tremendous effort to visualize a C domain with substrates or substrates mimic in the active site. Attempts to co-crystallize or soak C domain-containing crystals with aminoacyl- or peptidyl-*N*-acetylcysteamine thioesters or aminoacyl- or peptidyl-PPE molecules have never had success (Ehmann et al., 2000, Bloudoff et al., 2013). Indeed, structures showing the PPE arm extending into the C domain or C domain homologues show limited interaction between the PPE arm and the C domain (Drake et al., 2016, Chen et al., 2016), suggesting that it is the PCP interaction with the C domain that heavily influences substrate delivery to the C domain. Even with a productive PCP-C interaction, our structures of F_1-A_1-$PCP_{1\text{-NH-fval}}$-C_2 took a great number of datasets to finally obtain a dataset that had revealed electron density for the donor substrate. The F_1-A_1-$PCP_{1\text{-SH}}$-C_2-A_2-$PCP_{2\text{-SH}}$ structure shows the PPE arms extending into the active site. If the current crystals can be optimized to diffract to higher

resolution, having both substrates in the active site at the same time may reduce substrate motion to allow full visualization of both donor and acceptor substrates. The additional two crystal forms may also provide useful conformations of LgrA, such as dimodular LgrA in the formylation state or a different condensation state.

Other constructs of LgrA may be useful for future crystallography efforts. Several of the LgrA structures have shown that the PCP domain has affinity for the C domain active site, even in the holo, unloaded state. A PCP_1-C_2-A_2-PCP_2 construct may be an easier crystallography target to strive for than optimizing the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ crystals given the reduced flexibility of the protein and the reduced size of the LgrA construct. A different strategy for loading the PCP domains with specific substrates would have to be used as Sfp cannot selectively modify individual PCP domains with different CoA molecules. A SNaPe (sortase mediated and native chemical ligation using synthetic peptide linkers) approach to creating fusion proteins between NRPS fragments has been developed that allows two separately purified proteins to be covalently linked together with a customizable linker (Ulrich and Cryle, 2017). This method could be used to produce either PCP_1 or PCP_2 separately, which could then be modified with appropriate substrate using Sfp and fused to either modified C_2-A_2-$PCP_{2\text{-}NH\text{-}gly}$ or $PCP_{1\text{-}NH\text{-}fval}$-$C_2$-$A_2$, respectively. Conversely, the PCP domains could be loaded with non-cognate substrates, such as ala-NH-PPE or gly-NH-PPE, that are similar to the native substrates and would still give an accurate depiction of the condensation reaction.

Following the theme of tailoring domains, another useful construct could be of the entire elongation module, C_2-A_2-PCP_2-E_2. Epimerization domains are common tailoring domains in NRPSs and it would be interesting to know how an E domain is incorporated into the overall modular architecture as current structural knowledge is limited to individual E (Samel et al., 2014) and PCP-E didomain (Chen et al., 2016) structures. Although the E_2 domain is inactive (catalytic His is mutated to Gln) and probably an evolutionary artifact from the initial assembly of the Lgr synthetase, the structure would still be representative of a functional E domain and illustrate its placement in the module.

5.4.3 The future of NRPS structural biology

It will be interesting to see whether the coming structures of multi-modular NRPSs will be determined by crystallography or cryo-electron microscopy (cryo-EM). Cryo-EM is now the technique of choice for large macromolecules but the moderate size of NRPS domains and especially the massive array of domain-domain and module-module conformations which are absolutely required for NRPS function (many of which remain even after targeted crosslinking), make NRPSs extremely challenging targets. A sample with a near continuum of conformations is problematic, even for the "new EM" (Frank, 2017), but will eventually be conquered. Chemical biology tools were key to obtaining structures, and new chemical biology tools will likely be needed access novel structures (Miller and Gulick, 2016, Tarry et al., 2017, Meier and Burkart, 2009, Bloudoff et al., 2016).

Additional inter-module structures, including a dimodular protein consisting of two elongation modules, would certainly be informative, but in light of the LgrA and other recent structural studies (Tarry et al., 2017, Drake et al., 2016), may not be the most beneficial target going forward in the field. Instead, other types of megaenzymes may be more desirable. In a study of 3339 gene clusters for polyketides or peptides found in 2699 genomes, 46% of gene clusters were associated with NRPS clusters, 34% with hybrid NRPS-PKS clusters and only 20% with PKS clusters (Wang et al., 2014a). Hybrid NRPS-PKS synthetases follow the same modular assembly line logic as NRPSs where carrier proteins shuttle substrates between active sites. Hybrid NRPS-PKSs have both NRPS and PKS modules in the same megaenzyme, which necessitates modification to the synthetic pathway of either canonical enzyme module (Figure 5.3A). Hybrid enzymes pose fascinating architectural questions because PKSs are functional dimers (Staunton et al., 1996, Kao, 1996) while NRPSs are monomeric (Sieber et al., 2002). Structures of NRPS-PKSs would inform how the oligomeric state requirements are satisfied in hybrid enzymes, and how carrier proteins adapt to transfer substrate to a non-native module type. However, despite the prevalence of hybrid NRPS-PKSs, no structures have been published that contain both NRPS and PKS domains. The current structural biology knowledge of PKSs is at a similar stage as of NRPSs where years of structural and biochemical data are coalescing into a functional model of how these megaenzymes operate (reviewed in (Robbins et al., 2016, Weissman, 2015)). There are unanswered questions in

both systems, but focus needs to shift on bridging the gap between NPRSs and PKSs to understand how these two related synthetic factories work together to produce their prevalent hybrid products. Additionally, structures of hybrid components will still undoubtedly be informative to NRPS and PKS function outside of the hybrid context.

A second benefit to obtaining hybrid NRPS-PKS structures is that NRPS and PKS research have the same end goal – harnessing biosynthetic machinery to bioengineer designer compounds. By understanding how NRPS and PKS components productively interact with each other without disrupting the synthetic cycle, chemistries specific to PKS synthesis can be added to the already impressive and diverse array of NRP products. The fact that hybrid NRPS-PKSs already exist in nature indicate that these hybrid natural products are useful and advantageous molecules, and worth pursuing in future studies (Figure 5.3B).

5.5 Outlook

X-ray crystallography has been indispensable for obtaining 'snapshots' of NRPSs as they undergo large conformational changes throughout their catalytic cycle. But to fully understand NRPS dynamics, other techniques must be used to bring these static snapshots to life with experiments that report on movement more directly, including fluorescence approaches (Alfermann et al., 2017), NMR experiments (Goodrich et al., 2015, Goodrich et al., 2017) and molecular dynamics simulations (Dowling et al., 2016).

5.5.1 Förster energy transfer

Förster energy transfer (FRET) is a classical method for observing protein conformational changes. The protein of interest is labelled with a donor and acceptor fluorophore and changes in fluorescence act as a readout for conformational change. FRET has only been used to study NRPSs in one previous study (Alfermann et al., 2017), and has great potential to deduce if the observed movements of the elongation module in the LgrA structures naturally occur in solution or if they are an artifact of crystallization. As an initial experiment, labelled F_1-A_1-PCP_1-C_2-A_{2core} could be obtained using a construct with terminal sortase recognition motifs. Sortases are transpeptidases that have been used extensively for labelling both the N- and C-termini of proteins (Pogliano et al.,

2012, Mao et al., 2004, Popp et al., 2007). The A_{2sub} and PCP_{2sub} domains would be omitted from the construct to limit conformational heterogeneity while maintaining the structural core of the elongation module. Following sortase treatment, termini could be labelled with a donor and acceptor pair of fluorophores with subsequent FRET measurements.

The Förster radius of the donor and acceptor fluorophores is ~60 Å, which means for energy transfer from the donor fluorophore to the acceptor fluorophore to occur, the elongation

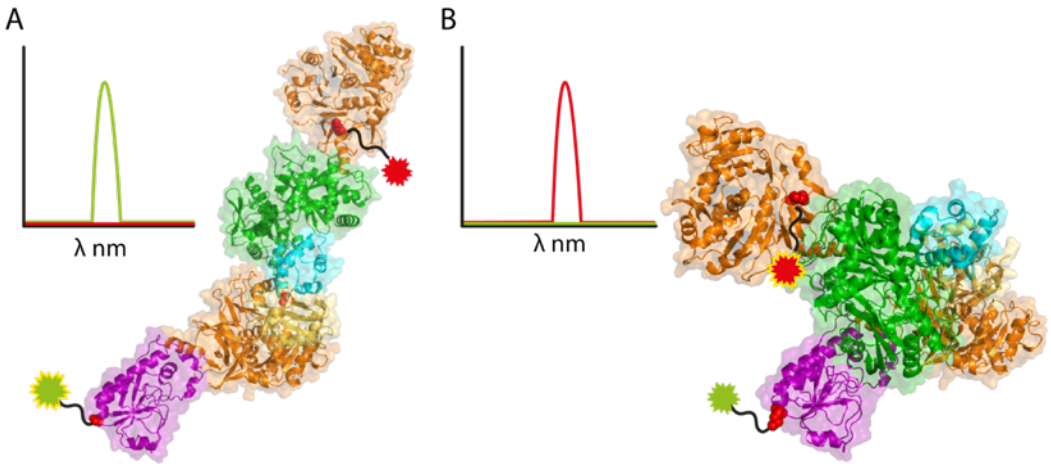


Figure 5.4 | Monitoring the LgrA catalytic cycle using FRET.
A, An extended conformation of F_1-A_1-PCP_{1}-C_2-A_{2core}, such as in the structure of F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$, would not allow energy transfer between the donor (green) and acceptor (red) fluorophores. **B**, A folded-back conformation of F_1-A_1-PCP_{1}-C_2-A_{2core}, such as in the structure of F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2, would permit energy transference.

module must be folded back against the initiation module in a conformation similar to that of the F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2-$PCP_{2\text{-}SH}$ and F_1-A_1-$PCP_{1\text{-}SH}$-C_2-A_2 structures to bring the N- and C- termini of the protein in close proximity. The extended conformations of LgrA, such as the ones observed in the F_1-A_1-$PCP_{1\text{-}val\text{-}AVS}$-$C_2$-$A_2$ structures, would exceed the Förster radius and energy transfer would not occur (Figure 5.4). The conformational change from an extended-like state to that of the folded back state is larger than the minimum required movement needed by the PCP to complete its synthetic cycle. If a FRET signal is observed, this will authenticate the flexible nature of the protein. Furthermore, the PCP_1 can be modified with coenzyme A and coenzyme A analogues to target different stages of the catalytic cycle, and the effects of those modifications tested.

One of the reoccurring questions in the NRPS field is the directionality of the PCP domain and what drives its conformational changes to move between active sites. To answer this question, an ambitious FRET strategy could be used to track the PCP_1 as it shuttles substrates between active sites by labelling F_1-A_1-PCP_1-C_2 with 4 different probes: acceptor fluorophores would be attached to the F_1 domain, A_1 domain and C_2 domain while the PCP_1 would be labelled with the donor fluorophore. Similar sortase labelling strategies as described earlier would be used to label the F_1 and C_2 domains. The A_1 domain would be labelled using either a surface cysteine or an engineered cysteine (with surface cysteines mutated to Ala or Ser). Finally, the PCP domain would be labelled using unnatural amino acid chemistry (Wang et al., 2001). By following the FRET signal, we would monitor where the PCP domain goes and if the PCP domain follows the predicted synthetic path (A domain to F domain to C domain back to A domain) or if it randomly moves between active sites. An alternative and more enthusiastic strategy would be to label our chimera construct of LgrA-BmdB (F_1-A_1-PCP_1-C_2-A_2-PCP_2-$\underline{C_{T3}}$, BmdB portion underlined), which would allow the PCP_1 domain to be followed in a multiple turnover environment. The C_2 domain would have to be labelled internally using an additional unnatural amino acid, opposed to the previous sortase-mediated label. Selective internal labels are certainly more challenging, but recent developments have shown they are possible using creative chemistry and quadruplet-decoding tRNAs for unnatural amino acids (Wang et al., 2014b).

5.5.2 Nuclear magnetic resonance

The past decade has seen an incredible advance in the field of nuclear magnetic resonance (NMR) and applications pertaining to higher molecular weight proteins. Once thought to be limited to <50 kDa proteins, new labelling techniques have opened up the realm of NMR for studying the correlation between structure and protein dynamics. Methyl-TROSY (transverse relaxation optimized spectroscopy) NMR is a technique that uses methyl group probes ($^{13}CH_3$) to monitor protein dynamics in a deuterated environment during ^{13}C-^{1}H heteronuclear multiple quantum correlation (HMQC) experiments (Tugarinov and Kay, 2004). Labelled methyl groups can easily be incorporated into the protein by providing labelled precursors for Ile, Val or Leu in the growth media (Rosenzweig and Kay, 2014). An intriguing approach to study PCP dynamics would be to use

methyl-TROSY NMR to monitor the conformational changes of the ~88 kDa LgrA initiation module. Each of the F_1, A_{1core}, A_{1sub} and PCP_1 domains contain Ile residues, allowing their movements to be detected and followed by NMR. Experiments could answer questions such as how often a given conformational state is occupied, address the time scale needed for conformational changes in the PCP_1 and A_{1sub} domains, how the presence of substrate on the PPE arm changes the conformational dynamics, and follow the PCP_1 domain as it travels between the F_1 and A_1 domain active sites.

5.5.3 Molecular dynamics

Our collection of LgrA structures make an exciting model for use in molecular dynamics simulations as they depict two adjacent modules within the same synthetase, negating the need to construct a dimodular model from domains/modules of different synthetases with potentially disruptive non-cognate interactions. Two different types of molecular dynamics simulations would be beneficial to try with LgrA: Brownian dynamics (BD) and Gaussian accelerated molecular dynamics (GAMD). BD is computationally lighter as it treats protein domains as rigid bodies to simulate how large macromolecules diffuse through an aqueous environment in up to millisecond time scales (Elcock, 2004). Using BD, the motions that LgrA, modified with different substrates on the PPE arms, may go through in solution could be simulated and compared to the conformations observed in our crystal structures. Additional information could also potentially be gained on which conformations are more favourable i.e. do modules prefer to be in an extended conformation (like in the structure of F_1-A_1-$PCP_{1\text{-val-AVS}}$-A_2) or do they prefer being folded into a more compact conformation (like in the structure of F_1-A_1-$PCP_{1\text{-SH}}$-A_2-$PCP_{2\text{-SH}}$). On the other hand, GAMD takes into account all of the atoms present in the model and uses an altered energy minimization algorithm to smooth the energy potential profile of the MD simulation, allowing for faster computations over longer time scales (Miao and McCammon, 2017). Using GAMD, the motions that LgrA may go through to transition between the conformational states observe in our crystal structures could be simulated. Both the BD and GAMD simulations would help to better understand the dynamics and motions that govern NRP synthesis.

5.5.4 Bioengineering

The modular organization of NRPSs combined with predictable products have made NRPSs attractive candidates for domain/module swapping experiments since their modular organization was first delineated. However, attempts to domain/module swap have often proven unsuccessful due to substrate specificity requirements and unproductive protein-protein interactions between swapped partners (Kraas et al., 2012, Duerfahrt et al., 2003) . Successful synthesis requires that the PCP_n domain be able to productively interact with the C_{n+1} domain to pass the nascent peptide to the next module. Our LgrA structures will be instructive for future bioengineering attempts and help inform decisions based on the PCP_1-C_2 interaction and overall modular organization observed in the structures. Furthermore, our LgrA-BmdB-C_3 chimera represents a successful instance of bioengineering through module supplementation, and addition of BmdB-C_3 to other NRPS chimeras may be beneficial for assaying their activity.

5.6 Final Statement

Nonribosomal peptide synthetase are truly fascinating mega-enzymes that have developed an elegant and versatile catalytic cycle. The experiments described in this thesis have contributed to the body of beautiful and insightful research aimed at understanding their function and structure. The NRPS field has reached an exciting crossroads where complimentary biophysical techniques must be used in conjugation with structural biology approaches to fully understand how these nanomachines operate in nature.

References

ACERBO, A. S., COOK, M. J. & GILLILAN, R. E. 2015. Upgrade of MacCHESS facility for X-ray scattering of biological macromolecules in solution. *Journal of Synchrotron Radiation,* 22**,** 180-186.

ADAMS, P. D., AFONINE, P. V., BUNKOCZI, G., CHEN, V. B., DAVIS, I. W., ECHOLS, N., HEADD, J. J., HUNG, L. W., KAPRAL, G. J., GROSSE-KUNSTLEVE, R. W., MCCOY, A. J., MORIARTY, N. W., OEFFNER, R., READ, R. J., RICHARDSON, D. C., RICHARDSON, J. S., TERWILLIGER, T. C. & ZWART, P. H. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr,* 66**,** 213-21.

ADAMS, P. D., GROSSE-KUNSTLEVE, R. W., HUNG, L. W., IOERGER, T. R., MCCOY, A. J., MORIARTY, N. W., READ, R. J., SACCHETTINI, J. C., SAUTER, N. K. & TERWILLIGER, T. C. 2002. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr,* 58**,** 1948-54.

AL-MESTARIHI, A. H., VILLAMIZAR, G., FERNÁNDEZ, J., ZOLOVA, O. E., LOMBÓ, F. & GARNEAU-TSODIKOVA, S. 2014. Adenylation and S-Methylation of Cysteine by the Bifunctional Enzyme TioN in Thiocoraline Biosynthesis. *Journal of the American Chemical Society,* 136**,** 17350-17354.

ALFERMANN, J., SUN, X., MAYERTHALER, F., MORRELL, T. E., DEHLING, E., VOLKMANN, G., KOMATSUZAKI, T., YANG, H. & MOOTZ, H. D. 2017. FRET monitoring of a nonribosomal peptide synthetase. *Nat Chem Biol,* 13**,** 1009-1015.

ALLEN, C. L. & GULICK, A. M. 2014. Structural and bioinformatic characterization of an Acinetobacter baumannii type II carrier protein. *Acta Crystallographica Section D,* 70**,** 1718-1725.

ALMASSY, R. J., JANSON, C. A., KAN, C. C. & HOSTOMSKA, Z. 1992. Structures of apo and complexed Escherichia coli glycinamide ribonucleotide transformylase. *Proc Natl Acad Sci U S A,* 89**,** 6114-8.

ALONZO, D. A., MAGARVEY, N. A. & SCHMEING, T. M. 2015. Characterization of cereulide synthetase, a toxin-producing macromolecular machine. *PLoS One,* 10**,** e0128569.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ANSARI, M. Z., SHARMA, J., GOKHALE, R. S. & MOHANTY, D. 2008. In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. *BMC Bioinformatics,* 9**,** 454.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

BALIBAR, C. J., VAILLANCOURT, F. H. & WALSH, C. T. 2005. Generation of D amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. *Chem Biol,* 12**,** 1189-200.

BELSHAW, P. J., WALSH, C. T. & STACHELHAUS, T. 1999. Aminoacyl-CoAs as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science,* 284**,** 486-9.
BERGENDAHL, V., LINNE, U. & MARAHIEL, M. A. 2002. Mutational analysis of the C-domain in nonribosomal peptide synthesis. *Eur J Biochem,* 269**,** 620-9.
BIASINI, M., BIENERT, S., WATERHOUSE, A., ARNOLD, K., STUDER, G., SCHMIDT, T., KIEFER, F., CASSARINO, T. G., BERTONI, M., BORDOLI, L. & SCHWEDE, T. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res,* 42**,** W252-8.
BIELING, P., BERINGER, M., ADIO, S. & RODNINA, M. V. 2006. Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nat Struct Mol Biol,* 13**,** 423-8.
BLOUDOFF, K., ALONZO, D. A. & SCHMEING, T. M. 2016. Chemical Probes Allow Structural Insight into the Condensation Reaction of Nonribosomal Peptide Synthetases. *Cell Chem Biol,* 23**,** 331-9.
BLOUDOFF, K., FAGE, C. D., MARAHIEL, M. A. & SCHMEING, T. M. 2017. Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. *Proc Natl Acad Sci U S A,* 114**,** 95-100.
BLOUDOFF, K., RODIONOV, D. & SCHMEING, T. M. 2013. Crystal structures of the first condensation domain of CDA synthetase suggest conformational changes during the synthetic cycle of nonribosomal peptide synthetases. *J Mol Biol,* 425**,** 3137-50.
BODE, H. B., BRACHMANN, A. O., JADHAV, K. B., SEYFARTH, L., DAUTH, C., FUCHS, S. W., KAISER, M., WATERFIELD, N. R., SACK, H., HEINEMANN, S. H. & ARNDT, H.-D. 2015. Structure elucidation and activity of kolossin A, the D-/L-pentadecapeptide product of a giant nonribosomal peptide synthetase. *Angewandte Chemie International Edition,* 54**,** 10352-10355.
BOSELLO, M., ZEYADI, M., KRAAS, F. I., LINNE, U., XIE, X. & MARAHIEL, M. A. 2013. Structural Characterization of the Heterobactin Siderophores from Rhodococcus erythropolis PR4 and Elucidation of Their Biosynthetic Machinery. *Journal of Natural Products,* 76**,** 2282-2290.
BOZHUYUK, K. A. J., FLEISCHHACKER, F., LINCK, A., WESCHE, F., TIETZE, A., NIESERT, C. P. & BODE, H. B. 2018. De novo design and engineering of non-ribosomal peptide synthetases. *Nat Chem,* 10**,** 275-281.
BREAZEALE, S. D., RIBEIRO, A. A., MCCLERREN, A. L. & RAETZ, C. R. 2005. A formyltransferase required for polymyxin resistance in Escherichia coli and the modification of lipid A with 4-Amino-4-deoxy-L-arabinose. Identification and function oF UDP-4-deoxy-4-formamido-L-arabinose. *J Biol Chem,* 280**,** 14154-67.
BREAZEALE, S. D., RIBEIRO, A. A. & RAETZ, C. R. 2002. Oxidative decarboxylation of UDP-glucuronic acid in extracts of polymyxin-resistant Escherichia coli. Origin of lipid a species modified with 4-amino-4-deoxy-L-arabinose. *J Biol Chem,* 277**,** 2886-96.
BRUNER, S. D., WEBER, T., KOHLI, R. M., SCHWARZER, D., MARAHIEL, M. A., WALSH, C. T. & STUBBS, M. T. 2002. Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. *Structure,* 10**,** 301-10.
BRUNGER, A. T. 2007. Version 1.2 of the crystallography and NMR system. *Nat Protoc,* 2**,** 2728-33.

CABOCHE, S., PUPIN, M., LECLERE, V., FONTAINE, A., JACQUES, P. & KUCHEROV, G. 2008. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res,* 36**,** D326-31.
CALCOTT, M. J. & ACKERLEY, D. F. 2014. Genetic manipulation of non-ribosomal peptide synthetases to generate novel bioactive peptide products. *Biotechnol Lett,* 36**,** 2407-16.
CHAKRABORTY, S., VENKATRAMANI, R., RAO, B. J., ASGEIRSSON, B. & DANDEKAR, A. M. 2013. Protein structure quality assessment based on the distance profiles of consecutive backbone Calpha atoms. *F1000Res,* 2**,** 211.
CHALLIS, G. L., RAVEL, J. & TOWNSEND, C. A. 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol,* 7**,** 211-24.
CHALUT, C., BOTELLA, L., DE SOUSA-D'AURIA, C., HOUSSIN, C. & GUILHOT, C. 2006. The nonredundant roles of two 4'-phosphopantetheinyl transferases in vital processes of Mycobacteria. *Proc Natl Acad Sci U S A,* 103**,** 8511-6.
CHEN, W.-H., LI, K., GUNTAKA, N. S. & BRUNER, S. D. 2016. Interdomain and Intermodule Organization in Epimerization Domain Containing Nonribosomal Peptide Synthetases. *ACS chemical biology,* 11**,** 2293-303.
CHENG, Y. Q. 2006. Deciphering the biosynthetic codes for the potent anti-SARS-CoV cyclodepsipeptide valinomycin in Streptomyces tsusimaensis ATCC 15141. *Chembiochem,* 7**,** 471-7.
CHENG, Y. Q. & WALTON, J. D. 2000. A eukaryotic alanine racemase gene involved in cyclic peptide biosynthesis. *J Biol Chem,* 275**,** 4906-11.
CILIA LA CORTE, A. L., PHILIPPOU, H. & ARIENS, R. A. 2011. Role of fibrin structure in thrombosis and vascular disease. *Adv Protein Chem Struct Biol,* 83**,** 75-127.
CLARDY, J., FISCHBACH, M. A. & WALSH, C. T. 2006. New antibiotics from bacterial natural products. *Nat Biotechnol,* 24**,** 1541-50.
CLUGSTON, S. L., SIEBER, S. A., MARAHIEL, M. A. & WALSH, C. T. 2003. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a (D)C(L) catalyst. *Biochemistry,* 42**,** 12095-104.
COEFFET-LE GAL, M. F., THURSTON, L., RICH, P., MIAO, V. & BALTZ, R. H. 2006. Complementation of daptomycin dptA and dptD deletion mutations in trans and production of hybrid lipopeptide antibiotics. *Microbiology,* 152**,** 2993-3001.
CONTI, E., FRANKS, N. P. & BRICK, P. 1996. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure,* 4**,** 287-98.
CONTI, E., STACHELHAUS, T., MARAHIEL, M. A. & BRICK, P. 1997. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J,* 16**,** 4174-83.
COOPER, E. D. 2014. Horizontal gene transfer: accidental inheritance drives adaptation. *Curr Biol,* 24**,** R562-r564.
CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res,* 14**,** 1188-90.
CRUMP, M. P., CROSBY, J., DEMPSEY, C. E., PARKINSON, J. A., MURRAY, M., HOPWOOD, D. A. & SIMPSON, T. J. 1997. Solution structure of the actinorhodin polyketide synthase acyl carrier protein from Streptomyces coelicolor A3(2). *Biochemistry,* 36**,** 6000-8.

DAI, M., FENG, Y. & TONGE, P. J. 2001. Synthesis of crotonyl-oxyCoA: a mechanistic probe of the reaction catalyzed by enoyl-CoA hydratase. *J Am Chem Soc,* 123**,** 506-7.

DE CRECY-LAGARD, V., MARLIERE, P. & SAURIN, W. 1995. Multienzymatic non ribosomal peptide biosynthesis: identification of the functional domains catalysing peptide elongation and epimerisation. *C R Acad Sci III,* 318**,** 927-36.

DE MAAYER, P. & COWAN, D. A. 2016. Comparative genomic analysis of the flagellin glycosylation island of the Gram-positive thermophile Geobacillus. *BMC Genomics,* 17**,** 913.

DELVAUX, N. A., THODEN, J. B. & HOLDEN, H. M. 2017. Molecular Architectures of Pen and Pal: Key Enzymes Required for CMP-Pseudaminic Acid Biosynthesis in Bacillus thuringiensis. *Protein Sci*.

DITTMANN, J., WENGER, R. M., KLEINKAUF, H. & LAWEN, A. 1994. Mechanism of cyclosporin A biosynthesis. Evidence for synthesis via a single linear undecapeptide precursor. *J Biol Chem,* 269**,** 2841-6.

DOWLING, D. P., KUNG, Y., CROFT, A. K., TAGHIZADEH, K., KELLY, W. L., WALSH, C. T. & DRENNAN, C. L. 2016. Structural elements of an NRPS cyclization domain and its intermodule docking domain. *Proc Natl Acad Sci U S A,* 113**,** 12432-12437.

DRAKE, E. J., MILLER, B. R., SHI, C., TARRASCH, J. T., SUNDLOV, J. A., ALLEN, C. L., SKINIOTIS, G., ALDRICH, C. C. & GULICK, A. M. 2016. Structures of two distinct conformations of holo-non-ribosomal peptide synthetases. *Nature,* 529**,** 235-8.

DU, L., CHEN, M., SANCHEZ, C. & SHEN, B. 2000. An oxidation domain in the BlmIII non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in Streptomyces verticillus ATCC15003. *FEMS Microbiol Lett,* 189**,** 171-5.

DU, L., HE, Y. & LUO, Y. 2008. Crystal structure and enantiomer selection by D-alanyl carrier protein ligase DltA from Bacillus cereus. *Biochemistry,* 47**,** 11473-80.

DUERFAHRT, T., DOEKEL, S., SONKE, T., QUAEDFLIEG, P. J. & MARAHIEL, M. A. 2003. Construction of hybrid peptide synthetases for the production of alpha-l-aspartyl-l-phenylalanine, a precursor for the high-intensity sweetener aspartame. *Eur J Biochem,* 270**,** 4555-63.

DUERFAHRT, T., EPPELMANN, K., MULLER, R. & MARAHIEL, M. A. 2004. Rational design of a bimodular model system for the investigation of heterocyclization in nonribosomal peptide biosynthesis. *Chem Biol,* 11**,** 261-71.

DUNSIRN, M. M., THODEN, J. B., GILBERT, M. & HOLDEN, H. M. 2017. Biochemical Investigation of Rv3404c from Mycobacterium tuberculosis. *Biochemistry,* 56**,** 3818-3825.

DUUS, J., GOTFREDSEN, C. H. & BOCK, K. 2000. Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chem Rev,* 100**,** 4589-614.

EHMANN, D. E., TRAUGER, J. W., STACHELHAUS, T. & WALSH, C. T. 2000. Aminoacyl-SNACs as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chem Biol,* 7**,** 765-72.

ELCOCK, A. H. 2004. Molecular simulations of diffusion and association in multimacromolecular systems. *Methods Enzymol,* 383**,** 166-98.

EMSLEY, P., LOHKAMP, B., SCOTT, W. G. & COWTAN, K. 2010. Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography,* 66**,** 486-501.

EPPELMANN, K., STACHELHAUS, T. & MARAHIEL, M. A. 2002. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry,* 41**,** 9718-26.

EVANS, B. S., CHEN, Y., METCALF, W. W., ZHAO, H. & KELLEHER, N. L. 2011. Directed evolution of the nonribosomal peptide synthetase AdmK generates new andrimid derivatives in vivo. *Chem Biol,* 18**,** 601-7.

EVANS, P. R. & MURSHUDOV, G. N. 2013. How good are my data and what is the resolution? *Acta crystallographica Section D, Biological crystallography,* 69**,** 1204-14.

FELNAGLE, E. A., JACKSON, E. E., CHAN, Y. A., PODEVELS, A. M., BERTI, A. D., MCMAHON, M. D. & THOMAS, M. G. 2008. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol Pharm,* 5**,** 191-211.

FERRERAS, J. A., RYU, J. S., DI LELLO, F., TAN, D. S. & QUADRI, L. E. 2005. Small-molecule inhibition of siderophore biosynthesis in Mycobacterium tuberculosis and Yersinia pestis. *Nat Chem Biol,* 1**,** 29-32.

FINKING, R., NEUMULLER, A., SOLSBACHER, J., KONZ, D., KRETZSCHMAR, G., SCHWEITZER, M., KRUMM, T. & MARAHIEL, M. A. 2003. Aminoacyl adenylate substrate analogues for the inhibition of adenylation domains of nonribosomal peptide synthetases. *Chembiochem,* 4**,** 903-6.

FISCHBACH, M. A., WALSH, C. T. & CLARDY, J. 2008. The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 4601-4608.

FORREST, A. K., JARVEST, R. L., MENSAH, L. M., O'HANLON, P. J., POPE, A. J. & SHEPPARD, R. J. 2000. Aminoalkyl adenylate and aminoacyl sulfamate intermediate analogues differing greatly in affinity for their cognate Staphylococcus aureus aminoacyl tRNA synthetases. *Bioorg Med Chem Lett,* 10**,** 1871-4.

FORTIN, P. D., WALSH, C. T. & MAGARVEY, N. A. 2007. A transglutaminase homologue as a condensation catalyst in antibiotic assembly lines. *Nature,* 448**,** 824-7.

FRANK, J. 2017. Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nat Protoc,* 12**,** 209-212.

FRANKE, D. & SVERGUN, D. I. 2009. DAMMIF, a program for rapidab-initioshape determination in small-angle scattering. *Journal of Applied Crystallography,* 42**,** 342-346.

FRUEH, D. P., ARTHANARI, H., KOGLIN, A., VOSBURG, D. A., BENNETT, A. E., WALSH, C. T. & WAGNER, G. 2008. Dynamic thiolation-thioesterase structure of a non-ribosomal peptide synthetase. *Nature,* 454**,** 903-6.

GAHLOTH, D., DUNSTAN, M. S., QUAGLIA, D., KLUMBYS, E., LOCKHART-CAIRNS, M. P., HILL, A. M., DERRINGTON, S. R., SCRUTTON, N. S., TURNER, N. J. & LEYS, D. 2017. Structures of carboxylic acid reductase reveal domain dynamics underlying catalysis. *Nat Chem Biol,* 13**,** 975-981.

GALONIC, D. P., VAILLANCOURT, F. H. & WALSH, C. T. 2006. Halogenation of unactivated carbon centers in natural product biosynthesis: trichlorination of leucine during barbamide biosynthesis. *J Am Chem Soc,* 128**,** 3900-1.

GATZEVA-TOPALOVA, P. Z., MAY, A. P. & SOUSA, M. C. 2005. Crystal Structure and Mechanism of the Escherichia coli ArnA (PmrI) Transformylase Domain. An Enzyme for Lipid A

Modification with 4-Amino-4-deoxy-L-arabinose and Polymyxin Resistance. *Biochemistry,* 44**,** 5328-5338.
GAUDELLI, N. M. & TOWNSEND, C. A. 2014. Epimerization and substrate gating by a TE domain in beta-lactam antibiotic biosynthesis. *Nat Chem Biol,* 10**,** 251-8.
GENTHE, N. A., THODEN, J. B., BENNING, M. M. & HOLDEN, H. M. 2015. Molecular structure of an N-formyltransferase from Providencia alcalifaciens O30. *Protein Sci,* 24**,** 976-86.
GOODRICH, A. C., HARDEN, B. J. & FRUEH, D. P. 2015. Solution structure of a nonribosomal peptide synthetase carrier protein loaded with Its substrate reveals transient, well-defined contacts. *J Am Chem Soc,* 137**,** 12100-9.
GOODRICH, A. C., MEYERS, D. J. & FRUEH, D. P. 2017. Molecular impact of covalent modifications on nonribosomal peptide synthetase carrier protein communication. *J Biol Chem,* 292**,** 10002-10013.
GUERRY, P., EWING, C. P., SCHIRM, M., LORENZO, M., KELLY, J., PATTARINI, D., MAJAM, G., THIBAULT, P. & LOGAN, S. 2006. Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence. *Mol Microbiol,* 60**,** 299-311.
GULICK, A. M. 2009. Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem Biol,* 4**,** 811-27.
GULICK, A. M. 2016. Structural insight into the necessary conformational changes of modular nonribosomal peptide synthetases. *Curr Opin Chem Biol,* 35**,** 89-96.
GULICK, A. M. 2017. Nonribosomal peptide synthetase biosynthetic clusters of ESKAPE pathogens. *Nat Prod Rep,* 34**,** 981-1009.
GULICK, A. M., STARAI, V. J., HORSWILL, A. R., HOMICK, K. M. & ESCALANTE-SEMERENA, J. C. 2003. The 1.75 A crystal structure of acetyl-CoA synthetase bound to adenosine-5'-propylphosphate and coenzyme A. *Biochemistry,* 42**,** 2866-73.
HANCHI, H., HAMMAMI, R., FERNANDEZ, B., KOURDA, R., BEN HAMIDA, J. & FLISS, I. 2016. Simultaneous Production of Formylated and Nonformylated Enterocins L50A and L50B as well as 61A, a New Glycosylated Durancin, by Enterococcus durans 61A, a Strain Isolated from Artisanal Fermented Milk in Tunisia. *Journal of agricultural and food chemistry,* 64**,** 3584-90.
HASLINGER, K., PESCHKE, M., BRIEKE, C., MAXIMOWITSCH, E. & CRYLE, M. J. 2015. X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature,* 521**,** 105-9.
HERBST, D. A., BOLL, B., ZOCHER, G., STEHLE, T. & HEIDE, L. 2013. Structural basis of the interaction of MbtH-like proteins, putative regulators of nonribosomal peptide biosynthesis, with adenylating enzymes. *J Biol Chem,* 288**,** 1991-2003.
HISANAGA, Y., AGO, H., NAKAGAWA, N., HAMADA, K., IDA, K., YAMAMOTO, M., HORI, T., ARII, Y., SUGAHARA, M., KURAMITSU, S., YOKOYAMA, S. & MIYANO, M. 2004. Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-CoA synthetase dimer. *J Biol Chem,* 279**,** 31717-26.
HOFFMANN, K., SCHNEIDER-SCHERZER, E., KLEINKAUF, H. & ZOCHER, R. 1994. Purification and characterization of eucaryotic alanine racemase acting as key enzyme in cyclosporin biosynthesis. *J Biol Chem,* 269**,** 12710-4.
HOJATI, Z., MILNE, C., HARVEY, B., GORDON, L., BORG, M., FLETT, F., WILKINSON, B., SIDEBOTTOM, P. J., RUDD, B. A., HAYES, M. A., SMITH, C. P. & MICKLEFIELD, J. 2002.

Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. *Chem Biol,* 9**,** 1175-87.
HOLDEN, H. M., THODEN, J. B. & GILBERT, M. 2016. Enzymes required for the biosynthesis of N-formylated sugars. *Current Opinion in Structural Biology,* 41**,** 1-9.
HOPPERT, M., GENTZSCH, C. & SCHORGENDORFER, K. 2001. Structure and localization of cyclosporin synthetase, the key enzyme of cyclosporin biosynthesis in Tolypocladium inflatum. *Arch Microbiol,* 176**,** 285-93.
HOSAKA, T., OHNISHI-KAMEYAMA, M., MURAMATSU, H., MURAKAMI, K., TSURUMI, Y., KODANI, S., YOSHIDA, M., FUJIE, A. & OCHI, K. 2009. Antibacterial discovery in actinomycetes strains with mutations in RNA polymerase or ribosomal protein S12. *Nat Biotechnol,* 27**,** 462-4.
HOTCHKISS, R. D., DUBOS R.J. 1940. Fractionation of the bactericidal agent from cultures of a soil bacillus. *Journal of Biological Chemistry,* 132**,** 791-792.
HOYER, K. M., MAHLERT, C. & MARAHIEL, M. A. 2007. The iterative gramicidin s thioesterase catalyzes peptide ligation and cyclization. *Chem Biol,* 14**,** 13-22.
HUR, G. H., VICKERY, C. R. & BURKART, M. D. 2012. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat Prod Rep,* 29**,** 1074-98.
ISHIYAMA, N., CREUZENET, C., MILLER, W. L., DEMENDI, M., ANDERSON, E. M., HARAUZ, G., LAM, J. S. & BERGHUIS, A. M. 2006. Structural studies of FlaA1 from Helicobacter pylori reveal the mechanism for inverting 4,6-dehydratase activity. *J Biol Chem,* 281**,** 24489-95.
JAITZIG, J., LI, J., SUSSMUTH, R. D. & NEUBAUER, P. 2014. Reconstituted biosynthesis of the nonribosomal macrolactone antibiotic valinomycin in Escherichia coli. *ACS Synth Biol,* 3**,** 432-8.
JAREMKO, M. J., LEE, D. J., OPELLA, S. J. & BURKART, M. D. 2015. Structure and substrate sequestration in the pyoluteorin type II peptidyl carrier protein PltL. *J Am Chem Soc,* 137**,** 11546-9.
JIN, M., FISCHBACH, M. A. & CLARDY, J. 2006. A biosynthetic gene cluster for the acetyl-CoA carboxylase inhibitor andrimid. *J Am Chem Soc,* 128**,** 10660-1.
JULIEN, B., SHAH, S., ZIERMANN, R., GOLDMAN, R., KATZ, L. & KHOSLA, C. 2000. Isolation and characterization of the epothilone biosynthetic gene cluster from Sorangium cellulosum. *Gene,* 249**,** 153-60.
KATZENELLENBOGEN, E., ROMANOWSKA, E., KOCHAROVA, N. A., SHASHKOV, A. S., KNIREL, Y. A. & KOCHETKOV, N. K. 1995. Structure of the O-specific polysaccharide of Hafnia alvei 1204 containing 3,6-dideoxy-3-formamido-D-glucose. *Carbohydr Res,* 273**,** 187-95.
KEATING, T. A., EHMANN, D. E., KOHLI, R. M., MARSHALL, C. G., TRAUGER, J. W. & WALSH, C. T. 2001. Chain Termination Steps in Nonribosomal Peptide Synthetase Assembly Lines: Directed Acyl-S-Enzyme Breakdown in Antibiotic and Siderophore Biosynthesis. *ChemBioChem,* 2**,** 99-107.
KEATING, T. A., MARSHALL, C. G., WALSH, C. T. & KEATING, A. E. 2002. The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nat Struct Biol,* 9**,** 522-6.
KEGLER, C., NOLLMANN, F. I., AHRENDT, T., FLEISCHHACKER, F., BODE, E. & BODE, H. B. 2014. Rapid determination of the amino acid configuration of xenotetrapeptide. *Chembiochem,* 15**,** 826-8.

KENYON, J. J., MARZAIOLI, A. M., DE CASTRO, C. & HALL, R. M. 2015. 5,7-di-N-acetyl-acinetaminic acid: A novel non-2-ulosonic acid found in the capsule of an Acinetobacter baumannii isolate. *Glycobiology,* 25**,** 644-54.

KENYON, J. J., NOTARO, A., HSU, L. Y., DE CASTRO, C. & HALL, R. M. 2017. 5,7-Di-N-acetyl-8-epiacinetaminic acid: A new non-2-ulosonic acid found in the K73 capsule produced by an Acinetobacter baumannii isolate from Singapore. *Sci Rep,* 7**,** 11357.

KESSLER, N., SCHUHMANN, H., MORNEWEG, S., LINNE, U. & MARAHIEL, M. A. 2004. The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. *J Biol Chem,* 279**,** 7413-9.

KITTILA, T., MOLLO, A., CHARKOUDIAN, L. K. & CRYLE, M. J. 2016. New Structural Data Reveal the Motion of Carrier Proteins in Nonribosomal Peptide Synthesis. *Angew Chem Int Ed Engl,* 55**,** 9834-40.

KOBEL, H. & TRABER, R. 1982. Directed biosynthesis of Cyclosporins. *European journal of applied microbiology and biotechnology,* 14**,** 237-240.

KOGLIN, A., LOHR, F., BERNHARD, F., ROGOV, V. V., FRUEH, D. P., STRIETER, E. R., MOFID, M. R., GUNTERT, P., WAGNER, G., WALSH, C. T., MARAHIEL, M. A. & DOTSCH, V. 2008. Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. *Nature,* 454**,** 907-11.

KOGLIN, A., MOFID, M. R., LOHR, F., SCHAFER, B., ROGOV, V. V., BLUM, M. M., MITTAG, T., MARAHIEL, M. A., BERNHARD, F. & DOTSCH, V. 2006. Conformational switches modulate protein interactions in peptide antibiotic synthetases. *Science,* 312**,** 273-6.

KOHLI, R. M., TAKAGI, J. & WALSH, C. T. 2002. The thioesterase domain from a nonribosomal peptide synthetase as a cyclization catalyst for integrin binding peptides. *Proc Natl Acad Sci U S A,* 99**,** 1247-52.

KOHLI, R. M. & WALSH, C. T. 2003. Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chem Commun (Camb)***,** 297-307.

KOKETSU, K., MINAMI, A., WATANABE, K., OGURI, H. & OIKAWA, H. 2012. Pictet-Spenglerase involved in tetrahydroisoquinoline antibiotic biosynthesis. *Curr Opin Chem Biol,* 16**,** 142-9.

KONDAKOVA, A. N., KIRSHEVA, N. A., SHASHKOV, A. S., SHAIKHUTDINOVA, R. Z., ARBATSKY, N. P., IVANOV, S. A., ANISIMOV, A. P. & KNIREL, Y. A. 2012. Structure of the O-polysaccharide of Photorhabdus luminescens subsp. laumondii containing D-glycero-D-manno-heptose and 3,6-dideoxy-3-formamido-D-glucose. *Carbohydr Res,* 351**,** 134-7.

KONZ, D., KLENS, A., SCHORGENDORFER, K. & MARAHIEL, M. A. 1997. The bacitracin biosynthesis operon of Bacillus licheniformis ATCC 10716: molecular characterization of three multi-modular peptide synthetases. *Chem Biol,* 4**,** 927-37.

KRAAS, F. I., GIESSEN, T. W. & MARAHIEL, M. A. 2012. Exploring the mechanism of lipid transfer during biosynthesis of the acidic lipopeptide antibiotic CDA. *FEBS Lett,* 586**,** 283-8.

KRIES, H., NIQUILLE, D. L. & HILVERT, D. 2015. A subdomain swap strategy for reengineering nonribosomal peptides. *Chem Biol,* 22**,** 640-8.

KUHLENKOETTER, S., WINTERMEYER, W. & RODNINA, M. V. 2011. Different substrate-dependent transition states in the active site of the ribosome. *Nature,* 476**,** 351-4.

LABBY, K. J., WATSULA, S. G. & GARNEAU-TSODIKOVA, S. 2015. Interrupted adenylation domains: unique bifunctional enzymes involved in nonribosomal peptide biosynthesis. *Nat Prod Rep,* 32**,** 641-53.

LAMBALOT, R. H., GEHRING, A. M., FLUGEL, R. S., ZUBER, P., LACELLE, M., MARAHIEL, M. A., REID, R., KHOSLA, C. & WALSH, C. T. 1996. A new enzyme superfamily - the phosphopantetheinyl transferases. *Chem Biol,* 3**,** 923-36.

LASKOWSKI, R. A. 2009. PDBsum new things. *Nucleic Acids Res,* 37**,** D355-9.

LAWEN, A. & ZOCHER, R. 1990. Cyclosporin synthetase. The most complex peptide synthesizing multienzyme polypeptide so far described. *J Biol Chem,* 265**,** 11355-60.

LAWRENCE, J. G. & ROTH, J. R. 1996. Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. *Genetics,* 143**,** 1843-1860.

LEDUC, D., BATTESTI, A. & BOUVERET, E. 2007. The hotdog thioesterase EntH (YbdB) plays a role in vivo in optimal enterobactin biosynthesis by interacting with the ArCP domain of EntB. *J Bacteriol,* 189**,** 7112-26.

LEE, T. V., JOHNSON, L. J., JOHNSON, R. D., KOULMAN, A., LANE, G. A., LOTT, J. S. & ARCUS, V. L. 2010. Structure of a eukaryotic nonribosomal peptide synthetase adenylation domain that activates a large hydroxamate amino acid in siderophore biosynthesis. *J Biol Chem,* 285**,** 2415-27.

LESLIE, A. G. W. & POWELL, H. R. 2007. Processing diffraction data with MOSFLM. *Evolving Methods for Macromolecular Crystallography,* 245**,** 41-51.

LI, L., DENG, W., SONG, J., DING, W., ZHAO, Q. F., PENG, C., SONG, W. W., TANG, G. L. & LIU, W. 2008. Characterization of the saframycin A gene cluster from Streptomyces lavendulae NRRL 11002 revealing a nonribosomal peptide synthetase system for assembling the unusual tetrapeptidyl skeleton in an iterative manner. *J Bacteriol,* 190**,** 251-63.

LI, Z., HWANG, S., ERICSON, J., BOWLER, K. & BAR-PELED, M. 2015. Pen and Pal Are Nucleotide-Sugar Dehydratases That Convert UDP-GlcNAc to UDP-6-Deoxy-d-GlcNAc-5,6-ene and Then to UDP-4-Keto-6-deoxy-l-AltNAc for CMP-Pseudaminic Acid Synthesis in Bacillus thuringiensis. *Journal of Biological Chemistry,* 290**,** 691-704.

LIN, S., VAN LANEN, S. G. & SHEN, B. 2007. Regiospecific chlorination of (S)-beta-tyrosyl-S-carrier protein catalyzed by SgcC3 in the biosynthesis of the enediyne antitumor antibiotic C-1027. *J Am Chem Soc,* 129**,** 12432-8.

LING, L. L., SCHNEIDER, T., PEOPLES, A. J., SPOERING, A. L., ENGELS, I., CONLON, B. P., MUELLER, A., SCHABERLE, T. F., HUGHES, D. E., EPSTEIN, S., JONES, M., LAZARIDES, L., STEADMAN, V. A., COHEN, D. R., FELIX, C. R., FETTERMAN, K. A., MILLETT, W. P., NITTI, A. G., ZULLO, A. M., CHEN, C. & LEWIS, K. 2015. A new antibiotic kills pathogens without detectable resistance. *Nature,* 517**,** 455-9.

LINNE, U. & MARAHIEL, M. A. 2000. Control of directionality in nonribosomal peptide synthesis: role of the condensation domain in preventing misinitiation and timing of epimerization. *Biochemistry,* 39**,** 10439-47.

LIU, Y. & BRUNER, S. D. 2007a. Rational manipulation of carrier-domain geometry in nonribosomal peptide synthetases. *Chembiochem,* 8**,** 617-21.

LIU, Y., ZHENG, T. & BRUNER, S. D. 2011. Structural basis for phosphopantetheinyl carrier domain interactions in the terminal module of nonribosomal peptide synthetases. *Chem Biol,* 18**,** 1482-8.

LOHMAN, J. R., MA, M., CUFF, M. E., BIGELOW, L., BEARDEN, J., BABNIGG, G., JOACHIMIAK, A., PHILLIPS, G. N. & SHEN, B. 2014. The crystal structure of BlmI as a model for nonribosomal peptide synthetase peptidyl carrier proteins. *Proteins,* 82**,** 1210-1218.

LOSEY, H. C., PECZUH, M. W., CHEN, Z., EGGERT, U. S., DONG, S. D., PELCZER, I., KAHNE, D. & WALSH, C. T. 2001. Tandem action of glycosyltransferases in the maturation of vancomycin and teicoplanin aglycones: novel glycopeptides. *Biochemistry,* 40**,** 4745-55.

LUO, L., BURKART, M. D., STACHELHAUS, T. & WALSH, C. T. 2001. Substrate recognition and selection by the initiation module PheATE of gramicidin S synthetase. *J Am Chem Soc,* 123**,** 11208-18.

MAGARVEY, N., FORTIN, P., THOMAS, P., KELLEHER, N. & WALSH, C. 2008. Gatekeeping versus promiscuity in the early stages of the andrimid biosynthetic assembly line. *ACS Chem Biol,* 3**,** 542-54.

MAGARVEY, N. A., EHLING-SCHULZ, M. & WALSH, C. T. 2006. Characterization of the cereulide NRPS alpha-hydroxy acid specifying modules: activation of alpha-keto acids and chiral reduction on the assembly line. *J Am Chem Soc,* 128**,** 10698-9.

MAO, H., HART, S. A., SCHINK, A. & POLLOK, B. A. 2004. Sortase-mediated protein ligation: a new method for protein engineering. *J Am Chem Soc,* 126**,** 2670-1.

MARAHIEL, M. A. 2016. A structural model for multimodular NRPS assembly lines. *Nat Prod Rep,* 33**,** 136-40.

MARKS, D. S., COLWELL, L. J., SHERIDAN, R., HOPF, T. A., PAGNANI, A., ZECCHINA, R. & SANDER, C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One,* 6**,** e28766.

MARSHALL, C. G., HILLSON, N. J. & WALSH, C. T. 2002. Catalytic mapping of the vibriobactin biosynthetic enzyme VibF. *Biochemistry,* 41**,** 244-50.

MAY, J. J., KESSLER, N., MARAHIEL, M. A. & STUBBS, M. T. 2002. Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A,* 99**,** 12120-5.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *J Appl Crystallogr,* 40**,** 658-674.

MCNALLY DAVID, J., SCHOENHOFEN IAN, C., MULROONEY ERIN, F., WHITFIELD DENNIS, M., VINOGRADOV, E., LAM JOSEPH, S., LOGAN SUSAN, M. & BRISSON, J. R. 2006. Identification of Labile UDP-Ketosugars in Helicobacter pylori, Campylobacter jejuni and Pseudomonas aeruginosa: Key Metabolites used to make Glycan Virulence Factors. *ChemBioChem,* 7**,** 1865-1868.

MCNALLY, D. J., HUI, J. P., AUBRY, A. J., MUI, K. K., GUERRY, P., BRISSON, J. R., LOGAN, S. M. & SOO, E. C. 2006. Functional characterization of the flagellar glycosylation locus in Campylobacter jejuni 81-176 using a focused metabolomics approach. *J Biol Chem,* 281**,** 18489-98.

MEIER, J. L. & BURKART, M. D. 2009. The chemical biology of modular biosynthetic enzymes. *Chem Soc Rev,* 38**,** 2012-45.

MIAO, V., COEFFET-LE GAL, M. F., NGUYEN, K., BRIAN, P., PENN, J., WHITING, A., STEELE, J., KAU, D., MARTIN, S., FORD, R., GIBSON, T., BOUCHARD, M., WRIGLEY, S. K. & BALTZ, R. H. 2006. Genetic engineering in Streptomyces roseosporus to produce hybrid lipopeptide antibiotics. *Chem Biol,* 13**,** 269-76.

MIAO, Y. & MCCAMMON, J. A. 2017. Chapter Six - Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *In:* DIXON, D. A. (ed.) *Annual Reports in Computational Chemistry.* Elsevier.

MILLER, B. R., DRAKE, E. J., SHI, C., ALDRICH, C. C. & GULICK, A. M. 2016. Structures of a Nonribosomal Peptide Synthetase Module Bound to MbtH-like Proteins Support a Highly Dynamic Domain Architecture. *J Biol Chem,* 291**,** 22559-22571.

MILLER, B. R. & GULICK, A. M. 2016. Structural Biology of Nonribosomal Peptide Synthetases. *Methods Mol Biol,* 1401**,** 3-29.

MISHRA, P. K. & DRUECKHAMMER, D. G. 2000. Coenzyme A analogues and derivatives: synthesis and applications as mechanistic probes of coenzyme A ester-utilizing enzymes. *Chemical Reviews,* 100**,** 3283-3310.

MITCHELL, C. A., SHI, C., ALDRICH, C. C. & GULICK, A. M. 2012. Structure of PA1221, a nonribosomal peptide synthetase containing adenylation and peptidyl carrier protein domains. *Biochemistry,* 51**,** 3252-63.

MOOTZ, H. D., KESSLER, N., LINNE, U., EPPELMANN, K., SCHWARZER, D. & MARAHIEL, M. A. 2002. Decreasing the ring size of a cyclic nonribosomal peptide antibiotic by in-frame module deletion in the biosynthetic genes. *J Am Chem Soc,* 124**,** 10980-1.

MORCOS, F., PAGNANI, A., LUNT, B., BERTOLINO, A., MARKS, D. S., SANDER, C., ZECCHINA, R., ONUCHIC, J. N., HWA, T. & WEIGT, M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A,* 108**,** E1293-301.

MORI, S., PANG, A. H., LUNDY, T. A., GARZAN, A., TSODIKOV, O. V. & GARNEAU-TSODIKOVA, S. 2018. Structural basis for backbone N-methylation by an interrupted adenylation domain. *Nature Chemical Biology*.

MORIARTY, N. W., GROSSE-KUNSTLEVE, R. W. & ADAMS, P. D. 2009. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr D Biol Crystallogr,* 65**,** 1074-80.

MORRISON, J. P., SCHOENHOFEN, I. C. & TANNER, M. E. 2008. Mechanistic studies on PseB of pseudaminic acid biosynthesis: a UDP-N-acetylglucosamine 5-inverting 4,6-dehydratase. *Bioorg Chem,* 36**,** 312-20.

NAZI, I., KOTEVA, K. P. & WRIGHT, G. D. 2004. One-pot chemoenzymatic preparation of coenzyme A analogues. *Anal Biochem,* 324**,** 100-5.

NGUYEN, K. T., RITZ, D., GU, J. Q., ALEXANDER, D., CHU, M., MIAO, V., BRIAN, P. & BALTZ, R. H. 2006. Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc Natl Acad Sci U S A,* 103**,** 17462-7.

NIELSEN, S. S., TOFT, K. N., SNAKENBORG, D., JEPPESEN, M. G., JACOBSEN, J. K., VESTERGAARD, B., KUTTER, J. P. & ARLETH, L. 2009. BioXTAS RAW, a software program for high-throughput automated small-angle X-ray scattering data reduction and preliminary analysis. *Journal of Applied Crystallography,* 42**,** 959-964.

OTWINOWSKI, Z. & MINOR, W. 1997. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol,* 276**,** 307-26.

OVCHINNIKOV, S., KAMISETTY, H. & BAKER, D. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife,* 3**,** e02030.

OVCHINNIKOV, S., KINCH, L., PARK, H., LIAO, Y., PEI, J., KIM, D. E., KAMISETTY, H., GRISHIN, N. V. & BAKER, D. 2015. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife,* 4**,** e09248.

PATEL, H. M. & WALSH, C. T. 2001. In vitro reconstitution of the Pseudomonas aeruginosa nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and N-methyltransferase activities. *Biochemistry,* 40**,** 9023-31.

PEI, J. & GRISHIN, N. V. 2014. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol,* 1079**,** 263-71.

PERRY, R. D., BALBO, P. B., JONES, H. A., FETHERSTON, J. D. & DEMOLL, E. 1999. Yersiniabactin from Yersinia pestis: biochemical characterization of the siderophore and its role in iron transport and regulation. *Microbiology,* 145 (Pt 5)**,** 1181-90.

PFEIFER, B. A., ADMIRAAL, S. J., GRAMAJO, H., CANE, D. E. & KHOSLA, C. 2001. Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science,* 291**,** 1790-2.

PLOSKON, E., ARTHUR, C. J., KANARI, A. L., WATTANA-AMORN, P., WILLIAMS, C., CROSBY, J., SIMPSON, T. J., WILLIS, C. L. & CRUMP, M. P. 2010. Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. *Chem Biol,* 17**,** 776-85.

POGLIANO, J., POGLIANO, N. & SILVERMAN, J. A. 2012. Daptomycin-mediated reorganization of membrane architecture causes mislocalization of essential cell division proteins. *J Bacteriol,* 194**,** 4494-504.

POPP, M. W., ANTOS, J. M., GROTENBREG, G. M., SPOONER, E. & PLOEGH, H. L. 2007. Sortagging: a versatile method for protein labeling. *Nat Chem Biol,* 3**,** 707-8.

QIAO, C., GUPTE, A., BOSHOFF, H. I., WILSON, D. J., BENNETT, E. M., SOMU, R. V., BARRY, C. E., 3RD & ALDRICH, C. C. 2007a. 5'-O-[(N-acyl)sulfamoyl]adenosines as antitubercular agents that inhibit MbtA: an adenylation enzyme required for siderophore biosynthesis of the mycobactins. *J Med Chem,* 50**,** 6080-94.

QIAO, C., WILSON, D. J., BENNETT, E. M. & ALDRICH, C. C. 2007b. A mechanism-based aryl carrier protein/thiolation domain affinity probe. *J Am Chem Soc,* 129**,** 6350-1.

QUADRI, L. E., WEINREB, P. H., LEI, M., NAKANO, M. M., ZUBER, P. & WALSH, C. T. 1998. Characterization of Sfp, a Bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry,* 37**,** 1585-95.

RADKOV, A. D. & MOE, L. A. 2014. Bacterial synthesis of D-amino acids. *Appl Microbiol Biotechnol,* 98**,** 5363-74.

REGER, A. S., WU, R., DUNAWAY-MARIANO, D. & GULICK, A. M. 2008. Structural characterization of a 140 degrees domain movement in the two-step reaction catalyzed by 4-chlorobenzoate:CoA ligase. *Biochemistry,* 47**,** 8016-25.

REIMER, J. M., ALOISE, M. N., HARRISON, P. M. & SCHMEING, T. M. 2016a. Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. *Nature,* 529**,** 239-42.

REIMER, J. M., ALOISE, M. N., POWELL, H. R. & SCHMEING, T. M. 2016b. Manipulation of an existing crystal form unexpectedly results in interwoven packing networks with pseudo-translational symmetry. *Acta Crystallographica Section D,* 72**,** 1130-1136.

REIMER, J. M., HAQUE, A. S., TARRY, M. J. & SCHMEING, T. M. 2018. Piecing together nonribosomal peptide synthesis. *Curr Opin Struct Biol,* 49**,** 104-113.

REIMMANN, C., PATEL, H. M., SERINO, L., BARONE, M., WALSH, C. T. & HAAS, D. 2001. Essential PchG-dependent reduction in pyochelin biosynthesis of Pseudomonas aeruginosa. *J Bacteriol,* 183**,** 813-20.

RIEGERT, A. S., CHANTIGIAN, D. P., THODEN, J. B., TIPTON, P. A. & HOLDEN, H. M. 2017. Biochemical Characterization of WbkC, an N-Formyltransferase from Brucella melitensis. *Biochemistry,* 56**,** 3657-3668.

ROBBEL, L. & MARAHIEL, M. A. 2010. Daptomycin, a bacterial lipopeptide synthesized by a nonribosomal machinery. *J Biol Chem,* 285**,** 27501-8.

ROBBINS, T., LIU, Y.-C., CANE, D. E. & KHOSLA, C. 2016. Structure and mechanism of assembly line polyketide synthases. *Current Opinion in Structural Biology,* 41**,** 10-18.

ROCHE, E. D. & WALSH, C. T. 2003. Dissection of the EntF condensation domain boundary and active site residues in nonribosomal peptide synthesis. *Biochemistry,* 42**,** 1334-44.

ROSENZWEIG, R. & KAY, L. E. 2014. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu Rev Biochem,* 83**,** 291-315.

ROSS, A. C., XU, Y., LU, L., KERSTEN, R. D., SHAO, Z., AL-SUWAILEM, A. M., DORRESTEIN, P. C., QIAN, P. Y. & MOORE, B. S. 2013. Biosynthetic multitasking facilitates thalassospiramide structural diversity in marine bacteria. *J Am Chem Soc,* 135**,** 1155-62.

ROUHIAINEN, L., PAULIN, L., SUOMALAINEN, S., HYYTIÄINEN, H., BUIKEMA, W., HASELKORN, R. & SIVONEN, K. 2000. Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in Anabaena strain 90. *Molecular Microbiology,* 37**,** 156-167.

ROUJEINIKOVA, A., SIMON, W. J., GILROY, J., RICE, D. W., RAFFERTY, J. B. & SLABAS, A. R. 2007. Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J Mol Biol,* 365**,** 135-45.

SALAH UD-DIN, A. I. M. & ROUJEINIKOVA, A. 2017. Flagellin glycosylation with pseudaminic acid in Campylobacter and Helicobacter: prospects for development of novel therapeutics. *Cell Mol Life Sci*.

SAMEL, S. A., CZODROWSKI, P. & ESSEN, L. O. 2014. Structure of the epimerization domain of tyrocidine synthetase A. *Acta Crystallogr D Biol Crystallogr,* 70**,** 1442-52.

SAMEL, S. A., SCHOENAFINGER, G., KNAPPE, T. A., MARAHIEL, M. A. & ESSEN, L. O. 2007. Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure,* 15**,** 781-92.

SCHMEING, T. M., HUANG, K. S., STROBEL, S. A. & STEITZ, T. A. 2005. An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-tRNA. *Nature,* 438**,** 520-4.

SCHMITT, E., PANVERT, M., BLANQUET, S. & MECHULAM, Y. 1998. Crystal structure of methionyl-tRNAfMet transformylase complexed with the initiator formyl-methionyl-tRNAfMet. *EMBO J,* 17**,** 6819-26.

SCHNEIDER, T. L., SHEN, B. & WALSH, C. T. 2003. Oxidase domains in epothilone and bleomycin biosynthesis: thiazoline to thiazole oxidation during chain elongation. *Biochemistry,* 42**,** 9722-30.

SCHOENAFINGER, G., SCHRACKE, N., LINNE, U. & MARAHIEL, M. A. 2006. Formylation domain: an essential modifying enzyme for the nonribosomal biosynthesis of linear gramicidin. *J Am Chem Soc,* 128**,** 7406-7.

SCHOENHOFEN, I. C., MCNALLY, D. J., BRISSON, J.-R. & LOGAN, S. M. 2006. Elucidation of the CMP-pseudaminic acid pathway in Helicobacter pylori: synthesis from UDP-N-acetylglucosamine by a single enzymatic reaction. *Glycobiology,* 16**,** 8C-14C.

SCHUBERT, H. L., BLUMENTHAL, R. M. & CHENG, X. 2003. Many paths to methyltransfer: a chronicle of convergence. *Trends Biochem Sci,* 28**,** 329-35.

SCHWARZER, D., FINKING, R. & MARAHIEL, M. A. 2003. Nonribosomal peptides: from genes to products. *Nat Prod Rep,* 20**,** 275-87.

SCHWARZER, D., MOOTZ, H. D., LINNE, U. & MARAHIEL, M. A. 2002. Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. *Proc Natl Acad Sci U S A,* 99**,** 14083-8.

SHEN, B., DU, L., SANCHEZ, C., EDWARDS, D. J., CHEN, M. & MURRELL, J. M. 2002. Cloning and characterization of the bleomycin biosynthetic gene cluster from Streptomyces verticillus ATCC15003. *J Nat Prod,* 65**,** 422-31.

SHERMAN, D. H. 2005. The Lego-ization of polyketide biosynthesis. *Nat Biotechnol,* 23**,** 1083-4.

SHI, R., LAMB, S. S., ZAKERI, B., PROTEAU, A., CUI, Q., SULEA, T., MATTE, A., WRIGHT, G. D. & CYGLER, M. 2009. Structure and function of the glycopeptide N-methyltransferase MtfA, a tool for the biosynthesis of modified glycopeptide antibiotics. *Chem Biol,* 16**,** 401-10.

SHRESTHA, S. K. & GARNEAU-TSODIKOVA, S. 2016. Expanding Substrate Promiscuity by Engineering a Novel Adenylating-Methylating NRPS Bifunctional Enzyme. *Chembiochem,* 17**,** 1328-32.

SIEBER, S. A., LINNE, U., HILLSON, N. J., ROCHE, E., WALSH, C. T. & MARAHIEL, M. A. 2002. Evidence for a monomeric structure of nonribosomal Peptide synthetases. *Chem Biol,* 9**,** 997-1008.

SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SODING, J., THOMPSON, J. D. & HIGGINS, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol,* 7**,** 539.

SILAKOWSKI, B., SCHAIRER, H. U., EHRET, H., KUNZE, B., WEINIG, S., NORDSIEK, G., BRANDT, P., BLOCKER, H., HOFLE, G., BEYER, S. & MULLER, R. 1999. New lessons for combinatorial biosynthesis from myxobacteria. The myxothiazol biosynthetic gene cluster of Stigmatella aurantiaca DW4/3-1. *J Biol Chem,* 274**,** 37391-9.

SKOU, S., GILLILAN, R. E. & ANDO, N. 2014. Synchrotron-based small-angle X-ray scattering of proteins in solution. *Nat. Protocols,* 9**,** 1727-1739.

SOCHA, A. M., LONG, R. A. & ROWLEY, D. C. 2007. Bacillamides from a hypersaline microbial mat bacterium. *Journal of Natural Products,* 70**,** 1793-1795.

STACHELHAUS, T., MOOTZ, H. D. & MARAHIEL, M. A. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol,* 6**,** 493-505.

STACHELHAUS, T. & WALSH, C. T. 2000. Mutational analysis of the epimerization domain in the initiation module PheATE of gramicidin S synthetase. *Biochemistry,* 39**,** 5775-87.

STAUNTON, J., CAFFREY, P., APARICIO, J. F., ROBERTS, G. A., BETHELL, S. S. & LEADLAY, P. F. 1996. Evidence for a double-helical structure for modular polyketide synthases. *Nat Struct Biol,* 3**,** 188-92.

SUNDARAM, S. & HERTWECK, C. 2016. On-line enzymatic tailoring of polyketides and peptides in thiotemplate systems. *Curr Opin Chem Biol,* 31**,** 82-94.

SUNDLOV, J. A. & GULICK, A. M. 2013. Structure determination of the functional domain interaction of a chimeric nonribosomal peptide synthetase from a challenging crystal with noncrystallographic translational symmetry. *Acta Crystallogr D Biol Crystallogr,* 69**,** 1482-92.

SUNDLOV, J. A., SHI, C., WILSON, D. J., ALDRICH, C. C. & GULICK, A. M. 2012. Structural and functional investigation of the intermolecular interaction between NRPS adenylation and carrier protein domains. *Chem Biol,* 19**,** 188-98.

SUO, Z., TSENG, C. C. & WALSH, C. T. 2001. Purification, priming, and catalytic acylation of carrier protein domains in the polyketide synthase and nonribosomal peptidyl synthetase modules of the HMWP1 subunit of yersiniabactin synthetase. *Proc Natl Acad Sci U S A,* 98**,** 99-104.

SVERGUN, D. 1992. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography,* 25**,** 495-503.

SVERGUN, D., BARBERATO, C. & KOCH, M. H. J. 1995. CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography,* 28**,** 768-773.

TAN, X. F., DAI, Y. N., ZHOU, K., JIANG, Y. L., REN, Y. M., CHEN, Y. & ZHOU, C. Z. 2015. Structure of the adenylation-peptidyl carrier protein didomain of the Microcystis aeruginosa microcystin synthetase McyG. *Acta Crystallogr D Biol Crystallogr,* 71**,** 873-81.

TANOVIC, A., SAMEL, S. A., ESSEN, L. O. & MARAHIEL, M. A. 2008. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science,* 321**,** 659-63.

TARRY, M. J., HAQUE, A. S., BUI, K. H. & SCHMEING, T. M. 2017. X-Ray Crystallography and Electron Microscopy of Cross- and Multi-Module Nonribosomal Peptide Synthetase Proteins Reveal a Flexible Architecture. *Structure,* 25**,** 783-793 e4.

TARRY, M. J. & SCHMEING, T. M. 2015. Specific disulfide cross-linking to constrict the mobile carrier domain of nonribosomal peptide synthetases. *Protein Eng Des Sel,* 28**,** 163-70.

TEDESCO, D. & HARAGSIM, L. 2012. Cyclosporine: a review. *J Transplant,* 2012**,** 230386.

THIRLWAY, J., LEWIS, R., NUNNS, L., AL NAKEEB, M., STYLES, M., STRUCK, A. W., SMITH, C. P. & MICKLEFIELD, J. 2012. Introduction of a non-natural amino acid into a nonribosomal peptide antibiotic by modification of adenylation domain specificity. *Angew Chem Int Ed Engl,* 51**,** 7181-4.

THODEN, J. B., GONEAU, M. F., GILBERT, M. & HOLDEN, H. M. 2013. Structure of a sugar N-formyltransferase from Campylobacter jejuni. *Biochemistry,* 52**,** 6114-26.

TOWNSLEY, L. E., TUCKER, W. A., SHAM, S. & HINTON, J. F. 2001. Structures of gramicidins A, B, and C incorporated into sodium dodecyl sulfate micelles. *Biochemistry,* 40**,** 11676-86.

TRABER, R., HOFMANN, H. & KOBEL, H. 1989. Cyclosporins--new analogues by precursor directed biosynthesis. *J Antibiot (Tokyo),* 42**,** 591-7.

TRIA, G., MERTENS, H. D. T., KACHALA, M. & SVERGUN, D. I. 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ,* 2**,** 207-217.

TUFAR, P., RAHIGHI, S., KRAAS, F. I., KIRCHNER, D. K., LOHR, F., HENRICH, E., KOPKE, J., DIKIC, I., GUNTERT, P., MARAHIEL, M. A. & DOTSCH, V. 2014. Crystal structure of a PCP/Sfp complex reveals the structural basis for carrier protein posttranslational modification. *Chem Biol,* 21**,** 552-62.

TUGARINOV, V. & KAY, L. E. 2004. An isotope labeling strategy for methyl TROSY spectroscopy. *J Biomol NMR,* 28**,** 165-72.
ULRICH, V. & CRYLE, M. J. 2017. SNaPe: a versatile method to generate multiplexed protein fusions using synthetic linker peptides for in vitro applications. *J Pept Sci,* 23**,** 16-27.
VAILLANCOURT, F. H., YIN, J. & WALSH, C. T. 2005. SyrB2 in syringomycin E biosynthesis is a nonheme FeII alpha-ketoglutarate- and O2-dependent halogenase. *Proc Natl Acad Sci U S A,* 102**,** 10111-6.
VATER, J., STEIN, T., VOLLENBROICH, D., KRUFT, V., WITTMANN-LIEBOLD, B., FRANKE, P., LIU, L. & ZUBER, P. 1997. The modular organization of multifunctional peptide synthetases. *J Protein Chem,* 16**,** 557-64.
VOLKOV, V. V. & SVERGUN, D. I. 2003. Uniqueness ofab initioshape determination in small-angle scattering. *Journal of Applied Crystallography,* 36**,** 860-864.
WALLACE, B. A. 2000. Common structural features in gramicidin and other ion channels. *Bioessays,* 22**,** 227-34.
WALLIN, G. & AQVIST, J. 2010. The transition state for peptide bond formation reveals the ribosome as a water trap. *Proc Natl Acad Sci U S A,* 107**,** 1888-93.
WALSH, C. T. 2004. Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science,* 303**,** 1805-10.
WALSH, C. T., CHEN, H., KEATING, T. A., HUBBARD, B. K., LOSEY, H. C., LUO, L., MARSHALL, C. G., MILLER, D. A. & PATEL, H. M. 2001. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Curr Opin Chem Biol,* 5**,** 525-34.
WANG, H., FEWER, D. P., HOLM, L., ROUHIAINEN, L. & SIVONEN, K. 2014a. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc Natl Acad Sci U S A,* 111**,** 9259-64.
WANG, K., SACHDEVA, A., COX, D. J., WILF, N. M., LANG, K., WALLACE, S., MEHL, R. A. & CHIN, J. W. 2014b. Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nature Chemistry,* 6**,** 393.
WANG, L., BROCK, A., HERBERICH, B. & SCHULTZ, P. G. 2001. Expanding the Genetic Code of <em>Escherichia coli</em&gt. *Science,* 292**,** 498.
WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.
WEBER, G., SCHÖRGENDORFER, K., SCHNEIDER-SCHERZER, E. & LEITNER, E. 1994. The peptide synthetase catalyzing cyclosporine production in Tolypocladium niveum is encoded by a giant 45.8-kilobase open reading frame. *Current Genetics,* 26**,** 120-125.
WEBER, T., BAUMGARTNER, R., RENNER, C., MARAHIEL, M. A. & HOLAK, T. A. 2000. Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure,* 8**,** 407-18.
WEINIG, S., HECHT, H. J., MAHMUD, T. & MULLER, R. 2003. Melithiazol biosynthesis: further insights into myxobacterial PKS/NRPS systems and evidence for a new subclass of methyl transferases. *Chem Biol,* 10**,** 939-52.
WEISSMAN, K. J. 2015. The structural biology of biosynthetic megaenzymes. *Nat Chem Biol,* 11**,** 660-70.

WENZEL, S. C., KUNZE, B., HOFLE, G., SILAKOWSKI, B., SCHARFE, M., BLOCKER, H. & MULLER, R. 2005. Structure and biosynthesis of myxochromides S1-3 in Stigmatella aurantiaca: evidence for an iterative bacterial type I polyketide synthase and for module skipping in nonribosomal peptide biosynthesis. *Chembiochem,* 6**,** 375-85.

WHITFIELD, C. & TRENT, M. S. 2014. Biosynthesis and export of bacterial lipopolysaccharides. *Annu Rev Biochem,* 83**,** 99-128.

WILLIAMS, G. J., BREAZEALE, S. D., RAETZ, C. R. & NAISMITH, J. H. 2005. Structure and function of both domains of ArnA, a dual function decarboxylase and a formyltransferase, involved in 4-amino-4-deoxy-L-arabinose biosynthesis. *J Biol Chem,* 280**,** 23000-8.

WINN, M., FYANS, J. K., ZHUO, Y. & MICKLEFIELD, J. 2016. Recent advances in engineering nonribosomal peptide assembly lines. *Nat Prod Rep,* 33**,** 317-47.

WINTER, G., WATERMAN, D. G., PARKHURST, J. M., BREWSTER, A. S., GILDEA, R. J., GERSTEL, M., FUENTES-MONTERO, L., VOLLMAR, M., MICHELS-CLARK, T., YOUNG, I. D., SAUTER, N. K. & EVANS, G. 2018. DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr D Struct Biol,* 74**,** 85-97.

WOODFORD, C. R., THODEN, J. B. & HOLDEN, H. M. 2015. New role for the ankyrin repeat revealed by a study of the N-formyltransferase from Providencia alcalifaciens. *Biochemistry,* 54**,** 631-8.

WOODFORD, C. R., THODEN, J. B. & HOLDEN, H. M. 2017. Molecular architecture of an N-formyltransferase from Salmonella enterica O60. *J Struct Biol,* 200**,** 267-278.

WORTHINGTON, A. S. & BURKART, M. D. 2006. One-pot chemo-enzymatic synthesis of reporter-modified proteins. *Org Biomol Chem,* 4**,** 44-6.

WORTHINGTON, A. S., RIVERA, H., TORPEY, J. W., ALEXANDER, M. D. & BURKART, M. D. 2006. Mechanism-based protein cross-linking probes to investigate carrier protein-mediated biosynthesis. *ACS Chem Biol,* 1**,** 687-91.

XU, Y., OROZCO, R., WIJERATNE, E. M., GUNATILAKA, A. A., STOCK, S. P. & MOLNAR, I. 2008. Biosynthesis of the cyclooligomer depsipeptide beauvericin, a virulence factor of the entomopathogenic fungus Beauveria bassiana. *Chem Biol,* 15**,** 898-907.

YAMADA, M. & KURAHASHI, K. 1968. Adenosine triphosphate and pyrophosphate dependent phenylalanine racemase of Bacillus brevis Nagano. *J Biochem,* 63**,** 59-69.

YANG, W. & DRUECKHAMMER, D. G. 2000. Computational studies of the aminolysis of oxoesters and thioesters in aqueous solution. *Org Lett,* 2**,** 4133-6.

YEH, E., COLE, L. J., BARR, E. W., BOLLINGER, J. M., JR., BALLOU, D. P. & WALSH, C. T. 2006. Flavin redox chemistry precedes substrate chlorination during the reaction of the flavin-dependent halogenase RebH. *Biochemistry,* 45**,** 7904-12.

YONUS, H., NEUMANN, P., ZIMMERMANN, S., MAY, J. J., MARAHIEL, M. A. & STUBBS, M. T. 2008. Crystal structure of DltA. Implications for the reaction mechanism of non-ribosomal peptide synthetase adenylation domains. *J Biol Chem,* 283**,** 32484-91.

ZHANG, J., LIU, N., CACHO, R. A., GONG, Z., LIU, Z., QIN, W., TANG, C., TANG, Y. & ZHOU, J. 2016. Structural basis of nonribosomal peptide macrocyclization in fungi. *Nat Chem Biol,* 12**,** 1001-1003.

ZHAO, C., COUGHLIN, J. M., JU, J., ZHU, D., WENDT-PIENKOWSKI, E., ZHOU, X., WANG, Z., SHEN, B. & DENG, Z. 2010. Oxazolomycin biosynthesis in Streptomyces albus JA3453 featuring

an "acyltransferase-less" type I polyketide synthase that incorporates two distinct extender units. *J Biol Chem,* 285**,** 20097-108.

ZIMMER, A. L., THODEN, J. B. & HOLDEN, H. M. 2014. Three-dimensional structure of a sugar N-formyltransferase from Francisella tularensis. *Protein Sci,* 23**,** 273-83.