Structural characterization of iterative polyketide synthases

Tan, Jackie Yen

2020

Tan, J. Y. (2020). Structural characterization of iterative polyketide synthases. Doctoral thesis, Nanyang Technological University, Singapore

<https://hdl.handle.net/10356/145736>

<https://doi.org/10.32657/10356/145736>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 18:55:21 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

Structural Characterization of Iterative Polyketide Synthases

Jackie Tan Yen

SCHOOL OF BIOLOGICAL SCIENCES

2020

Structural Characterization of Iterative Polyketide Synthase

Jackie Tan Yen

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological
University in partial fulfilment of the requirement for the
degree of Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined.

I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

.....23/1/2020.
Date

.....
Jackie Tan Yen

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

2020-01-21

.....
Date

.....
Dr Gao Yonggui

Authorship Attribution Statement

*(A) This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

.....21/1/2020.
Date

.....
Jackie Tan Yen

Acknowledgement

First and foremost, I would like to express my sincere gratitude to Prof Gao for many reasons. He first took a chance in me as a project officer when we first met 6 years ago, without any reservation of doubt. Two years later, he generously extended an offer to be a PhD candidate in his lab. Throughout it all, he has been nothing but a generous, knowledgeable, and patient mentor and friend. Thank you for the invaluable guidance, inspiration, and faith in me and my abilities. I would also like to express my sincere gratitude to Dr Liang Zhao-xun for being a close collaborator and support throughout my entire journey here. In addition, I'd like to thank members of my Thesis Advisory Committee as well for taking time off their busy schedule to evaluate my candidate – Dr Yang Liang, Dr Liang Zhao-xun, and Dr Scott Rice.

The journey to PhD has not been easy, and it was only thanks to everyone around me in the Gao Lab that I was able to come so far. Special thanks to also, 1) my coffee buddy, cat lover, and meme dealer, Rya, for being a wonderful, generous, and big-hearted labmate, 2) Xin Fu for being a fellow rap and cat enthusiast and salvager of my failed projects, 3) Chairman Qiao for being a steadfast lab seminar partner and office neighbour, 4) Wenting for teaching me the true power of Western Blot and taking good care of the lab, 5) Shengyang for being a generally cool helpful aloof guy and a source of scientific inspiration, and 6) KJ + Zhiyang for our snack breaks and for shooting the breeze. I'd also like to express my appreciation to my current and past labmates for the enjoyable experience.

Big thanks to my other collaborators, Dr Shashi Bhushan for teaching me the joy of electron microscopy – I will miss the times we spent talking about EM structures side-by-side, Dr Andrew Wong for his endless enthusiasm and help with all things EM, Prof Gruber and Dr

Sony for her help with SAXS analysis, Prof Newman for offering his cross-linking mass spectrometry expertise, and Dr Lu Lanyuan for his expertise on molecular dynamics.

Saving the best for the last, I would like to thank my partner, best friend, cute cat, pillar of support, confidante, and mother to my best boys Char Siew Mao and Siew Long Bao – Michelle. Without you, I would not have achieved as much as I did nor wanted, and finally I understand the saying that behind a successful man is a woman supporting him.

Table of contents

Acknowledgement.....	i
Table of contents.....	iii
List of figures.....	vii
List of tables.....	xii
List of abbreviations.....	xiii
Abstract.....	xvii
Chapter 1: Generalized predictive model for polyketide chain length synthesized by iterative polyketide synthases.....	1
1.1 Introduction.....	1
1.2 Aims and objectives.....	4
1.3 Materials and methods.....	5
1.3.1 Dataset.....	5
1.3.2 KS domain prediction and phylogenetic analysis.....	5
1.3.3 KS domain modelling.....	6
1.3.4 Active site analysis.....	6
1.3.5 Statistical analysis and machine learning.....	6
1.4 Results.....	7
1.4.1 iPKS dataset.....	7
1.4.2 Phylogenetic tree of iPKS KS domains.....	10
1.4.3 KS domain modelling.....	12
1.4.4 Exploratory data analysis of KS domains.....	15
1.4.5 Bivariate analyses in KS domain biochemical properties.....	24
1.4.6 Multiple linear regression analyses.....	27
1.4.7 Machine learning modelling to predict polyketide product length.....	31

1.5 Discussion.....	34
Chapter 2: Structural characterization of partially reducing iterative polyketide synthase NcsB.....	
2.1 Introduction.....	37
2.2 Aims and significance.....	42
2.3 Materials and methods.....	43
2.3.1 Reagents and chemicals.....	43
2.3.2 Expression and purification of NcsB.....	43
2.3.3 Small-angle X-ray scattering.....	44
2.3.4 Dynamic light scattering.....	45
2.3.5 Negative stain EM grid preparation and staining.....	45
2.3.6 2D and 3D reconstruction of NcsB.....	46
2.3.7 Homology domain generation and fitting.....	47
2.3.8 Cross-linking mass spectrometry of NcsB.....	47
2.4 Results.....	48
2.4.1 Purification of NcsB.....	48
2.4.2 SAXS and DLS analysis.....	52
2.4.3 Negative stain EM data collection.....	56
2.4.4 Negative stain EM data processing.....	58
2.4.5 NcsB dimer initial model reconstruction.....	60
2.4.6 NcsB monomer initial model reconstruction.....	61
2.4.7 NcsB dimer and monomer negative stain EM structures.....	63
2.4.8 Homology modelling and domain arrangement.....	66
2.4.9 Cross-linking mass spectrometry.....	68
2.5 Discussion.....	73

Chapter 3: Structural characterization of highly reducing iterative polyketide

synthase DynE8.....	77
3.1 Introduction.....	77
3.2 Aims and significance.....	79
3.3 Materials and methods.....	81
3.3.1 Reagents and chemicals.....	81
3.3.2 Expression and purification of DynE8 and DynU15.....	81
3.3.3 Gradient Fixation (GraFix).....	82
3.3.4 Negative stain EM grid preparation and staining	82
3.3.5 2D and 3D reconstruction of DynE8 based on negative stain EM.....	82
3.3.6 Cryo-EM grid preparation and data collection.....	83
3.3.7 Cryo-EM data processing.....	84
3.4 Results.....	85
3.4.1 Expression and purification of DynE.....	85
3.4.2 Expression and purification of DynU15.....	86
3.4.3 Expression and purification of co-expressed DynE8 and DynU15.....	89
3.4.4 Negative stain EM data collection of DynE8.....	91
3.4.5 Sucrose-based gradient centrifugation.....	92
3.4.6 Negative stain EM data collection of GraFix DynE8.....	94
3.4.7 Negative stain EM data processing.....	96
3.4.8 DynE8 negative stain EM structure.....	99
3.4.9 Cryo-EM screening and data collection.....	103
3.4.10 Cryo-EM data processing.....	108
3.5 Discussion.....	113

Chapter 4: Summary.....	116
4.1 Conclusion.....	116
4.2 Future work and directions.....	117
References.....	118

List of figures

Figure 1. Simplified domain organization of Type I, II, and III PKS.....	1
Figure 2. Representative domain architectures of three subtypes of iterative PKS.....	2
Figure 3. KS domain genealogy of characterized iPKSs.....	11
Figure 4. Representative homology models of the different types of KS domains.....	13
Figure 5. A small window for the Gblocks output for the 97 KS domains.....	14
Figure 6. Representative output of a model submitted into the CASTp server.....	15
Figure 7. Histogram of all cavity volumes in the dataset.....	15
Figure 8. Histograms of the distribution of the KS domain cavity volumes.....	16
Figure 9. The boxplot for the cavity volumes of NR-, PR-, and HR-KS domains.....	17
Figure 10. Histogram of the GRAVY scores of the KS domains.....	18
Figure 11. Histograms of the distribution of the KS domain GRAVY scores.....	19
Figure 12. The boxplot for the GRAVY scores of NR-, PR-, and HR-KS domains.....	20
Figure 13. Histogram of the polyketide product lengths in the dataset.....	21
Figure 14. The histograms of the distributions of PKS product length.....	22
Figure 15. The boxplot for the polyketide product lengths for NR-, PR-, and HR-PKS.....	23
Figure 16. Plot between KS domain cavity volume vs KS domain GRAVY score.....	24
Figure 17. The OLS regression for Gravy score against product length.....	25
Figure 18. The OLS regression for KS domain cavity volume against product length.....	26
Figure 19. The plots for cavity volume against product length for the three types of PKS...	26
Figure 20. The MLR of actual vs predicted polyketide chain length based on log cavity volume and GRAVY score.....	28
Figure 21. The MLR plot of actual polyketide chain length vs the predicted chain length.....	29
Figure 22. Comparison of the results of the two MLRs.....	30

Figure 23. Predicted polyketide chain lengths vs actual polyketide lengths for machine learning techniques.....	32
Figure 24. Predicted polyketide chain lengths vs actual polyketide lengths for boosted tree machine learning algorithms.....	33
Figure 25. The chemical structure of neocarzinostatin.....	39
Figure 26. Partially-reducing iPKS NcsB and its catalytic activity.....	40
Figure 27. A simplified 2D reaction cycle of NPA biosynthesis in NcsB and the domains involved.....	41
Figure 28. Nickel affinity chromatograph of NcsB + Sfp.....	48
Figure 29. Anion exchange chromatography of NcsB + Sfp.....	49
Figure 30. Gel filtration chromatograph of NcsB.....	50
Figure 31. Gel filtration chromatography using Superdex 200 10/300 GL Increase.....	51
Figure 32. Biochemical assays for NcsB activity.....	51
Figure 33. Solution X-ray experimental scattering pattern.....	53
Figure 34. Overlapping of pair-distance distribution function $P(r)$ of NcsB at the three different concentrations.....	54
Figure 35. Normalized Kratky plot of NcsB.....	55
Figure 36. DLS profile of NcsB in solution.....	56
Figure 37. Representative view of the NcsB negative stain EM grid.....	57
Figure 38. 2D classification of NcsB dimer particles.....	58
Figure 39. Representative NcsB classes with manual outline tracing.....	59
Figure 40. Representative 2D classes of NcsB monomer.....	59
Figure 41. Initial 3D model of NcsB dimer constructed using 2D classes.....	60
Figure 42. The micrograph and its tilted view.....	61
Figure 43. Representative particles with tilt pairs and 2D classes.....	62

Figure 44. Random conical tilt model reconstruction of NcsB.....	62
Figure 45. 3D reconstruction of the NcsB dimer using the dimer particles, refined to a resolution of about 18.6 Å.....	63
Figure 46. Refinement progress for dimeric NcsB using non-NcsB model.....	64
Figure 47. 3D reconstruction of the NcsB monomer using monomeric particles, refined to a resolution of 14.8 Å.....	64
Figure 48. Fourier shell correlation (FSC) graph of NcsB monomer and dimer structures...	65
Figure 49. Solid rendering of two monomeric NcsB fitted inside the meshed representation of dimeric NcsB.....	66
Figure 50. Proposed NcsB model with domains.....	67
Figure 51. General CLMS workflow.....	68
Figure 52. The amino acid sequence of NcsB with highlighted lysine (K) residues.....	69
Figure 53. SDS-PAGE results of NcsB cross-linking with DSS and DSSO.....	70
Figure 54. Visualization of the cross-linked pairs in NcsB.....	72
Figure 55. A comparison between the existing PKS structures.....	74
Figure 56. Chemical structure of dynemicin A.....	77
Figure 57. Domain arrangement of highly-reducing iPKS and DynE8.....	78
Figure 58. Nickel affinity chromatography results of DynE8.....	85
Figure 59. Gel filtration chromatography results of DynE8.....	86
Figure 60. Nickel affinity chromatography results of DynU15.....	87
Figure 61. Gel filtration chromatography results of DynU15.....	88
Figure 62. DynU15 distribution in gel filtration chromatography fractions.....	89
Figure 63. Nickel affinity chromatography results of DynE8 + DynU15 co-expression.....	90
Figure 64. Gel filtration chromatography results of DynE8 + DynU15 co-expression.....	91
Figure 65. Representative micrograph of purified DynE8.....	92

Figure 66. SDS-PAGE of post-GraFix sucrose density gradient centrifugation fractions...	93
Figure 67. SDS-PAGE of GraFix fractions for DynE8 + U15 co-expression.....	94
Figure 68. Representative micrograph of DynE8 ^{GA+}	95
Figure 69. Representative micrograph of DynE8 ^{GA-}	96
Figure 70. 2D classes of DynE8 ^{GA+} particles.....	97
Figure 71. 2D classes of DynE8 ^{GA-} particles.....	97
Figure 72. 2D class average of Wiener filtered particles.....	99
Figure 73. Representative initial models generated from the 2D classes of DynE8 particles.....	100
Figure 74. Representative initial models generated from 2D classes of DynE8 Wiener-filtered particles.....	100
Figure 75. 3D reconstruction of DynE8 particles.....	101
Figure 76. Fourier shell correlation (FSC) graph of DynE8 ^{GA+} and DynE ^{GA-} structures.....	102
Figure 77. Representative micrograph of the first cryo-EM screening.....	103
Figure 78. Representative micrograph of the second cryo-EM screening.....	104
Figure 79. SDS-PAGE gel of DynE8 ^{GA-} fractions after GraFix.....	105
Figure 80. Representative micrograph of the cyro-EM micrograph with DynE ^{GA-}	105
Figure 81. Representative micrograph of the cyro-EM micrograph with DynE ^{GA+}	106
Figure 82. Representative micrograph of the cryo-EM micrograph with higher concentrations of DynE ^{GA-}	107
Figure 83. Individual particles picked manually from micrographs.....	108
Figure 84. 2D classes of particles using EMAN2.....	109
Figure 85. 2D classes of cryo-EM ^{GA-} particles using RELION.....	109
Figure 86. A sample of the DynE8 ^{GA-} 2D classes from automated particle picking.....	110
Figure 87. 2D classes after three rounds of particle selection.....	111

Figure 88. 2D classification of the second round of DynE8 ^{GA-} dataset using EMAN2.....	112
Figure 89. An example of the 2D classes for DynE ^{GA-}	112
Figure 90. Comparison between DynE8 ^{GA-} , PikAIII, and NcsB.....	114

List of tables

Table 1 Python packages and their purpose used in current study.....	7
Table 2 List of PKS dataset used in this study.....	8
Table 3 PDB models used as template for modelling.....	12
Table 4 Summary of the exploratory data analysis.....	23
Table 5 Summary of the bivariate analyses performed between dataset features.....	30
Table 6 Summary of the model performances in this study.....	34
Table 7 NcsB and homologous bacterial iPKS.....	39
Table 8 Data collection and scattering derived parameters for NcsB.....	52
Table 9 NcsB domain homology modelling templates.....	67
Table 10 Representative cross-linked peptide sequences from the CLMS.....	71

List of abbreviations

2YT	2x YT media
6-MSA	6-methylsalicylic acid
6-MSAS	6-methylsalicylic acid synthase
B-ME	β -mercaptoethanol
Ψ KR	Pseudo-ketoreductase
ACP	Acetyl carrier protein
AT	Acyl transferase
CoA	Coenzyme A
C-MeT	C-methyltransferase
CLF	Chain length factor
CLMS	Cross-link mass spectrometry
CTF	Contrast transfer function
D_H	Hydrodynamic diameter
DH	Dehydratase
DLS	Dynamic light scattering
Dmax	Maximum particle dimension
DTT	Dithiothreitol
DSS	Disuccinimidyl suberate
DSSO	Disuccinimidyl sulfoxide
DynE8	Dynemicin polyketide synthase
DynU15	Dynemicin acetylenase
EDTA	Ethylenediaminetetra acetic acid
EM	Electron microscopy
EMDB	Electron microscopy database

ER	Enoyl reductase
FAS	Fatty acid synthase
FSC	Fourier shell correlation
GA	Glutaraldehyde
GBM	Gradient boosting machine
GraFix	Gradient fixation
GRAVY	Grand average of hydropathy
HCl	Hydrochloric acid
HPLC	High performance liquid chromatography
HR-PKS	Highly-reducing polyketide synthase
iPKS	Iterative polyketide synthase
IPTG	Isopropyl β -D-1-thiogalactopyranoside
I(q)	Scattering intensity
kDa	KiloDalton
k-NN	k-Nearest Neighbours
KR	Ketoreductase
KS	Keto-acyl synthase
kV	Kilovolt
LB	Luria-Bertani media
mFAS	Mammalian fatty acid synthase
MAS	MAS-like polyketide synthase
modPKS	Modular polyketide synthase
ML	Machine learning
MLR	Multiple linear regression
MW	Molecular weight

MWCO	Molecular weight cut-off
NADPH	Nicotinamide adenine dinucleotide phosphate
NCS	Neocarzinostatin
NcsB	Neocarzinostatin naphthoic acid polyketide synthase
NcsE	Neocarzinostatin enediyne polyketide synthase
NR-PKS	Non-reducing polyketide synthase
NRPS	Non-ribosomal peptide synthase
NS-EM	Negative stain electron microscopy
OD600	Optical density at 600 nm
OLS	Ordinary least squares
PKS	Polyketide synthase
PDB	Protein Data Bank
PT	Product template domain
PPTase	Phosphopantetheiny transferase
P(r)	Distance distribution function
R^2	Correlation coefficient
RCT	Random conical tilt
R_g	Radius of gyration
RMSE	Root mean square error
SAT	Starter acyl transferase
SAXS	Small angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
Sfp	Surfactin production
SMA	Poly(styrene-co-maleic acid)
SMANCS	Poly(styrene-co-maleic acid) neocarzinostatin

TE	Thioesterase
TH	Thiohydrolase
V_p	Porod volume
V_c	Volume of correlation

Abstract

Polyketide synthases are large multi-domain proteins found in bacterial and fungal species that synthesize structurally complex compounds called polyketides. Polyketides have strong antibacterial, antitumoral, and anticancer effect, and have therapeutic potential. Iterative polyketide synthases (iPKS) are one of the types of polyketide synthase and uses its domain iteratively to synthesize a polyketide chain product. While the individual domains in iPKSs are well studied, the full-length structures of iPKSs remain largely uncharacterized. As such, structural characterization of iPKS for the purpose of rational engineering and fine-tuning polyketide products remains an ongoing effort.

The results in this thesis discuss the use of *in silico* approaches to characterize iPKS. Firstly, the relationship between the physicochemical properties of iPKS ketosynthase domain and its product length was investigated, using multiple linear regression and machine learning. Predictive modelling of iPKS product was achieved by using catalytic site area and volume, and hydropathy scores as features. Secondly, biochemical characterization of partially-reducing iPKS NcsB was performed. Negative stain electron microscopy (EM) was used to elucidate low-resolution structures of monomeric and dimeric NcsB. The proposed domain arrangement in the low-resolution model of NcsB provides an insight into iPKS conformation and architecture. Lastly, biochememical characterization of highly-reducing DynE8 was performed. Optimization of sample preparation was carried out to improve homogeneity for EM experiments. Improved samples of DynE8 was obtained using sucrose gradient ultracentrifugation, which advances efforts in future characterization of iPKS. Negative stain EM models of DynE8 and its overall conformation are shown.

Chapter 1: Generalized predictive model for polyketide chain length synthesized by iterative type I polyketide synthases

1.1 Introduction

Fungi are remarkable organisms that synthesize a variety of structurally complex and bioactive compounds named polyketides. Polyketides have valuable pharmaceutical properties and are often antimicrobial, e.g., doxycycline and erythromycin¹. These polyketide compounds are synthesized by proteins named polyketide synthases (PKSs), a family of enzymes that work in concert. PKSs are split into three main types – Type I, Type II, and Type III PKSs.

Figure 1. Simplified domain organization of Type I, II, and III PKS. The PKS domains are represented as circles. In modular Type I PKSs, groups of domains for a module with each module responsible for extending the polyketide chain. In iterative Type I PKSs, the domains are arranged in a single module, and each domain is used iteratively to extend the polyketide chain. Type II PKSs are dissociable proteins that contain single domains that work iteratively to extend the polyketide chain. Type III PKSs are similarly iterative, without using ACP. KS: ketosynthase, AT: acyltransferase, ACP: acyl carrier protein, KR: ketoreductase, DH: dehydratase, ER: enoyl reductase, TE: thioesterase, CoA: Coenzyme A.

Type I PKS is a group of large (>150 kDa) multidomain proteins and is further classified into modular and iterative polyketide synthases (iPKS). Compared to modular PKSs, which work similarly to factory lines where the polyketide chain grows and is modified as it is passed from one protein module to the next, iPKSs are large proteins containing domains that work iteratively to extend the polyketide chain². Type II PKSs are dissociable single-domain and independent enzymes that work iteratively for polyketide synthesis³. During chain lengthening, tailoring enzymes modify parts of the polyketide chain, e.g., methyltransferase, oxygenase, reductase, which introduce diversity in the end polyketide product. Type III PKSs are composed of homodimer ketosynthase domains which are responsible for iterative chain extension⁴. Type III PKSs can work with both coenzyme A-tethered and ACP-tethered extender units to extend the polyketide chain.

Figure 2. Representative domain architectures of three subtypes of iterative PKS. (A) Non-reducing iPKS generally contains starter unit ACP transacylase (SAT), ketosynthase (KS), malonyl-acetyl transferase (MAT), product template domain (PT), acyl carrier protein (ACP), and thioesterase/Claisen cyclase (TE/CLC). (B) Partially-reducing iPKS generally contains KS, acyl transferase (AT), dehydratase (DH), ketoreductase (KR), and ACP. (C) Highly-reducing iPKS generally contains KS, MAT, DH, methyltransferase (MT), enoyl reductase (ER), KR, and ACP.

There are three subtypes of iPKSs based on the products synthesized and phylogeny: non-reducing PKS (NR-PKS), partially-reducing PKS (PR-PKS), and highly-reducing PKS (HR-PKS). The three types of iPKSs share several core domains; ketosynthase (KS), acyltransferase (AT), and acyl carrier protein (ACP). They then differ with additional domains for additional polyketide chain modifications. NR-PKSs do not have product reduction ability, but instead have domains such as starter unit-ACP transacylase (SAT), product template (PT), and thioesterase (TE). Both PR-PKS and HR-PKS have ketoreductase (KR) for product ketoreduction, and a combination of dehydratase (DH), thiohydrolase (TH), and enoyl reductase (ER) domains.

Given the structural diversity of iPKSs secondary metabolites, there is great interest in understanding the enzyme mechanism for rational design of synthetic metabolites. Over the last few years, bioinformatics tools have been used to perform secondary metabolite genome mining and to predict the chemical structures of NRPS, modular type I and type II polyketides such as PRISM³, TransATor⁶, NaPDOS⁷, and antiSMASH⁸. Phylogenetic-based product prediction has been attempted for NRPS/PKS-NRPS (non-ribosomal peptide synthase) hybrids as well⁹. However, these tools are not suitable for the predictions for iPKS products.

Several factors are understood to affect polyketide backbone length. This includes the ketosynthase chain length factor (KS-CLF) in Type II PKSs¹⁰, the starter unit acyltransferases (SAT) in NR-PKS¹¹, and combination of three domains: C-methyltransferase (C-MeT), pseudo-ketoreductase (Ψ KR), and ketoreductase (KR) in HR-PKS-NRPS hybrids¹². Liu *et al*¹³ discovered that there was a correlation between NR-PKS product length and the thioesterase, reductase, and metallo- β -lactamase-type TE domains of NR-PKS, with a

correlation coefficient of 0.75, or a R^2 value of 0.563. While the correlation was useful in predicting lengths of polyketide intermediates, the domains analyzed in these studies are not found across the different reducing types of type I PKSs. NR-PKS domain swapping experiments have also supported the role of KS domain in chain length control¹⁴⁻¹⁷.

Furthermore, there was correlation found between the number of iterations performed by KS domains and the cavity volumes¹⁸. Using KS domain physicochemical properties is an attractive tool to predict polyketide chain backbone length as all type I polyketide synthases have a KS domain.

1.2 Aims and objectives

This study aims to elucidate the relationship between KS domain physicochemical properties and polyketide product length across different types of iterative type I PKS and different species. The objectives of the study are the following:

- A list of HR-, PR-, and NR-PKS will be compiled from extensive literature review
- Physiochemical properties of the list of PKSs will be derived and calculated using *in silico* methods
- Mathematical approaches and machine learning techniques will be applied on the physiochemical properties of the PKS to develop a generalized model that can predict polyketide chain backbone length.

Developing a generalized predictive model for polyketide chain backbone length provides an additional tool in the rational tuning of polyketide products. Expanding the chemical space of polyketide products may lead to discovery of products with clinical use.

1.3 Materials and methods

1.3.1 Dataset

An extensive review of literature involving iPKS published up to December 2019 was conducted. The search terms used were “iterative”, “Type I”, and “polyketide synthase” and excluded “Type II”, “Type III”. Publications that investigated modular type I PKS and NRPS were excluded. Only publications that experimentally characterized the role of an iPKS and its product were chosen. From the reviewed literature, the following information was extracted: name of iPKS, the source organism, type of organism, the type of PKS, the pre-release product, and the number of carbons atoms in the pre-release product. The full amino acid sequence of the iPKS in the selected publications were retrieved from Uniprot and NCBI. Information on the type of PKS was also obtained from Uniprot annotation.

1.3.2 KS domain prediction and phylogenetic analysis

KS domain sequences of the full-length iPKS protein sequences were identified using the PKS/NRPS Analysis Web Server (<http://nrps.igs.umaryland.edu/>)¹⁹. In cases where there was no automatic detection of the KS domain boundaries, manual inspection of the server scores was conducted and the KS region identified with the highest level of confidence and score was used. The phylogenetic analysis on the set of KS domain sequences was conducted using the Phylogeny.fr (<http://www.phylogeny.fr>) web server. The workflow consists of sequence alignment using MUSCLE (v3.8.31) with standard mode and default parameters, removal of ambiguous regions using Gblocks (v0.91b), and construction of the phylogenetic tree using the maximum likelihood method implemented by the PhyML 3.1/3.0 aLRT. The constructed trees were bootstrapped and TreeDyn (v198.3) was used for tree visualization.

1.3.3 KS domain modelling

The KS domains identified in each of the iPKSs was modelled using the homology-modelling server SWISS-MODEL (<https://swissmodel.expasy.org/>)²⁰. The highest scoring model was chosen based on the GMQE and QMEAN score.

1.3.4 Active site analysis

Cavity area and volume were calculated using the CASTp (Computed Atlas of Surface Topography of proteins) version 3.0 server (<http://sts.bioe.uic.edu/castp/index.html>)²¹. In the analysis of the cavity volumes in the KS domain models, only the cavities which contained the catalytic triad (Cys-His-Asn), were selected. KS domain residue hydropathy was calculated using the Protein GRAVY server (https://www.bioinformatics.org/sms2/protein_gravy.html). The GRAVY (grand average of hydropathy) value was calculated by summing up the hydropathy value for each residue and dividing the total value by the length of the amino acid sequence²². The models were visualized using USCF Chimera²³.

1.3.5 Statistical analysis and machine learning

The data in this study was analysed using Python v3.6.5, and open-source Python libraries for statistical and machine learning. The libraries are listed in Table 1 along with the purpose of each library²⁴⁻²⁹.

Table 1. Python packages and their purpose used in current study.

Package name	Version	Purpose
sklearn	0.21.3	General-purpose machine learning library
xgboost	0.82	XGBoost machine learning model library
numpy	1.15.4	Library for numerical and mathematical operations
pandas	0.23.0	Library to manipulate tabular data
scipy	0.21.3	Scientific and statistical methods library
seaborn	0.9.0	Data visualization library
statsmodel	0.9.0	Library for statistical methods
statannot	0.2.1	Annotation library for statistical results
matplotlib	2.2.2	Data visualization library
catboost	0.14.2	Gradient boosting on decision tree library
lightgbm	2.2.3	Fast and light gradient boosting machine library

1.4 Results

1.4.1 iPKS dataset

Based on the literature search, a total of 97 iPKSs were identified from 79 publications³⁰⁻¹⁰⁸ and analysed in this study (Table 2). Out of the 97 sequences, 48 (49.5%) were NR-PKS, 19 (19.6%) were PR-PKS, and 30 (30.9%) were HR-PKS. 79 (81.4%) iPKS belonged to fungi and 18 (18.6%) iPKS belonged to bacteria. The range of the number of the carbon atoms in the pre-release polyketide backbone is between 5 to 28.

Table 2. The list of PKS dataset used in this study.

Source	Name	Template	Identity (%)	Organism	Reduction	Product	Product length	Accession
Zhang et al, 2003	Pks1	6FIJ	54.99	<i>Glarea lozoyensis</i>	NR	tetrahydroxynaphthalene	10	AAN59953.1
Ma et al, 2007	Pks4	6FIJ	62.47	<i>Fusarium fujikuroi</i>	NR	nonaketide SMA76a	18	CAB92399.1
Fujii et al, 2000	PKS1	6FIJ	54.52	<i>Colletotrichum lagenarium</i>	NR	1,3,6,8-tetrahydroxynaphthalene	10	BAA18956.1
Sanchez et al, 2010	OrsA	6FIJ	43.85	<i>Aspergillus nidulans</i>	NR	orsellinic acid	8	CBF73505.1
Wang et al, 2008	Rads2	5E5N	35.80	<i>Floropilus chiversii</i>	NR	radicicol precursor	17	C5H882.1
Zhou et al, 2008	ZEA1	5E5N	34.61	<i>Gibberella zeae</i>	NR	zearalenone precursor	18	A0A098D6U0.1
Zhou et al, 2010	Hmp3	5E5N	32.38	<i>Hypomyces subiculosus</i>	NR	(6'S, 10'S)-trans-7',8'-dehydrozearalenol	18	B3FWS8.1
Zhou et al, 2010	Rdc1	4Z37	35.29	<i>Pochonia chlamydosporia</i>	NR	(R)-monocillin II	18	ACD39770
Xu et al, 2014	CurS2	5E5N	34.83	<i>Aspergillus terreus</i>	NR	10,11-dehydrocurvularin	16	AGC95321
Moriwaki et al, 2004	PKS1	6FIJ	56.35	<i>Bipolaris oryzae</i>	NR	1,3,6,8-tetrahydroxynaphthalene	10	BAD22832
Zhang et al, 2017	PfmaE	6FIJ	55.92	<i>Pestalotiopsis fici</i>	NR	1,3,6,8-tetrahydroxynaphthalene	10	ETS82099.1
Tsai et al, 1998	Alb1	6FIJ	61.63	<i>Neosartorya fumigata</i>	NR	heptaketide naphthopyrone YWA1	14	AAC39471
Fujii et al, 2001	WA	6FIJ	61.63	<i>Emericella nidulans</i>	NR	YWA1	14	Q03149
Chiang et al, 2011	AlbA	6FIJ	60.00	<i>Aspergillus niger</i>	NR	heptaketide	14	EHA28527
Frandsen et al, 2006	PKS12	6FIJ	60.28	<i>Fusarium graminearum</i>	NR	nor-rubrofusarin	14	AAU10633
Cary et al, 2009	AflC	6FIJ	58.14	<i>Aspergillus ochraceoroseus</i>	NR	norsolorinic acid anthrone	20	ACH72912
Newman et al, 2012	CTB1	6FIJ	100.00	<i>Cercospora nicotianae</i>	NR	trihydroxynaphthalene	14	AAT69682
Studt et al, 2012	FSR1	6FIJ	62.36	<i>Gibberella fujikuroi</i>	NR	6-O-demethylfusarubinaldehyde	14	CCE67070
Chettri et al, 2013	NSAS	6FIJ	57.31	<i>Dothistroma septosporum</i>	NR	norsolorinic acid	20	AAZ95017
Awakawa et al, 2012	NhPKS1	6FIJ	63.66	<i>Nectria haematococca mpVI 77-13-4</i>	NR	3-acetonyl-1,6,8-trihydroxy-2-naphthaldehyde	14	XP_003039929
Chooi et al, 2010	GSFA	6FIJ	50.82	<i>Penicillium aethiopicum</i>	NR	heptaketide	14	ADI24953
Lim et al, 2012	EncA	6FIJ	55.22	<i>Neosartorya fumigata</i>	NR	atrochrysone thioester	16	XP_746435
Chiang et al, 2010	MdpG	6FIJ	54.76	<i>Emericella nidulans</i>	NR	atrochrysone carboxylic acid	16	XP_657754
Li et al, 2011	AptA	6FIJ	49.65	<i>Emericella nidulans</i>	NR	2,3,6,8,9-pentahydroxy-1-oxo-3-(2-oxopropyl)-1,2,3,4-tetrahydroanthracene-2-carboxylic acid	18	XP_663604
Li et al, 2011	AdaA	6FIJ	51.97	<i>Aspergillus niger</i>	NR	decaaketide	20	XP_001394705
Awakawa et al, 2009	ACAS	6FIJ	57.24	<i>Aspergillus terreus</i>	NR	atrochrysone carboxylic acid precursor	16	XP_001217072
Chooi et al, 2010	VrtA	6FIJ	51.04	<i>Penicillium aethiopicum</i>	NR	polyketide backbone'	19	ADI24926
Klejnstrup et al, 2012	AusA	6MHK	37.77	<i>Emericella nidulans</i>	NR	3,5-dimethylorsellinic acid	10	XP_681652
Regueira et al, 2011	MapC	6MHK	38.13	<i>Penicillium brevicompactum</i>	NR	5-methylorsollinic acid	9	ADY00130
Ahuja et al, 2012	PkbA	5E5N	38.02	<i>Aspergillus nidulans FGSC A4</i>	NR	3-methylorsellinic acid	9	XP_664052
Chiang et al, 2009	AfoE	4Z37	35.84	<i>Emericella nidulans</i>	NR	2,4-dihydroxy-6-(5,7-dimethyl-2-oxo-trans-3-trans-5-nonadienyl)-3-methylbenzaldehyde	19	XP_658638
Ahuja et al, 2012	PkfA	5E5N	35.31	<i>Emericella nidulans</i>	NR	orsellinaldehyde	8	XP_660834
Zabala et al, 2012	AzaA	4MZ0	38.13	<i>Aspergillus niger</i>	NR	azanigerone aldehyde precursor	13	EHA28237
Storm et al, 2017	PksCT	4MZ0	40.38	<i>Monascus purpureus</i>	NR	keto-aldehyde intermediate	13	BAD44749
Bailey et al, 2007	PksI	5E5N	33.10	<i>Sarocladium strictum</i>	NR	3-methylorcinolaldehyde	8	CAN87161
Lackner et al, 2013	ArmB	5E5N	38.48	<i>Armillaria mellea</i>	NR	orsellinic acid	8	AFL91703
Winter et al, 2012	CazM	4MZ0	36.58	<i>Chaetomium globosum</i>	NR	cazaldehyde A	16	AKA40069.1
Fahad et al, 2014	SorB	4Z37	34.61	<i>Penicillium rubens</i>	NR	sorbicillin	14	CAP95404.1
Chooi et al, 2015	SnPKS19	6FIJ	52.09	<i>Parastagonospora nodorum</i>	NR	alternariol	14	AKN45693.1
Chooi et al, 2008	XsePKS1	5E5N	34.71	<i>Xanthoparmelia semiviridis</i>	NR	C4-methylated tetraketide	9	EF547512.1
Yu et al, 2016	PKS63787	5E5N	36.17	<i>Antrrodia cinnamomea</i>	NR	orsellinic acid	8	AST08390.1
Abdel-Hameed et al, 2016	Cu-PKS-4	6FIJ	54.19	<i>Cladonia uncialis</i>	NR	2,3-dehydro-6-HM	10	AUW31184.1
Xu et al, 2014	LtLasS2	5E5N	33.73	<i>Lasiodiplodia theobromae</i>	NR	octaketide	16	AHV78247.1
Xu et al, 2014	AzResS2	4NA2	37.62	<i>Acremonium zeae</i>	NR	nonaketide	16	AHV78253.1
Kashiwa et al, 2016	PNK2	6FIJ	60.93	<i>Phoma sp.</i>	NR	nonaketide	18	BAW18768.1
Itoh et al, 2012	Trt4	6MHK	38.50	<i>Aspergillus terreus</i>	NR	3,5-dimethylorsellinic acid	10	EAU29529.1
Griffiths et al, 2018	CfPKs1	6FIJ	57.51	<i>Cladosporium fulvum</i>	NR	1,3,6,8-tetrahydroxynaphthalene	10	
Artigot et al, 2009	6MSAS	6MHK	38.79	<i>Aspergillus clavatus</i>	NR	6-methylsalicylic acid	8	EAW11667.1
Liu et al, 2005	NcsB	6MHK	39.52	<i>Streptomyces carzinostaticus subsp. Neocarzinostaticus</i>	PR	2-hydroxyl-5-methyl-naphthoic acid	12	AAM77986.1
Sun et al, 2012	SACE5548	2QO3	46.35	<i>Saccharopolyspora erythraea</i>	PR	(R)-mellein	10	WP_009944722.1
Zhao et al, 2008	AziB	6MHK	37.68	<i>Streptomyces sahachiroi</i>	PR	5-methyl-1-naphthoate	12	B4XYB8.1

Xiao et al, 2011	TiaB	5E5N	<i>Dactylosporangium</i>	PR	orsellinic acid	8	ADU85988.1
			40.76 <i>aurantiacum subsp. Hamdenensis</i>				
Ahlert et al, 2002	CalO5	2QO3	51.07 <i>Micromonospora echinospora</i>	PR	orsellinic acid	8	AAM70355.1
Ito et al, 2009	PctS	4MZ0	52.12 <i>Streptomyces pactum</i>	PR	6-methylsalicylic acid	8	BAF92601.1
Van Lanen et al, 2007	MdpB	4MZ0	50.95 <i>Actinomadura madurae</i>	PR	6-methylsalicylic acid	8	ABY66019.1
Daum et al, 2009	PokM1	4MZ0	52.24 <i>Streptomyces diastatochromogenes</i>	PR	6-methylsalicylic acid	8	ACN64831.1
Shao et al, 2006	ChlB1	4MZ0	54.82 <i>Streptomyces antibioticus</i>	PR	6-methylsalicylic acid	8	CAM04547.1
Chooi et al, 2015	SN477	6MHK	38.53 <i>Parastagonospora nodorum</i>	PR	(R)-mellein	10	AIW00670.1
Kage et al, 2013	RSc1806	4MZ0	50.72 <i>Ralstonia solanacearum</i>	PR	6-pentylsalicylic acid	12	CAD15508.1
Tannous et al, 2014	6MSAS	5E5N	39.67 <i>Penicillium expansum</i>	PR	6-methylsalicylic acid	8	AIG62146.1
Spencer and Jordan, 1992	6MSAS	6MHK	38.30 <i>Penicillium patulum</i>	PR	6-methylsalicylic acid	8	CAA39295.1
Kim et al, 2012	CmPKS1	5E5N	41.29 <i>Cladonia metacorallifera</i>	PR	6-methylsalicylic acid	8	ADQ27444.1
Gaisser et al, 1997	AviM	2QO3	50.47 <i>Streptomyces viridochromogenes</i>	PR	orsellinic acid	8	AAK83194.1
Karlsson et al, 2015	PksL2	5E5N	36.90 <i>Aspergillus parasiticus</i>	PR	6-methylsalicylic acid	8	AAC23536.1
Lohman et al, 2013	KedU38	4OQJ	57.96 <i>Streptoalloteichus sp. ATCC 53650</i>	PR	3-hydroxy-7,8-dimethoxy-6-isopropoxy-2-naphthoic acid	11	AFV52200.1
Fujii et al, 1996	ATX	4MZ0	44.26 <i>Aspergillus terreus</i>	PR	6-methylsalicylic acid	8	EAU32819.1
Bacha et al, 2009	AOMSAS	6MHK	38.95 <i>Aspergillus westerdijkiae</i>	PR	6-methylsalicylic acid	8	AAS98200.1
Gao and Thorson, 2008	DynE8	4QYR	34.15 <i>Micromonospora chersina</i>	HR	β -hydroxyhexaene	15	AAO25899.1
Sun et al, 2009	SgcE	4OPE	31.82 <i>Streptomyces globisporus</i>	HR	1,3,5,7,9,11,13-Pentadecaheptaene	15	WP_010056319.1
Liu et al, 2005	NcsE	4OPE	33.97 <i>Streptomyces carzinostaticus</i>	HR	1,3,5,7,9,11,13-Pentadecaheptaene	15	AAM78012.1
Van Lanen et al, 2007	MdpE	4OPE	33.97 <i>Actinomadura madurae</i>	HR	1,3,5,7,9,11,13-Pentadecaheptaene	15	AAQ17110.2
Sun et al, 2009	CalE8	4OPE	31.82 <i>Micromonospora echinospora</i>	HR	β -hydroxyhexaene	16	AAM94794.1
Zabala et al, 2014	Bref-PKS	6MHK	36.49 <i>Eupenicillium brefeldianum</i>	HR	octaketide	16	AIA58899.1
Butchko et al, 2003	FUM1	5E5N	37.86 <i>Fusarium verticillioides</i>	HR	fumonisin precursor	20	AAD43562.2
Xie et al, 2009	LovF	5E5N	38.55 <i>Aspergillus terreus</i>	HR	2-methylbutyryl	5	AAD34559.1
Ma and Tang, 2007	LovB	4QYR	41.59 <i>Aspergillus terreus</i>	HR	dihydromonacolin L	19	AAD39830.1
Sakai et al, 2009	MokB	5E5N	37.59 <i>Monascus pilosus</i>	HR	diketide	5	ABA02240.1
Winter et al, 2012	CazF	5E5N	40.61 <i>Chaetomium globosum</i>	HR	4-methyl-hex-2-enoate	7	EAU31921.1
Fahad et al, 2014	SorA	4QYR	38.89 <i>Penicillium rubens</i>	HR	methylated hexaketide	14	CAP95405.1
Chiang et al, 2009	AfoG	5E5N	39.19 <i>Emericella nidulans</i>	HR	3,5-dimethyloctadienone	10	CBF88289.1
Ugai et al, 2015	Bet1	5E5N	39.34 <i>Pleospora betae</i>	HR	betaenone C	21	BAQ25466.1
Chen et al, 2016	GloL	5E5N	39.07 <i>Glarea lozoyensis</i>	HR	10, 12-dimethyl myristoyl	16	EPE34340.1
Liu et al, 2003	EspE	4OPE	31.89 <i>Actinomadura verrucososporea</i>	HR	β -hydroxyhexaene	15	AAP92148.1
Zhou et al, 2010	Hmp8	5E5N	39.62 <i>Hypomyces subiculosus</i>	HR	(7S,11S,2E,8E)-7,11-dihydroxydodeca-2,8-dienoate	12	ACD39767.1
Zabala et al, 2012	AzaB	5E5N	39.57 <i>Aspergillus niger</i>	HR	2,4-Dimethylhexanoic acid	8	EHA28244.1
Zhou et al, 2010	Rdc5	5E5N	41.77 <i>Pochonia chlamydosporia</i>	HR	hexa/pentaketide	12	ACD39774.1
Kasahara et al, 2010	PSS	5E5N	35.00 <i>Alternaria solani</i>	HR	desmethylprosolanapyrone I	16	BAJ09789.1
Fujii et al, 2005	PKSN	4MZ0	45.61 <i>Alternaria solani</i>	HR	alternapyrone	28	BAD83684.1
Mao et al, 2015	AurA	4QYR	38.86 <i>Calcarisporium arbuscula</i>	HR	hexaene pyrone	22	ALD83627.1
Kasahara et al, 2006	PKSF	5E5N	38.42 <i>Alternaria solani</i>	HR	aslanipyronone / aslaniol	22	BAE80697.1
Saha et al, 2012	PksJ	5E5N	38.04 <i>Alternaria alternata</i>	HR	3,5,7,9,11,13-hexaoxotetradecanoic acid	14	AFN68301.1
Xu et al, 2014	Curs1	5E5N	39.33 <i>Aspergillus terreus</i>	HR	tetraketide	8	AGC95324.1
Janevska et al, 2016	Gpy1	5E5N	36.97 <i>Fusarium fujikuroi</i>	HR	gibepyrone A carbon backbone	10	CCT75967.1
Cox et al, 2004	SQHKS	5E5N	33.25 <i>Phoma sp.</i>	HR	hexaketide	19	AMY15068.1
Cox et al, 2004	SQTKS	5E5N	37.38 <i>Phoma sp.</i>	HR	tetraketide	10	AMY15057.1
Xu et al, 2014	LtLasS1	5E5N	39.76 <i>Lasiodiplodia theobromae</i>	HR	9(R)-hydroxydecanoic acid	10	AHV78245.1
Xu et al, 2014	AzResS1	5E5N	41.97 <i>Acremonium zeae</i>	HR	7(R)-hydroxyoct-2(E)-enoic acid	8	AHV78252.1

NR – non-reducing, PR – partially-reducing, and HR – highly-reducing.

1.4.2 Phylogenetic tree of iPKS KS domains

The purpose of the constructing the phylogenetic tree is to examine the diversity of the KS domains from the different types of iPKS. The tree starts with two major clades – non-reducing and reducing iPKSs (Fig 1).

The clade containing reducing iPKSs branch into two subclades containing PR-PKS and HR-PKS, and enediyne producing HR-PKS. The PR-PKS clade branches into two subclades. The first subclade contains 6-methylsalicyclic acid (6-MSA) producing KS from *Aspergillus* spp. and *Penicillium* spp. (Figure 3, upper line), whereas the second subclade contains the rest of the PR-PKS that produce cyclic polyketide compounds other than 6-MSA such as (R)-mellein, orsellinic acid, and naphthoic acid analogues. While the positions of the KS domains in the PR-PKS subclade were mostly correctly group in the same subclade, two 6-MSA-producing iPKS – 6MSAS from *Aspergillus clavatus* and 6MSAS from *Byssoschlamys nivea* – were found in this subclade even though these two proteins were classified as NR-PKS in literature.

The enediyne-producing HR-PKS KS domains are in a distinct subclade (Figure 3, lower line). The subclade is located away from the rest of the HR-PKS and PR-PKS subclades. Apart from the 6-MSAS and enediyne-producing PKS, there is no observable correlation between polyketide, polyketide length and the clade positions.

In summary, the phylogenetic tree based on KS domains alone showed that it is possible to group PKS into distinct reducing and non-reducing PKS clades.

Figure 3. KS domain genealogy of characterized iPKSs. The clades are colored in red (PR-PKS), blue (HR-PKS), and yellow (NR-PKS). 6-MSAS-type PR-PKS and enediayne-producing HR-PKS are highlighted.

1.4.3 KS domain modelling

The homology models for the 97 iPKSs were generated using the SWISS-MODEL server.

The templates used in the modelling is described in Table 3.

Table 3. PDB models used as template for modelling. The list contains the PDB code, source organism, and the description of the protein.

Template PDB	Organism	Purpose
4MZ0	<i>Moorea producens</i> 3L	Sixth module of curacin A synthesis (CurL)
2QO3	<i>Saccharopolyspora erythraea</i>	Third module of 6-deoxyerthronolide B synthesis (EryAII)
4NA2	<i>Bacillus subtilis</i>	Second module of bacillaene synthesis (PksJ)
4OPE	<i>Streptomyces albus</i>	Oxazolomycin iPKS (OzmH)
4OQJ	<i>Streptomyces albus</i>	Oxazolomycin iPKS (OzmQ)
4QYR	<i>Streptomyces platensis</i>	Isomigrastatin iPKS (MgsE)
4Z37	<i>Brevibacillus brevis</i>	Second module of C0ZGQ5
5E5N	<i>Bacillus subtilis</i>	Sixth module of bacillaene synthesis (PksL)
6FIJ	<i>Cercospora nicotianae</i>	Cercosporin iPKS (CTB1)
6MHK	<i>Bacillus amyloliquefaciens</i>	Ninth module of bacillaene synthesis (BaeM)

The modelling of the 97 KS domains used 10 published and unpublished crystal structures^{11,109-113}. Almost two-thirds of the HR-PKS KS domains (19/30) were modelled after PksL, the sixth module of bacillaene synthesis (PDB: 5E5N). For NR-PKS, around half of the KS domains (26/48) were modelled after NR-PKS cercosporin synthase CTB1 (PDB: 6FIJ). There was no observable preference for templates when PR-PKSs were modelled. Apart from the fungal cercosporin NR-PKS CTB1, the templates come from PKS found in

bacteria and more specifically there is a large proportion of *Bacillus* spp. and *Streptomyces* spp. proteins.

Figure 4. Representative homology models of the different types of KS domains. From left: OrsA from *Aspergillus nidulans* (yellow), DynE8 from *Micromonospora chersina* (green), NcsB from *Streptomyces carzinostaticus* (red), and Hmp8 from *Hypomyces subiculosus* (blue). Blue circle: extra loop found in enediynes-producing iPKSs.

While the homology models across the three different types looked globular and were generally similar to a ketosynthase domain (Figure 4), the homology models of HR-PKSs that produce enediynes, e.g., DynE8, have an extended loop on top of the structure (Figure 4, second from left).

After trimming the sequences from the alignment from MUSCLE using Gblocks, it was found that there are 56 amino acid-long regions that are conserved and found only in enediyne-producing HR-PKSs (Figure 5). This sequence corresponds to the loop that extends out of the KS main body in the corresponding homology models (Figure 4, blue circle).

	130	140	150	160	170	180
	=====	=====	=====	=====	=====	=====
SgcE_Streptomyces	LDTATRALADAGFPAGEGLPTERT	VVVVGNLT	TGEFSRANGLRLRPYVRRILADALQ-E			
NcsE_Streptomyces	LDTAGRALADAGFPAGEGLPHERT	VVMGNTLT	TGEFTRAMVRLRPYVRRVMAAALAGQ			
MdpE_Actinomadu	LDVAGRALADAGFPAGEGLPRERT	VIVGNTLT	TGEFTRANVLRRLRPYVRRVVAAGLREQ			
CalE8_Micromono	LDMAAQALADAGFPAGEGLPRERT	VVVVGNLT	TGEFTRAGMMRLRPYVRRVGAALGEQ			
EspE_Actinomadu	LDVAGDTLAAAGFAGGGGLDRRA	VILGNTLT	TGEFSRAALMLRLRPYVRRNLVSALAAR			
DynE8_Micromono	LETAADALVDAGLPDGTGDRSRV	VVIGNSLT	TGEFSRAAGLRIRWPYVRRAVAAGLVDG			
ArmB_Armillaria	LHTAYEALEDAGYVPN	----	STPT	-----	-----	-----
PKS63787_Antrond	LHVAYHALENAGYVPN	----	ATPC	-----	-----	-----
PkbA_AspERGillu	LQVAYQTLESAEYSG	----	SGI	-----	-----	-----
MapC_Penicilliu	LQVAYEAMESAGYLGW	----	QPN	-----	-----	-----
AusA_Emericella	LQVAYEAMESAGYCGL	----	RRS	-----	-----	-----
Trt4_AspERGillu	LQVAYEAIESAGYCGL	----	RAS	-----	-----	-----
AfoE_Emericella	LQAAYQAVEQSGYFVEDHNGYTPD					
CazM_Chaetomium	MQAAYQAVEQSGYFTETTA	--	APE	-----	-----	-----
PkFA_Emericella	LQAAYQAVEQSGYFLE	----	PE	-----	-----	HNE
SorB_Penicilliu	LQTAYQAVAQSGYYHR	----	PE	-----	-----	
Pks1_Sarocladiu	LQVAYQAVEGSGYFNK	----	SSSS	-----	-----	
AzaA_AspERGillu	LQVAYQAVEQSGYFTT	----	PKSD	-----	-----	
PksCT_Monascus	LQVAYQAVEQSGYFRN	----	NGTN	-----	-----	

Figure 5. A small window for the Gblocks output for the 97 KS domains. The underlined sequence in red is the highly conserved amino acids found on in enediyne-producing HR-PKS.

The cavity volume and area of each KS domain was measured using the CASTp server. A mathematical probe based on computational geometry methods is used by the server to map the various cavities in the model. The cavity space is filled with red spheres (Figure 6), and residues in contact with the spheres were identified. Since there was not setting to narrow the search for cavities, manual inspection was required to select the cavity that contained the catalytic triad in the set of cavities found and that the cavity mapped involved the conserved residues implicated in ketosynthase catalytic activity¹¹.

Figure 6. Representative output of a model submitted into the CASTp server. A rendering of 6-MSAS from *Aspergillus clavatus*. The red spheres denote the cavity space measured by the server.

1.4.4 Exploratory data analysis of KS domains

To examine the distribution of the cavity volumes in the KS dataset, the cavity volumes were plotted on a histogram (Figure 7).

Figure 7. Histogram of all cavity volumes in the dataset. (A) Distribution of the raw values of KS domains. (B) Distribution of the log transformed values of KS domains.

The resulting histogram showed a right skew distribution (Figure 7A). The average cavity volume is $163 \text{ \AA}^3$ with a standard deviation of $76 \text{ \AA}^3$, and a median of $145 \text{ \AA}^3$. There is a

slight outlier, the *Streptomyces carzinostaticus* NcsE with a cavity volume of 485 Å³. The cavity data was also log transformed and plotted, which resulted in a more normal distribution of data (Figure 7B). The KS domains were further divided into HR-, PR-, and NR-KS domains and the respective volumes were plotted in separate histograms (Figure 8A-D). PK-PKS cavity volumes do not have any observable distribution (Figure 8A) due to a relatively small sample size. HR-PKS and NR-PKS show a similar right skew (Figure 8B – C). The average cavity volume for HR-, PR-, and NR-KS domains is $187 \pm 101 \text{ Å}^3$, $169 \pm 72 \text{ Å}^3$, and $145 \pm 54 \text{ Å}^3$ respectively.

Figure 8. Histograms of the distribution of the KS domain cavity volumes. (A) PR-PKS (red). (B) HR-PKS (blue). (C) NR-PKS (yellow). (D) The three histograms plotted in a single plot. Number of bins = 10.

Normality was not assumed in these distributions and the Mann-Whitney test was conducted to test for the differences between the distribution of the cavity volumes in the KS domains. Mann-Whitey test was used as the populations of the KS domain cavity volumes are not

normally distributed. A boxplot was plotted to visualize the differences between the KS domains (Figure 9). There was statistical significance in the difference between the NR-KS domain cavity volumes and the HR-KS domain volumes, with a p -value of 0.02. There was no statistical significance in the difference between NR-KS domain and PR-KS domain volumes, and PR-KS and HR-KS domain volumes, with p -values of 0.171 and 0.853, respectively.

Figure 9. The boxplot for the cavity volumes of NR-, PR-, and HR-KS domains. The difference is marginally significant for $0.01 < p < 0.05$ (*), and not significant for $p > 0.05$ (ns).

The GRAVY score is an index that measures how hydrophobic/hydrophilic an amino acid sequence is. The distribution for GRAVY score was examined as well (Figure 10).

Figure 10. Histogram of the GRAVY scores of the KS domains. Number of bins = 10.

The more positive the number, the more hydrophobic the sequence would be whereas negative numbers mean a hydrophilic nature. The average GRAVY score is -0.177, with a standard deviation of 0.102 and median of -0.198. The GRAVY scores are normally distributed with a slight right skew. The GRAVY scores for the three types of KS domains are plotted for comparison (Figure 11). Similarly, due to a small sample size, there is no observable distribution for PR-PKS KR GRAVY scores. However, the GRAVY scores of HR-PKS and NR-PKS are normally distributed. The average GRAVY scores for HR-, PR-, and NR-KS domains is -0.16 ± 0.07 , -0.08 ± 0.10 , and -0.23 ± 0.08 , respectively.

GRAVY scores for the overall domain was used because it was challenging to identify the exact cavities using the homology models alone. On the other hand, it was easier to identify the entire KS domain from full-length sequences using the PKS/NRPS Analysis Web Server. Since the GRAVY score is a normalized average of the scores assigned to each residue, and if the hydrophilic solvent-facing residues were generally the same, the major differences would lie in the catalytic cavity and subsequently the hydrophobic residues found within.

Figure 11. Histograms of the distribution of the KS domain GRAVY scores. (A) PR-PKS (red). (B) HR-PKS (blue). (C) NR-PKS (yellow). (D) The three histograms plotted in a single plot. Number of bins = 10.

Statistical tests for difference in distributions were conducted. While the distribution appeared normal for the overall grouped GRAVY scores and for the respective types of KS, no distribution assumptions were made, and Mann-Whitney test was chosen. Mann-Whitney test was used as the GRAVY scores did not seem normally distributed. A boxplot was plotted to visualize the differences between the GRAVY scores (Figure 12). There was statistical significance in the differences between the GRAVY scores for NR-, PR-, and HR-KS domains. The p -values for the differences in GRAVY scores between NR- and PR-KS, NR- and HR-KS, and PR- and HR-KS are 4.27×10^{-6} , 3.06×10^{-3} , and 5.27×10^{-4} , respectively.

Figure 12. The boxplot for the GRAVY scores of NR-, PR-, and HR-KS domains. The difference is significant for $1 \times 10^{-3} < p \leq 1 \times 10^{-2}$ (**), $1 \times 10^{-4} < p \leq 1 \times 10^{-3}$ (***), and $p \leq 1 \times 10^{-4}$ (****).

These results imply that the hydrophobic environments of the three types of KS are different. Collectively, the KS domains of both reducing types of iPKS (PR- and HR-PKS) have relatively more positive GRAVY score compared to NR-PKS. This implies the KS domains of reducing PKSs are more hydrophobic than non-reducing KS.

The distribution of product lengths in the dataset was also examined. A histogram was plotted to visualize the distribution (Figure 13). The iPKS product lengths are distributed normally with a right skew. The average length of the polyketide product is 13, with a median of 13 and a standard deviation of 4.55.

Figure 13. Histogram of the polyketide product lengths in the dataset. Bins = 10.

To examine the differences between the product lengths from the three types of iPKSs, the product lengths were plotted in separate histograms (Figure 14A – D). The distribution of polyketide product lengths in HR-PKS is approximately normal. However, there is no normal distribution for the product lengths in NR-PKS and PR-PKS.

On top of that, the ranges in the three types of PKS are distinctly different (Figure 14D). HR-PKS produces the products with the largest range of lengths (5 – 28) whereas the range of lengths for PR-PKS is narrow (8 – 12), with 13/19 products 8 carbons long. The average product lengths for HR-, PR-, and NR-PKS are 14 ± 5 , 9 ± 2 , and 14 ± 4 respectively.

Figure 14. The histograms of the distributions of PKS product length. (A) PR-PKS (red). (B) HR-PKS (blue). (C) NR-PKS (yellow). (D) The three histograms plotted in a single plot. Number of bins for subplots A – C = 5. Number of bins for subplots D = 10.

Given that parametric conditions cannot be assumed in the product length distribution, the Mann-Whitney test was used to assess the differences in the three population. Mann-Whitney test was used as the populations of the PKS product lengths are not normally distributed. A boxplot was also plotted to visualize the distributions (Figure 15). There is no statistical significance in the difference between the product lengths in HR-PKS and NR-PKS. However, there is a statistical significance in the difference between NR- and PR-PKS product lengths, and PR- and HR-PKS product lengths. The p -values for the differences in GRAVY scores between NR- and PR-KS, NR- and HR-KS, and PR- and HR-KS are 5.16×10^{-6} , 0.804, and 4.72×10^{-4} , respectively.

The finding agrees with the observation that PR-PKS typically produces small cyclic polyketide products. On the other hand, both NR- and HR-PKS have an extensive product repertoire and synthesize a larger range of products.

Figure 15. The boxplot for the polyketide product lengths for NR-, PR-, and HR-PKS. The difference is statistically significant for $1 \times 10^{-4} < p \leq 1 \times 10^{-3}$ (***), $p \leq 1 \times 10^{-4}$ (****), and not significant for $p \geq 0.05$ (ns).

The summary of the exploratory data analysis performed for cavity volume, GRAVY scores, and product lengths is presented in Table 4.

Table 4. Summary of the exploratory data analysis. Information on the average and the standard deviation are given for cavity volume, GRAVY score, and product length of NR-, PR-, and HR-PKS.

	Volume (A_3)	GRAVY	Product length
NR-PKS	144.70 ± 53.94	-0.23 ± 0.08	13.67 ± 3.88
PR-PKS	168.86 ± 72.27	-0.08 ± 0.10	9.00 ± 1.60
HR-PKS	187.48 ± 100.62	-0.16 ± 0.07	14.10 ± 5.48

The exploratory data analysis was useful for investigating characteristics of individual features and to assess if there are statistical significances in the differences between the types of the PKS.

1.4.5 Bivariate analyses in KS domain biochemical properties

In statistics, bivariate analysis involves identifying and quantifying the empirical relationship between two different variables in a dataset¹¹⁴. For this study, ordinary least squares (OLS) regression was used to assess the relationship between the different features of the dataset, and with the polyketide product length. Residual plots were also plotted to assess model prediction performance. The equation for residual e is given as $e = y - \hat{y}$, where y is observed value and $\hat{y}$ is predicted value. The assumption is that the relationships between the KS domain physicochemical properties are linear.

Figure 16. Plot between KS domain cavity volume vs KS domain GRAVY score. (A) Volume against GRAVY score. (B) Log volume against GRAVY score.

Firstly, the KS domain volume was plotted against GRAVY score. There is no observable correlation ($R^2 = 0.03$) between KS domain volume and GRAVY score (Figure 16A). The log-transformed KS domain cavity volume is not correlated with the GRAVY score ($R^2 = 0.01$) either (Figure 16B). Both plots do not reveal discernible upwards or downwards trend and the coefficient of the independent variable is close to zero, 2.34×10^{-4} and 0.03 for Figure 16A and 16B respectively. The lack of correlation is anticipated because GRAVY score is calculated by dividing the total hydropathy score of the residues in an amino acid

sequence by the number of residues. This results in a score that is invariant to sequence length whereas KS cavity volume is positively correlated to the number of residues.

Figure 17. The OLS regression for Gravy score against product length. (A) Scatterplot with the regression line. (B) Residual plot of the regression between actual product length and predicted product length from the regression.

Secondly, an OLS regression was plotted for Gravy score against product length, and a weak negative correlation was observed (Figure 17A). The intercept was found to be 9.67 and the coefficient was -18.21. As such, the equation for the regression was $\hat{y} = 9.67 - 18.21x$. The correlation coefficient (R^2) value for the regression was 0.17, and the root mean squared error (RMSE) was 4.12, which suggests that the correlation is poor. To compare the predictions against the actual values, a residual plot was plotted (Figure 17B). The residual plot showed that the predicted product lengths are not clustered around the $y = 0$ line, indicating that there are large differences between the actual values and the OLS predictions.

Figure 18. The OLS regression for KS domain cavity volume against product length. (A) Volume vs product length. (B) Log volume vs product length.

The correlation between KS domain cavity volume and product length was investigated. In the OLS regression for both volume against product length and log volume against product length (Figure 18A – B), the correlation is weak with a R^2 of 0.01 and 0.02 with an RMSE of 4.49 and 4.48, respectively. The plotted regression lines for both plots are almost horizontal.

Figure 19. The plots for cavity volume against product length for the three types of PKS. (A) PR-PKS KS domain. (B) HR-PKS KS domain. (C) NR-PKS KS domain.

The cavity volumes of the KS domains from the three types were plotted against product length for a qualitative assessment on the correlation (Figure 19A – C). The plots for 19B – C are similar with Figure 18A, while the plot for PR-PKS KS domain (Figure 19A) shows no discernible trend.

Overall, the bivariate analyses using OLS showed that there is a weak linear correlation between cavity volume and product length, as well as between GRAVY score and product length.

1.4.6 Multiple linear regression analyses

Apart from the line of best fit obtained from the linear models in bivariate analyses, multiple linear regression (MLR) was used to examine the correlation between more than one physicochemical property of the KS domain and the product length. In this analysis, the assumption was that more than one feature is involved in product length for iPKS. The MLR used in this study was in the form of $\hat{y} = \beta_0 + \beta_1x_1 + \dots + \beta_nx_n$, where β_0 is the constant/intercept, and β_1 , and subsequent values β_n are the coefficients for independent variables. *statsmodel* library was used to perform the MLR through OLS.

Figure 20. The MLR of actual vs predicted polyketide chain length based on log cavity volume and GRAVY score. A line plot $y = x$ is plotted in black. The RMSE of the prediction and R^2 is shown in the lower right corner. The MLR equation is shown in the upper left.

Firstly, log cavity volume and GRAVY score were used together as independent variables to be fitted in the MLR (Figure 20). Upon fitting, the intercept β_0 was found to be 7.503, β_1 to be 1.170×10^{-2} , and β_2 to be -19.730. The resulting MLR equation is $\hat{y} = 7.503 + 0.0117x_1 - 19.730x_2$. To assess the quality of the MLR, polyketide product lengths were predicted using the MLR and compared to the actual lengths (Figure 20). A line with the equation $y = x$ was plotted to visually assess how well the predictions were made against the actual length of the polyketide product. The RMSE between the actual and predicted length is 4.03, with an R^2 value of 0.21. The MLR predictions were longer than the length of the polyketide chain for chain lengths below 14, and predictions were shorter for the length for chain lengths 15 carbon atoms and above.

Next, KS domain cavity area was added into the regression to assess whether any improvement of the correlation is observed. Upon fitting the data into the MLR, the intercept

β_0 was found to be 3.848, β_1 to be 4.5×10^{-3} , β_2 to be 0.835, and β_3 to be -19.767. The resulting equation is thus $\hat{y} = 3.848 + 0.00452x_1 + 0.835x_2 - 19.767x_3$.

Figure 21. The MLR plot of actual polyketide chain length vs the predicted chain length. A line plot $y = x$ is plotted in black. The RMSE of the prediction and the R^2 is shown in the lower right corner.

While plots shown in Figure 20 and Figure 21 looked similar, care was taken to ensure that the results were from different regressions. To assess whether the predictions from the two regressions were the same, the two sets of results were plotted against each other (Figure 22).

Figure 22. Comparison of the results of the two MLRs. The x-axis are results of the MLR with area, log cavity volume, and GRAVY score. The y-axis are results of the MLR with log cavity volume and GRAVY score.

The plots reveal a strong correlation between the results of the two regressions, but the values were not identical. The similarities between the two MLRs imply the additional independent variable – KS domain area – is not useful in improving the regression.

Table 5. Summary of the bivariate analyses performed between dataset features. Pairs of variables are listed with the R^2 values and the RMSE.

Independent variable (x)	Dependent variable (y)	R^2	RMSE
Cavity volume	Product length	0.01	4.49
Log cavity volume	Product length	0.02	4.48
GRAVY score	Product length	0.17	4.12
Cavity volume	GRAVY SCORE	0.03	
Log cavity volume	GRAVY SCORE	0.01	
Cavity volume, GRAVY SCORE	Product length	0.21	4.03
Area, Log cavity volume, GRAVY SCORE	Product length	0.21	4.03

Table 5 summarizes the performance for the OLS regressions and MLR plotted so far. The RMSE for the regressions are generally similar, and the OLS regressions are implied to be

poor performers. MLR performed slightly better in terms of R^2 and RMSE. This implies that there are non-linear factors involved in the prediction of product length.

1.4.7 Machine learning modelling to predict polyketide product length

Apart from linear and multiple linear regressions, machine learning was employed to predict iPKS product lengths in order to identify possible hidden patterns that linear regressions and multiple linear regressions are unable to capture.

The machine learning modelling process involves two stages – training and testing stage. In the training process, 80% of the data (77 samples) was randomly selected and used for model training, whereas 20% of the data (20 samples) was kept as unseen test data to assess model performance. In the model training, the same features used to build MLR were used as features for machine learning, i.e. cavity area, log cavity volume, and GRAVY score. The number of carbons in the polyketide product was used as the dependent variable.

The following machine learning techniques were used for the study:

- 1) Decision tree
- 2) Random forest
- 3) k-Nearest Neighbours (kNN)
- 4) XGBoost
- 5) LightGBM
- 6) CatBoost
- 7) Gradient Boosted Machine (GBM)

Figure 23. Predicted polyketide chain lengths vs actual polyketide lengths for machine learning techniques. (A, C, E) Plot between predictions and actual lengths for decision tree, random forest, and kNN respectively. (B, D, F) Residual plot for the predictions from decision tree, random forest, and kNN, respectively.

Figure 23 displays the results for decision tree, random forest, and k-Nearest Neighbour regressions (Figure 23A – E). Decision tree regression predictions have an RMSE of 2.44 and a R² value of 0.60. Random forest regression predictions performed similarly, with an RMSE of 2.44 and a R² value of 0.60. kNN on the other hand had an RMSE of 3.99 and an R² value of 0.23. The residuals for the predictions from the three regression are close to y = 0 and suggests that the predictions are close to the actual values.

Figure 24. Predicted polyketide chain lengths vs actual polyketide lengths for boosted tree machine learning algorithms. (A, C, D, F) Plot between predictions and actual lengths for XGBoost, LightGBM, CatBoost, and GBM respectively. (B, D, E, G) Residual plots from XGBoost, LightGBM, CatBoost, and GBM respectively.

Boosted tree algorithms were also explored to predict polyketide chain lengths (Figure 24A – G). XGBoost predictions had an RSME of 3.06 from the actual lengths and an R^2 value of 0.37. LightGBM predictions had an RMSE of 3.14 and an R^2 value of 0.34. CatBoost predictions had an RMSE of 3.36 and an R^2 value of 0.24. Finally, GBM predictions had an RMSE of 2.36 from the actual polyketide chain lengths, and an R^2 value of 0.63.

Table 6. Summary of the model performances in this study.

Technique	RSME	R^2 value
Decision Tree	2.44	0.6
Random Forest	2.44	0.6
k-Nearest Neighbour	3.39	0.23
XGBoost	3.06	0.37
LightGBM	3.14	0.34
CatBoost	3.36	0.24
GBM	2.36	0.63

The boosted tree algorithms generally performed poorly compared to linear regression, MLR, and the normal tree-based algorithms, except GBM which performed the best out of the seven techniques employed in this study (Table 6).

1.5 Discussion

In the past decade, there is increasing interest in polyketide synthases and rational engineering of PKSs to widen the existing polyketide structural diversity for therapeutic purposes. The accumulation of experimental characterization of PKSs and their products has allowed a quantitative understanding of the relationships between physicochemical properties of KS and its catalytic products. This study started with curating published iPKS information for phylogenetic and regression analysis.

Previous studies showed that polyketide phylogeny was largely driven by KS domains in *Aspergillus* spp., and the result agreed with the observation in the phylogenetic tree found in this study of KS domains across different species¹¹⁵. The first major division between NR- and PR-/HR-PKS also agreed with prior reports that KS domains group depending on the presence of reducing domains¹¹⁶. However, apart from PKSs that produce enediynes and 6-MSA, the phylogenetic analysis does not suggest that there is a correlation between clade position and polyketide chain length.

This observation is supported by the statistical testing of the differences in the product length from NR-, PR-, and HR-PKS, where there was statistical significance in the differences in the product lengths between PR- and NR-PKS, and PR- and HR-PKS but no statistical significance in the differences in the product lengths between NR- and HR-PKS. Other statistical tests comparing the physicochemical properties yielded insights in the differences between the three PKSs. Comparison in the GRAVY score suggests that the non-reducing and reducing PKSs differ significantly in hydrophobicity. However, statistical tests do not support the differences in the cavity volume from the three different PKS. This suggests that non-reducing PKSs and reducing PKSs have different hydrophobic environments in the domain cavity to accommodate the hydrophobicity of the extending polyketide chain.

We also demonstrate for the first time, a regression and machine learning based approach to predicting the polyketide product length based on the modelled KS domain physicochemical characteristics other than the KS domain cavity volume. The relationship between KS domain cavity volume and polyketide product length is not entirely linear as suggested by previous literature, as the correlation R^2 score is low. Instead, we required non-linear regression techniques to derive a predictive model from the dataset. The non-linear relationship is

explored through tree-based regressions, such as decision tree, random forests, and boosted tree algorithms. This approach was promising because even the worst performing machine learning model performed better than or were on par with the best MLR in terms of RMSE and R^2 .

While the current machine learning models have a decent R^2 scores and relatively low RMSE values compared to linear regressions, the models are stochastic and require more samples for better confidence in the accuracy of the model. Increasing the number of features for model training without increasing the number of samples available for training will lead to overfitting of the model. As such, improving the model will require more research in experimentally characterizing iPKS and its products.

Future work will also involve expanding the number of features by calculating other physicochemical properties of the KS and other domains. The training dataset can also be improved by extending the model to modular type I PKS and NRPS and their products. With more data and research, the model will be robust enough to predict the length of products for any iPKS – even new unseen iPKS with unknown structures in the future.

Chapter 2: Structural characterization of partially reducing iterative polyketide synthase NcsB

2.1 Introduction

Polyketides are a large class of structurally diverse natural products produced by bacteria, fungi, and plants. Many polyketides exhibit pharmaceutically attractive biological activities such as antimicrobial, antitumor, and immunosuppressing activities. Polyketides are synthesized by type I, type II, or type III PKSs¹¹⁷. Type I PKSs, which feature multiple catalytic domains, can be further divided into type I modular PKS (modPKS) and type I iterative PKS (iPKS). Type I modPKSs assemble their product using multiple PKS modules, each composed of several catalytic domains, whereas type I iPKSs synthesize their products iteratively using a single module.

The ketosynthase (KS) domain and acyltransferase (AT) domain form a condensing region to elongate the polyketide precursor. The AT domain loads extender units made up of acyl-CoA derivatives onto the acyl carrier proteins (ACP), which in turn requires phosphopantetheine modification for its shuttling function. The phosphopantetheine modification is either facilitated by an external phosphopantetheinyl transferase (PPTase) or an internal PPTase domain as a part of the PKS. The KS domain then catalyzes the lengthening of the polyketide precursor via a decarboxylative Claisen condensation with the acyl extender unit bound to ACP. While the KS, AT, and ACP form the minimal PKS, other domains such as ketoreductase (KR), dehydratase (DH), thiohydrolase (TH), and enoylreductase (ER) modify the polyketide precursor. The various combination of these modifying domains results in a variety of polyketide products with varying levels of oxidation and modification. Despite sharing the same repertoire of catalytic domains with type I modular PKSs, iterative type I

PKSs feature intriguing programmed reduction, dehydration, or methylation mechanisms that remain to be fully understood^{2,118}.

Recent X-ray crystallographic and electron microscopic (EM) studies have yielded a wealth of information regarding the full-length structures of type I modPKS PikAIII¹¹⁹ and fully-reducing iPKS mycocerosic acid synthase-like polyketide synthase (MAS-like)¹²⁰, in the backdrop of the mammalian fatty acid synthase (mFAS) structure reported earlier¹²¹. Despite significant sequence homology shared by the individual catalytic domains of PikAIII, MAS-like, and mFAS, the three proteins exhibit very different dimerization interface and domain organization. Given the differences observed for the three full-length protein structures, characterization of structures of other types of PKSs is essential to fully appreciate the divergent structure and function of PKSs. In this study, we examine partially-reducing iterative polyketide synthase NcsB.

Streptomyces carzinostaticus NcsB produces the hexaketide-derived 2-hydroxyl-5-methyl naphthoic acid moiety (Figure 25, red box) of antitumor compound neocarzinostatin (NCS), an enediyne polyketide with strong antitumoral activity. NCS consists of four parts – naphthoic acid, an enediyne warhead, *N*-methyl fucosamine, and an ethylene carbonate unit¹²². NCS binds to DNA in the cell, and the epoxide in the enediyne warhead (Figure 25, black box) forms a benzyne from undergoing the Bergman cyclization which then cleaves the bound DNA¹²³. In the clinical setting, neocarzinostatin was conjugated to a synthetic copolymer poly(styrene-co-maleic anhydride) (SMA) to form SMANCS. SMANCS was the first approved nanomedicine and used for the treatment of hepatocellular carcinoma¹²⁴. SMANCS is currently the only form of neocarzinostatin that is in clinical use.

Figure 25. The chemical structure of neocarzinostatin. The naphthoic acid moiety is boxed in red. The enediyne warhead is boxed in black, with the epoxide at the right tip. Figure modified from Popik Group (<http://vpopik.uga.edu/Enediynes.html>)

NcsB shares high sequence similarity with several other partially reducing bacterial iterative PKSs such as *Micromonospora echinospora* CalO5 and 6-MSAS in *Penicillium patulum*¹²². The identity and similarity of NcsB and other bacterial iPKS are summarized in Table 7.

Table 7. NcsB and homologous bacterial iPKS.

Gene	Accession number	Organism	Identity%/Similarity%
SACE5532	WP_009944722	<i>Saccharopolyspora erythraea</i>	51/65
AziB	ABY83164	<i>Streptomyces sahachiroi</i>	52/66
PokM1	ACN64831	<i>Streptomyces diastatochromogenes</i>	49/62
ChlB1	AAZ77673	<i>Streptomyces antibioticus</i>	47/61
MdpB	ABY66019	<i>Actinomadura madurae</i>	48/60
CalO5	AAM70355	<i>Micromonospora echinospora</i>	49/62
AviM	AAK83194	<i>Streptomyces viridochromogenes</i>	47/60

Figure 26. Partially-reducing iPKS NcsB and its catalytic activity. KS: ketosynthase, AT: acyltransferase, TH: thiohydrolase, KR: ketoreductase, ACP: acyl carrier protein.

NcsB consists of five domains – ketosynthase (KS), acyltransferase (AT), thiohydrolase (TH), ketoreductase (KR), and acyl carrier protein (ACP). NcsB does not possess an internal PPTase, unlike the other iterative PKS NcsE from the neocarzinostatin and other enediyne biosynthetic pathways. In the experimental setting, 4'-phosphopantetheinyl transferase Sfp is co-expressed to facilitate the phosphopantetheinylation of the ACP domain for polyketide synthesis¹²⁵. To produce 2-hydroxyl-5-methyl naphthoic acid (NPA), NcsB requires one unit of acetyl-CoA as a starter unit and five units of malonyl-CoA as extender units.

The NPA biosynthesis reaction cycle is initiated by the loading of the starter unit, the acetyl moiety from acetyl-CoA, to the KS domain. The extender unit, the malonyl moiety from malonyl-CoA, is loaded onto the AT domain, where it is transferred to the ACP. The location of ACP facilitates this transfer between these domains. Following that, the unit attached to the KS domain undergoes Claisen condensation, extending the polyketide chain. The chain tethered onto ACP is then moved to KS, where it is ready for the next cycle of extension. If the polyketide chain is of a certain length, it undergoes ketoreduction performed by the KR domain. When the length of the polyketide chain is long enough, it is released by the TH domain, forming the final compound naphthoic acid (NPA). The relative positions of the

domains in 2D were simplified from original 3D positions for clarity of the reaction (Figure 27).

Figure 27. A simplified 2D reaction cycle of NPA biosynthesis in NcsB and the domains involved.

There is a need to characterize the partially-reducing iPKSs full-length structures, as fully-reducing iPKS and mFAS produce wholly different products, i.e. mFAS and MAS-like produce long chained molecules, as opposed to the partially-reducing iPKS NcsB which

produces cyclic products. We hypothesize that the difference in products formed by iterative polyketide synthases requires a vastly different architecture.

2.2 Aims and objectives

The aim of the project is to characterize the structure of NcsB and elucidate its architecture.

The architecture of NcsB provides insights into its enzymatic mechanism and catalytic environments. The objectives of this study are the following:

- Express NcsB in *E. coli* expression system for protein purification
- Purify NcsB using chromatographic methods
- Perform characterization of NcsB biophysical properties using SAXS and DLS
- Use electron microscopy techniques to obtain electron micrographs of NcsB
- Reconstruct 3D models of NcsB using 3D reconstruction techniques

These insights guide future engineering efforts to expand the chemical space for neocarzinostatin derivatives with potentially tuneable antitumoral potency for improved clinical applicability.

2.3 Materials and methods

2.3.1 Reagents and chemicals

Coenzyme A (CoA), malonyl-CoA, acetyl-CoA, NADPH, and other chemical compounds were purchased from Sigma-Aldrich and stored at -20°C. Cross-linking reagents disuccinimidyl sulfoxide (DSSO) and disuccinimidyl suberate (DSS) were purchased from Thermo Fischer Scientific. The crystallization screening kits were purchased from Hampton Research and Rigaku Reagents. BL21(DE) *E. coli* strains used for protein expression were obtained from Novagen. pET28b-NcsB and pCDF-Sfp were obtained from Dr Liang Zhao-xun's lab in Nanyang Technological University.

2.3.2 Expression and purification of NcsB

pET28b-NcsB and pCDF-Sfp were co-transformed into *E. coli* BL21(DE3) chemocompetent cells. The cells were plated on 2TY plates supplemented with 50 µg/mL of kanamycin, chloramphenicol, and streptomycin. A single colony was used to inoculate 50 mL of LB medium supplemented with 10% (v/v) glycerol. The culture was incubated in a shaking incubator overnight at 37°C at 180 rpm. 8 mL of the culture was transferred to 800 mL of LB medium supplemented with 10% (v/v) glycerol, 50 µg/mL of kanamycin, chloramphenicol, and streptomycin. The culture was incubated at 37°C at 200 rpm until OD₆₀₀ reached 0.6 before it was cooled to 16°C and induced with 0.5 mM IPTG. The culture is further incubated at 16°C for another 16-20 hours before harvesting by centrifugation.

The cell pellet was resuspended in lysis buffer (50 mM NaH₂PO₄ (pH 8.0), 300 mM NaCl, 20 mM imidazole, 5 mM β-ME and 10% (v/v) glycerol) and lysed using French type pressure cell. After centrifugation at 22,000 rpm for 30 minutes at 4°C, the supernatant was filtered through 0.45 µm membrane filters and loaded onto a HiTrap™ Ni²⁺-NTA column (GE

Healthcare). The column was then washed with the lysis buffer before elution with elution buffer containing 500 mM imidazole. The elute was dialyzed in dialysis buffer (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 5 mM β -ME and 10% (v/v) glycerol) overnight at 4°C. The dialyzed sample was then loaded onto a HiTrap™ Q HP column (GE Healthcare). The loaded sample was eluted via gradient elution with elution buffer containing 1M of NaCl to separate NcsB from Sfp.

NcsB was then purified by size-exclusion chromatography using a HiLoad™ 26/60 Superdex™ 200 column (GE Healthcare) equilibrated with the final buffer (20 mM Tris-HCl (pH 8.0), 150 mM NaCl, and 5 mM β -ME). NcsB molecular weight was approximated using the Superdex 200 10/300 GL column (GE Healthcare). NcsB elution volume was compared to a prepared standard curve.

2.3.3 Small-angle X-ray scattering

The SAXS data were collected with a BRUKER NANOSTAR SAXS instrument equipped with a Metal-Jet X-ray source (Excillum) and VÅNTEC 2000-detector system. The scattering patterns were measured using a sample to detector distance of 0.67 m and a wavelength of $\lambda = 1.3414 \text{ \AA}$, which covered a range of momentum transfer of $0.016 < q < 0.4 \text{ \AA}^{-1}$ ($q = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle). The SAXS experiments were carried out at 15°C with a sample volume of 40 μl in a vacuum-tight quartz capillary. Radiation damage to the protein sample was monitored with six 5 min exposures and no radiation effect was observed. The data were normalized to the intensity of the transmitted beam. The scattering of the buffer was subtracted, and the difference curves were scaled with protein concentration. All the data processing steps were performed using the program PRIMUS19 from the ATSAS package. The experimental data obtained for all protein samples were analyzed for

aggregation using the Guinier region. The forward scattering, $I(0)$, and the radius of gyration, R_g , were computed using the Guinier approximation. These parameters were also computed from the extended scattering patterns using the indirect transform package GNOM, which provided the distance distribution function, $P(r)$, the maximum particle dimension, D_{max} , and the radius of gyration, R_g . The qualitative particle motion was inferred by plotting the scattering patterns in the normalized Kratky plot. The oligomeric state of the protein was confirmed from the molecular mass calculation based on Porod volume (V_p) and the volume of correlation (V_c). The experiment was assisted by Dr Sony of Gruber group.

2.3.4 Dynamic light scattering

Dynamic light scattering (DLS) of NcsB was carried out using a Malvern Zetasizer Nano ZS spectrophotometer (Malvern Instruments). DLS experiments were performed in a low-volume quartz batch cuvette using 12 μ L of NcsB. After 60 sec. equilibration time, the backscattering at 173° was collected for all proteins. Scattering intensities were analyzed using the in-built software, Zetasizer to calculate the hydrodynamic diameter (D_H), size, and volume distribution.

2.3.5 Negative stain EM grid preparation and staining

Carbon-coated copper grids were glow discharged in air for 30-60 sec. and 2 μ L of sample was applied onto the grid, incubated for 30s, and blotted away using Whatman No. 1 paper. 1% uranyl acetate was then applied on the grid, incubated for 30 sec., followed by blotting of the excess stain. Data was acquired using a Tecnai T12 transmission electron microscope operating at 120 kV in a defocus range of 1.0 - 2.0 μ m and recorded at 49,000 $\times$ nominal magnification on an FEI Eagle 4k x 4k camera, resulting in a final sampling of 2.11 $\text{\AA}$ /pixel.

2.3.6 2D and 3D reconstruction of NcsB

Particles were picked semi-automatically with *e2boxer.py* from the EMAN2 software package using a square box size of either 76 or 90 pixels. A total of 2,243 NcsB dimeric particles were manually picked from 72 micrographs; and a total of 40,934 of NcsB monomeric particles were semi-automatically picked from the same set of micrographs.

NcsB dimer particles were sorted into 25 classes using MSA-based reference-free 2D classification with *e2refine2d.py*, and an initial model was generated with *e2initialmodel.py* using those initial classes. The best model that was consistent with the 2D classes was selected as the model for 3D refinement. The final model of NcsB dimer was obtained using *e2refine_easy.py* by refining the initial model using the particles.

NcsB monomer particles were sorted into 50 initial classes using *e2refine2d.py* for quality inspection. After that, the monomer particles were processed using the maximum likelihood-based reference-free 2D classification using RELION 1.4. The classification was performed for 25 iterations with a 140-pixel mask at 2.11 Å/pixel. The generated class averages were visually inspected for incorrectly selected particles (i.e. possible degradation products) and overlapping particles and large aggregates, which were then removed from the dataset, leaving 27,417 particles.

The initial model for NcsB refinement was constructed using random conical tilt (RCT), with *e2rct.py*. 28 micrographs were collected along with the 40°-tilt pair. From the 28 micrographs, 1,418 particle pairs were manually picked. The particles were then classified into 25 classes and used as reference for RCT model generation. The random conical tilt initial volume generated by EMAN2 was used for 3D classification against 27,417 particles,

yielding a set of five preliminary 3D maps. The preliminary 3D maps were inspected, and particle sets were picked from maps which agreed with the 2D class averages. A representative preliminary map was then low-pass filtered to 60 Å and used as a starting model for 3D auto-refinement against 12,040 NcsB particles to obtain the final model of NcsB monomer. Reprojections from the two structures showed agreement with the reference-free class averages.

2.3.7 Homology domain generation and fitting

I-TASSER¹²⁶ was used to generate homology models of KS (PDB ID: 4MZ0¹⁰⁹), AT (PDB ID: 4RL1¹²⁷), TH (PDB ID: 4LN9¹²⁸), KR (PDB ID: 4HXY¹²⁹), and ACP (PDB ID: 4HKG). Visualization and docking of PKS domain homology models were performed using UCSF Chimera²³. The map was segmented into parts using SEGGER¹³⁰, and domains were manually fitted into each part via rigid fitting.

2.3.8 Cross-linking mass spectrometry of NcsB

NcsB was buffer exchanged into 20 mM HEPES pH 7.5 at a final concentration of 10 µM. DSSO and DSS were dissolved to a stock concentration of 50 mM. Cross-linking for DSSO was performed at 50- to 100-fold molar excess whereas DSS was performed at 20- to 50-fold molar excess of NcsB. The reaction was quenched by adding Tris pH 8.0 to a final concentration of 20 mM. The cross-linked NcsB samples were then separated using a 7.5% SDS-PAGE, and bands corresponding to dimeric NcsB (~400 kDa) and monomeric NcsB (~200 kDa) were excised and prepared for cross-linking mass spectrometry. The cross-linking mass spectrometry of the cross-linked mass spectrometry was performed by the Proteomic Core Facility by Sze Group of Nanyang Technological University.

2.4 Results

2.4.1 Purification of NcsB

Nickel affinity chromatography was performed as the first step in the purification of bacterial lysate containing co-expressed NcsB (~190 kDa) and Sfp (~30 kDa) (Figure 28).

Figure 28. Nickel affinity chromatograph of NcsB + Sfp. (Inset). SDS-PAGE gel results of the purification with the corresponding fractions in red line.

The nickel affinity purification profile of NcsB showed contained heterogenous fractions of NcsB and Sfp. The fractions contained unspecified impurities occurring at 50 kDa. Fractions that contained NcsB and Sfp were collected, concentrated using a 30 MWCO spin concentrator, and prepared for anion exchange chromatography.

Figure 29. Anion exchange chromatography of NcsB + Sfp. (Inset). SDS-PAGE gel results of fractions in peak 1 (blue) and peak 2 (red).

Anion exchange chromatography was performed in steps of increasing concentration of NaCl (Figure 29). Two peaks were observed in the chromatogram – the first peak eluted at around 20% elution buffer (~200 mM NaCl) and contained Sfp as well as impurities from the nickel affinity chromatography (Figure 29, inset, blue line). The second peak eluted at around 40% elution buffer (~400 mM NaCl) and contained mostly NcsB. Only fractions containing NcsB were collected for gel filtration chromatography

Figure 30. Gel filtration chromatography of NcsB. (Inset). SDS-PAGE gel results of fractions in the peak (red line).

The gel filtration chromatography using Superdex 200 10/60 column yielded homogenous fractions of NcsB (Figure 30) eluting at a volume of 70 mL. The fractions containing NcsB were concentration again and snap frozen in liquid nitrogen for further experiments or storage at -80°C .

The molecular weight of the purified NcsB was estimated by gel filtration chromatography using Superdex 200 10/300 (Figure 31). NcsB eluted at 10 mL, which corresponded to around 450 kDa based on calibration standard. This would suggest that NcsB is a dimer, and similar to other PKSs which are also dimeric in solution.

Figure 31. Gel filtration chromatography using Superdex 200 10/300 GL Increase. NcsB elutes at around 10 mL.

Despite the anomalous molecular weight, the purified NcsB was free of contamination and the SDS-PAGE band was present at the expected size (Figure 32A). The purified protein was also functional when tested by members of the Liang Group using an NADPH assay, and HPLC to assess whether the polyketide product was produced (Figure 32B, C).

Figure 32. Biochemical assays for NcsB activity. (A) SDS-PAGE results of NcsB protein. (B) NADPH absorbance assay. NADPH is consumed during naphthoic acid synthesis, which results in decrease at 260 nm. (C) HPLC retention time for NcsB product, naphthoic acid. Data provided by Liang Group.

2.4.2 SAXS and DLS analysis

Two other experiments were conducted to assess the molecular weight and dimensions of NcsB: small-angle X-ray Scattering (SAXS) and dynamic light scattering (DLS). SAXS was performed on purified NcsB to determine the size and shape in solution and to provide information on the radius of gyration (R_g), maximum particle dimension (D_{max}), conformation, and assembly state. The results of the SAX experiment are summarized in Table 7.

Table 8. Data collection and scattering derived parameters for NcsB. Data provided by the Gruber Group.

Sample details	NcsB		
Organism	<i>Streptomyces carzinostaticus</i> subsp. <i>neocarzinostaticus</i>		
Genbank	AAM77986.1		
Extinction coefficient (A ₃₈₀ , M ⁻¹ cm ⁻¹)	236290		
MM from chemical composition (Da)	187220		
Solvent	20 mM Tris-HCl (pH 8.0), 150 mM NaCl and 5 mM β-ME		
Data collection parameters			
Instrument (source & detector)	Bruker NanoStar equipped with MetalJet eXcillum X-ray source and VÅNTEC-2000 detector		
Beam geometry	100 μm slit		
Wavelength (Å)	0.134		
q range (Å ⁻¹)	0.016 - 0.4		
Exposure time (min)	30 (6 frames x 5 min)		
Temperature (K)	288.15		
Concentration range (mg·ml ⁻¹)	0.7	1.7	2.5
Structural parameters			
I(0) (arbitrary units) [from P(r)]	726.10 ± 11.14	725.70 ± 5.50	719.00 ± 4.50
R _g (Å) [from P(r)]	63.44 ± 0.76	64.48 ± 0.35	64.56 ± 0.28
I(0) (arbitrary units) (from Guinier)	721.78 ± 39.64	736.40 ± 22.68	744.88 ± 18.33
R _g (Å) (from Guinier)	62.16 ± 3.53	63.85 ± 1.93	65.00 ± 1.53
D _{max} (Å)	197 ± 10	197 ± 10	197 ± 10
Porod volume estimate (V _p) (Å ³)	~758,526	~763,379	~738,0561

To ascertain whether there was concentration-dependent change in particle size, NcsB was measured at three different concentrations of 0.7, 1.5, and 2.5 mg/ml (Figure 33). Upon measurement, a Guinier plot ($\ln I(q)$ vs. q^2) was plotted where the square of momentum transfer q was plotted against the natural logarithm of the scattering intensity $I(q)$.

The full equation is given as such:

$$\ln(I(q)) = \ln(I_0) - (R_g^2/3) q^2$$

where $I(q)$ is scattering intensity, R_g radius of gyration, and q is momentum transfer.

The Guinier plot was plotted for low q values and the plot appeared linear and revealed good data quality (Figure 33, inset).

Figure 33. Solution X-ray experimental scattering pattern. Solution X-ray experimental scattering pattern ($\circ$) of NcsB at concentrations of 0.7 (black), 1.7 (green) and 2.5 mg/ml (blue). (Inset) Guinier plots show linearity, indicating no aggregation. The scattering profiles are offset for clarity by applying arbitrary scale factors.

This indicates that there was no aggregation or concentration-dependent increase in particle size¹³¹. From the slope of the linear fit, the experimental R_g -value was derived to be 65.00 ± 1.53 Å at a protein concentration of 2.5 mg/ml (Table 7).

Figure 34. Overlapping of pair-distance distribution function $P(r)$ of NcsB at the three different concentrations.

The extended scattering curve was converted using the indirect Fourier transform to provide the distance distribution function ($P(r)$), which is a histogram of distances between all possible pairs of atoms within a particle. The $P(r)$ of NcsB exhibited a local maximum with a right-skewed distribution (Figure 34) with a D_{max} of 197 ± 10 Å, indicating an elongated particle in solution. The R_g -value extracted from the $P(r)$ function (54.56 ± 0.28 Å) agreed with the R_g -value extracted from the Guinier region (Table 7). The $P(r)$ of NcsB adopts the same distribution and shape in the three different concentrations.

Figure 35. Normalized Kratky plot of NcsB. The plot contains NcsB (●; blue) and the compact globular lysozyme (●; grey) with a peak (---; grey), which represents the theoretical peak and assumes an ideal Guinier region of a globular particle.

The scattering curve was transformed to a normalized Kratky plot¹³². NcsB displayed a profile that was slightly shifted from a well-defined bell curve profile, usually seen in a globular protein like lysozyme, and indicates that NcsB has a slightly extended conformation (Figure 35). The molecular mass was calculated for NcsB based on the volume extracted from higher angle of scattering data and the volume of correlation. It was calculated to be 461 ± 40 kDa and 395 ± 40 kDa using Porod volume and volume of correlation (V_c), respectively. The data indicate that recombinant NcsB exists as a dimer in solution at the tested range of concentrations (Table 7). Based on the SAXS data, NcsB exists as dimer in solution in the three concentrations tested.

Figure 36. DLS profile of NcsB in solution. Solid line the DLS profile for NcsB at 0.7 mg/mL and dashed line for NcsB at 2.0 mg/mL.

A second biophysical approach, dynamic light scattering (DLS), was used to study the oligomer formation of NcsB. The hydrodynamic diameter of the particle was determined to be 18.04 ± 3.60 nm and 20.11 ± 5.32 nm at 0.7 mg/ml and 2.0 mg/ml respectively (Figure 36). The DLS profile also revealed a dimeric species with an estimated molecular weight of 536 ± 123 kDa and 646 ± 183 kDa at 0.7 mg/mL and 2.0 mg/mL respectively.

Together, the measured diameter and estimated molecular weight agree with SAXS findings, confirming that NcsB was dimeric in solution, which is consistent with characteristics of a PKS, i.e. dimeric due to the KS domain¹³³. A possible reason for the anomalous elution volume is NcsB taking on a globular shape that is not spherical, which affects the migration throughout the column.

2.4.3 Negative stain EM data collection

Structural characterization using X-ray crystallography was unsuccessful because NcsB was recalcitrant to crystallization efforts (data not shown). As the molecular weight of NcsB and

its dimensions were suitable for electron microscopy (EM), EM was used to visualize NcsB directly.

Figure 37. Representative view of the NcsB negative stain EM grid. There are single globular shapes, paired globular shapes, and aggregates present. Scale bar, 50 nm.

The NS-EM results show that the particles in the electron microscopy images were mostly monodisperse, with a contrast showing discernible particles (Figure 37). The particles were mostly globular, slightly elongated, and in random orientations. There were single particles and particles which occurred in pairs, indicating that there was a mixed population of NcsB monomers and dimers. This mixed population may be caused due to negative staining as the uranyl acetate dye is strongly charged and can cause artifacts¹³⁴. The longest dimension of the NcsB monomer measured at 10 nm and the NcsB dimer measured at close to 20 nm. In addition, there was a relatively higher population of NcsB monomers compared to NcsB dimers. The dimensions of the NcsB dimer in NS-EM agree with both the SAXS and DLS data. The dimeric NcsB has an elongated shape which may account for the anomalous elution volume in the gel filtration chromatography. On the other hand, monomeric NcsB is globular and slightly rectangular in shape. For data analysis, 72 micrographs were collected for subsequent image processing.

2.4.4 Negative stain EM data processing

2,243 NcsB dimer particles were manually picked for 2D averaging into 25 classes (Figure 38). The classification also allowed for removal of bad classes, such as protein aggregates, and visual inspection of the heterogeneity of the dataset.

Figure 38. 2D classification of NcsB dimer particles. The circled 2D projection shows the two masses in contact via their “tips”.

The resulting 2D class averages displayed two globular densities closely attached together (Figure 38). There is a slight protrusion out of the two globular densities at the point of contact, and the resulting shape resembles a cassette tape or a binocular symbol (Figure 38, circled class). There were also classes containing one particle that suggests that it contains either only monomeric NcsB or a different view of the particle. Those particles were not chosen for removal in the case that these globular classes represented a genuine view of the dimer.

The dimer particles in the 2D classes were measured to be roughly 180 Å at its longest dimension. The NcsB dimer shape is consistent with the observation that it is globular and elongated and agrees with the D_{max} measurement of 197 ± 10 Å determined for NcsB in solution by the SAXS data and the hydrodynamic diameter of 18.04 ± 3.60 nm determined by the DLS data.

Figure 39. Representative NcsB classes with manual outline tracing. The trace (red outline) of the NcsB classes that identify one and two grooves.

The shape of the NcsB dimer class was manually traced (Figure 39). The trace suggests that the classes have either one or two grooves. It is possible that these two classes represent different views of the dimer. Another feature of the classes is that there is a dark spot in the center of the 2D classes, which suggests that there is a cavity present in the particle.

Figure 40. Representative 2D classes of NcsB monomer.

2D classification of the NcsB monomers was performed as well. The particles were semi-automatically picked, and larger numbers of the monomer particles were obtained from the micrographs as a result. A total of 40,934 of NcsB monomers was picked from the same set of micrographs and classified into 50 classes. The resulting 2D classes of the monomers from 2D classification showed a globular shape, with a small defined “tip” extending out of the main body (Figure 40). This observation is consistent with 2D classes of NcsB dimers, i.e. two particles joined at their tips.

As the monomers were picked semi-automatically, a large population of the particles were not suitable for 3D modelling, e.g., aggregates, and were removed through two rounds of 2D classification using RELION 1.4. The final population of particles used in subsequent analyses was 12,040.

2.4.5 NcsB dimer initial model reconstruction

The generated 2D classes were used to construct the initial models for 3D refinement.

EMAN2 *e2initialmodel.py* generated initial models for the NcsB dimer. Out of the generated initial models, the models that were picked were the ones with the highest score and most consistent with the 2D classes (Figure 41).

Figure 41. Initial 3D model of NcsB dimer constructed using 2D classes. Front and bottom views are shown.

The initial model for the NcsB dimer was asymmetrical and elongated, resembling two cherries stuck together at their stalks. The initial models were used for the 3D refinement step to obtain the final 3D models.

2.4.6 NcsB monomer initial model reconstruction

As the particles of NcsB monomer were small, random conical tilt reconstruction was used to reconstruct the NcsB monomer. Micrographs containing the NcsB monomer were collected and for each micrograph, a corresponding 40°-tilt images was taken (Figure 42).

Figure 42. The micrograph and its tilted view. The numbered blue squares from the left panel corresponds with the numbered green squares.

Figure 43. Representative particles with tilt pairs and 2D classes. (A) Particles (columns 1, 3, 5) with its tilt pair (columns 2, 4, 6). (B) 2D classes of the picked particles.

The monomer particles were mono-dispersed and globular (Figure 43A), and the 2D classes of the picked particles constructed using EMAN2 (Figure 43B) were consistent with the 2D classes from RELION 1.4 and there were no anomalous 2D classes generated using the particles. The 2D classes and paired particles were then used to construct the RCT model (Figure 44).

Figure 44. Random conical tilt model reconstruction of NcsB.

The reconstructed model is globular and resembles a number “6”, with a minor cavity in the center. The volume agrees with the 2D classes and the individual particles. The volume was then used for 3D classification in RELION 1.4.

2.4.7 NcsB dimer and monomer negative stain EM structures

Using the initial model, the dimer was refined against using the dimer particles. The full 3D reconstruction of the dimeric NcsB revealed an elongated volume with a dimension of 86 Å x 168 Å x 87 Å (Figure 45).

Figure 45. 3D reconstruction of the NcsB dimer using the dimer particles, refined to a resolution of about 18.6 Å. “Front” view (left panel) and “bottom” view (right panel) display a notched cavity, along with the corresponding 2D class (left) and 2D project of 3D volume (right) below each view.

The dimer structure is slightly asymmetrical, with one half of the volume slightly larger than the other side. The 3D reprojections and 2D class averages for the NcsB dimer were in good agreement. The volume also contains a notch in the center, which also agrees with grooves identified in the 2D class averages of dimeric NcsB.

Figure 46. Refinement progress for dimeric NcsB using non-NcsB model. PikAIII cryo-EM model (EMDB ID: 5647) was used as model for refinement against the dimeric NcsB particles over ten cycles.

To ensure that there was no *ab initio* model bias during refinement, the 3D model of PikAIII dimer was used in refinement as well (Figure 46). The resulting model converged to the same elongated conformation as the *ab initio* model over 10 refinements. The notch in the center of the dimer returned, and the reaction chamber in PikAII was also filled up with a density at the end of the refinement.

Figure 47. 3D reconstruction of the NcsB monomer using monomeric particles, refined to a resolution of 14.8 Å. Solid rendering of NcsB monomer. Below each view is the respective class averages (left) and corresponding reprojected images (right).

The 3D reconstruction of the monomeric NcsB using a final set of 12,040 particles revealed an overall globular pear-shaped body (Figure 47), at a dimension of roughly 77 Å x 110 Å x 60 Å. The EM model did not have any observable symmetry and had a roughly triangular base. This suggested that the domains in NcsB monomer were tightly packed, except one domain that is protruding out of the structure. The 3D reprojection and class averages were in good agreement (Figure 47, bottom panels).

Figure 48. Fourier shell correlation (FSC) graph of NcsB monomer and dimer structures. Density map resolution calculated for the FSC value of approximately 0.143. Global resolution based on FSC for dimer structure is $1/0.538 = 18.6$ Å and monomer structure is $1/0.675 = 14.8$ Å.

The final resolutions of NcsB dimer and NcsB monomer were 18.6 Å and 14.8 Å, respectively (Figure 48). While the resolutions of both monomeric and dimeric structures were low, the respective models' conformation were distinct enough to observe the overall shapes. As such, the monomeric EM models were manually positioned into the dimeric model.

Figure 49. Solid rendering of two monomeric NcsB fitted inside the meshed representation of dimeric NcsB.

To probe the domain organization of NcsB, the reconstructed monomers were fitted into the dimer structure. Due to the lower resolution of dimeric NcsB, the shape was not as defined as that of monomeric NcsB. However, there was enough space in the dimeric NcsB to fit two monomers (Figure 49). Based on the 2D projections, the two monomers were arranged in a clockwise point symmetry around monomers' tips which were arranged to be in contact with one another. The arrangement of the two monomers accounted for the notched middle in the dimer structure.

2.4.8 Homology modelling and domain arrangement

Homology modelling server I-TASSER was used to model the five NcsB domains given the lack of crystal structures. The individual domains in NcsB have high homology with the published modular type I PKS structures. The KS domain was modelled using the KS domain from modular PKS CurL of the curacin A biosynthesis, the AT domain with the acyltransferase from the avermectin modular PKS, the TH domain with the dehydratase domain from the terminal module of the rifamycin PKS, the KR domain with the first module of the phoslactomycin PKS Plm1, and ACP from a NRPS-containing operon in *Acinetobacter baumannii*. Table 8 summarizes the published structures used in the homology modelling.

Table 9. NcsB domain homology modelling templates. The domains in NcsB and the corresponding templates using in homology modelling by I-TASSER.

Domain	Template PDB	Organism	Purpose
KS	4MZ0	<i>Moorea producens</i> 3L	Sixth module of curacin A synthesis
AT	4RL1	<i>Streptomyces avermitilis</i>	First module of avermectin synthesis
TH	4LN9	<i>Amycolatopsis mediterranei</i>	Terminal module of rifamycin synthesis
KR	4HXY	<i>Streptomyces</i> sp. HK803	First module of phoslactomycin synthesis
ACP	4HKG	<i>Acinetobacter baumannii</i>	Uncharacterized NRPS secondary metabolite pathway

Based on the arrangement of the monomers in the dimer and the fact that the KS domain occurs as a dimer and is involved in the dimerization in polyketide synthases and FAS, the two tips in the monomer models in contact were designated as the KS domain. The monomer volume was then segmented using Segger¹³⁰ to determine the placement of the remainder of the domains (Figure 50). Areas of high EM density were designated as one segment, and six segments were obtained from the volume segmentation (Figure 50A).

Figure 50. Proposed NcsB model with domains. (A) Segmented monomeric NcsB inside the meshed representation of dimeric NcsB. (B) Homology model arrangement in NcsB.

All domains fit reasonably within the EM structure after volume segmentation (Figure 50B). With reference to the KS domain in the left monomer in Figure 46, the subsequent domains were arranged in a counter-clockwise fashion. When viewed from the side, the triangular face is composed of the KR, AT, and TH domains (Figure 50B, right side). Overall, the domains are arranged in a compact manner with the overall structure resembling the number “6”, which is notably different from the cross-hatchet form of mFAS and MAS-like, and the linear conformation of PikAIII.

2.4.9 Cross-linking mass spectrometry

To confirm the validity of the determined domain rearrangement in the fitted NcsB models, cross-linking mass spectrometry (CLMS) was performed. CLMS is an increasingly popular technique to complement structural studies¹³⁵.

Figure 51. General CLMS workflow. Image from O'Reilly and Rappsilber (2018)

In short, a protein complex is incubated with a crosslinking reagent with fixed spacer sequence that provides a distance restraint. These spacers, such as DSS and DSSO, contain reactive NHS esters on both ends and target lysine residues on proteins. After cross-linking,

the protein complex is digested by trypsin into cross-linked peptides and enriched using chromatographic methods. The peptides then undergo mass spectrometry and are analysed for cross-links based on the mass spectrometry spectra (Figure 51). NcsB contains 13 lysine residues on its surface (Figure 52), and most of its lysine residues are located on the ketosynthase domain. There are lysine residues on the ACP domain as well. Interestingly, there were no lysine residues on the TH and KR domain. If the arrangement of the domains is correct, we anticipate only cross-links between the ketosynthase domains and not other domains.

```
>tr|Q84HC6|Q84HC6_STRCZ Iterative type I polyketide synthase
MSCRYAPDLDSPDKFWDFLVNGRSTVGDMPDKRWEPYASGSPQATAAMRDTVRRGAFLDD
IEGFDAEFFGISPREADFLDPQQRFMLELAWEALADAGVPPLTLRNTDTGVYAAANSNDY
GRRLLIEDIPRTGAYAVNGTTLYGIANRVSYFLDLHGSPMAVDACAGALTALHLARQSLL
TGETPLAIVGGLNIMSTPSLNVALNDAGAMSPDGRSKAFDEADADGYGRGEGAGVLVLKRL
SDARRDNDPVHAVIRGSGVFQDGRSDGMMAPDGDQAEHMLRQAYHRAGVDPATVDYVEAH
GTGTPTDGDEEITALAKVFGAGRSPHAPCLIGSVKPNVGHVEGGSGITGVIKTVLALRNE
LIPPTLHDRPRTDWDWDAGVRLVGQVQEWPSGCRPRRAGVSSYGVGGTISHVILEESPV
PAATSSADASTGVRTPALFPLSAASEAGLRALAGEAAGWVASRPDTPLPVSGHTLTQRRS
HLAQRAAVVADSAEQLVDRLEVAEGRNGPGIVSARTSAGRADDAVWVFSGHGAQWSGMG
RRLLESEPVFAATLDALDEVFREELGWTPREAVTEGGPWTAAHVQALFAVQIGLADVWR
SKGLRPGAVIGHSVGEIAAAVVAGSLDRDEAARFACRRAAALQRLDGRGAMVMVGLPFEE
AALRLGDRRDVEAAISAGPHSTVLSGDRSAVLRVAEEWQASEVWTRTVDSIAFHSVHVD
EVTGDIESAARLLTPRPPTVPLYTTALSDPRSAPRDSGYWAANLRKPVRFTEAVRAAAE
DGHRLFLEVSSHPVVAHSVETLLDLGIEDAAVAGTLRRDTDEVESLLENLAELHCHGVA
VDWARHHTDGELVGLPAAVWQHRPYWIFPETTADAGLGRGHDPASHSLLGGRMTVSGSPT
RQVWQTRLDMDSRPYPQSHGLVGVEVTPAASIINTFAAAVEEDGPSALTDIVLRTPLAVE
PPRVVQVREGRSLSLATRAEDADAGSEWITHTAAVTPGVRPAGGRDTEAIRSRLP
EGSLTRADEMFERMGVEGYAFPWDLEELRHDDHEQLAVLQIEPSPAQRATSWAHVIDGAL
TISAMVVSPGDATVLWMSRSIDQVTWSGEPARLTVHSTRSLRSPHDTVDRVADERGDV
VCEVVGRLFAAVEHIGA AVLPSSELVHEIVWRPWPDPEDHEGAEDAPVEQVILVGDEATVP
LAEQLESAGMTCVQVGDSPETGLRPDLFARPGAVVAPALAGSDTAPEEEAERVSWLLVR
TVQRVAEIRADVTGDAPAQRVWCVTSDVRRARDERSVAHGPLWGLARIVAGDHPELWGGA
VDIGPSADGIGARLVALLDGAAGTEDVISLTGEGAEVARLSRIDRSADGTPLQCSPSGTV
LITGGLGALGLEVARWLVDRGARRLVLSRRALPNRTEWPAVTD AETRRRIDGVLAL EAL
GVTVRVLALDITDQVSAALAPEALGLPPVRGVVHAAGVVNNALVDKVDLEGLREVLAP
KVRGAMVLHRLFPFGDLDFVLFSSCGQFARLSGQASYAAANSFLDTLASHRNAGEHTET
VSLGWTAWRGLGMSSNIDTTMFEANSRGL EAVSATEAFGAWSFGDRFQSDYQAILRVVPT
PVHTPRLPVFRDLPVSGETDGPTGDQLFTTTLEGLPEQEARERITADVREQVAGVLNFD
SEVEVKRPLVELGVDSVMTVALRVRLQRRYGLELPPTILWAKPTVAALSEHVCDSLRWDG
EEHGDLAAPTAAA
```

Figure 52. The amino acid sequence of NcsB with highlighted lysine (K) residues. The KS domain is highlighted in blue, KR domain is highlighted in green, and ACP is highlighted in purple.

NcsB was with 50 and 200 μ M of DSS, 50 and 100 μ M of DSSO, with control (NcsB alone). After quenching the reaction with Tris buffer, the cross-linked fractions are separated using SDS-PAGE and visualized (Figure 53).

Figure 53. SDS-PAGE results of NcsB cross-linking with DSS and DSSO. (1) Control. (2) 50 μ M DSS. (3) 200 μ M DSS. (4) 50 μ M DSSO. (5) 100 μ M DSSO. The protein ladder marker maximum size is 270 kDa, followed by 175 kDa. (Red box) The hypothesized band where cross-linked NcsB dimer (~390 kDa).

There were non-specific cross-linking and high molecular weight aggregates, and faint bands were observed at a location that corresponds to around 400 kDa, above the 270 kDa marker (Figure 53, red box) that is not present in the control lane. The gel section containing the cross-linked bands was then excised for further CLMS experiments.

The CLMS experiments and subsequent spectra analysis revealed 27 cross-links for DSSO and 3 cross-links for DSS (Table 9). In the 27 DSSO cross-links, there are two pairs of cross-linked peptides found. The first general peptide pair involves the 32nd and 217th lysine residues in the KS domain and whereas the second peptide pair involves the 1488th and 1501st residue in the KR domain. In the 3 DSS cross-links, only KS domain cross-links are found.

Table 10. Representative cross-linked peptide sequences from the CLMS. The cross-links are for KS and KR domain are shown in red and dark blue lines respectively.

Sequence	Crosslink	Crosslink	Confidence	Leading P	Leading P
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGR-GEGAGVLVLKR	DSSO	Intra	Unknown	217	238
SKAFDEADGYGR-GEGAGVLVLKR	DSSO	Intra	Unknown	217	238
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
FWDFLVNGRSTVGDMPPDKR-STVGDMPPDKR	DSSO	Intra	Unknown	32	32
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGR-GEGAGVLVLKR	DSSO	Intra	Unknown	217	238
SKAFDEADGYGR-GEGAGVLVLKR	DSSO	Intra	Unknown	217	238
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
GVVHAAGVVNNALVDKVDLEGLR-EVLAPKVR	DSSO	Intra	Unknown	1488	1501
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGR-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSSO	Intra	Unknown	217	32
SKAFDEADGYGRGEGAGVLVLK-STVGDMPPDKR	DSS	Intra	Unknown	217	32
SKAFDEADGYGR-GEGAGVLVLKR	DSS	Intra	Unknown	217	238
SKAFDEADGYGR-GEGAGVLVLKR	DSS	Intra	Unknown	217	238

Mathematically, the possible combinations of unidirectional pairs in a population of size N is given by:

$$\text{number of pairs} = n(n - 1)/2$$

where n is the number of items in a sample population, and that the pairings are unidirectional, i.e. residue 217 and residue 32 cross-link is the same as residue 32 and residue 217 crosslinks. As such, in a protein with 13 lysine residues, we expect to see a maximum of $13(13 - 1)/2 = 78$ pairs of crosslinks. However, there were only 4 unique pairs in the CLMS

results. This is because cross-linking experiments involving homodimers are challenging as most CLMS approaches are optimized for heteromers and not for homodimers¹³⁶.

The cross-linking results from CLMS was visualized to assess the validity of the proposed NcsB domain arrangement (Figure 54).

Figure 54. Visualization of the cross-linked pairs in NcsB. Lysine residues are presented as pink ball figures, and the cross-linked pairs of lysine residues are measured (yellow lines).

Most of the cross-links are intra-monomer and are clustered around the KS domain. The cross-links within the KS domain measure 12.62 Å, 23.21 Å, and 34.85 Å. The KR cross-link is measured to be 25.93 Å whereas the lysine residues in the AT domain are too far away from other domains, and as such do not have any cross-links. The ACP lysine residues did not cross-link with any other residues, which can be attributed to the positions of the residues which are facing outwards of the protein.

Interestingly, while the theoretical length of DSSO spacer is 10.1 Å, the cross links are 2-3x that of the spacer length. However, each lysine residue adds 6.4 Å to each side, which is around 23 Å when the full cross-link is measured. On top of that, the dynamic movement of residues and domains in solution can lead to seemingly longer cross-links compared to theoretical calculations¹³⁷. The cross-linking results support that dimerization does occur in the KS domain for NcsB and does not disprove the current model.

2.5 Discussion

Here, we demonstrate for the first time the EM reconstruction model of a partially reducing polyketide synthase NcsB. We obtained only the low-resolution structure of NcsB via negative stain EM which suggests the overall globular shape of the protein. The arrangement of the KS domain follows the symmetry of KS dimers from other PKS KS structures, as the dimerization interface is highly conserved across PKS and its homologs.

Figure 55. A comparison between the existing PKS structures. (A) Cryo-EM structure of PikAIII fitted with corresponding crystal structures (EMDB ID: 5647). (B) The crystal structures of dimeric DH-ER-KR (PDH ID: 5BP4) arranged with dimeric KS-AT (PDB ID: 5BP1). (C) The crystal structure of mammalian FAS (PDB ID: 2VZ8). (D) NcsB (this study). The colours of the domains in the structures are as follows: KS (red), AT (blue), TH (brown), KR (green), ACP (yellow), MAT (silver), ER (cyan), DH (gold), and linkers (grey). Not shown on the mammalian FAS structure are the TE and ACP domains due to their inherent flexibility.

Despite the high homology between NcsB and the rest of the homologs, the final NcsB dimer and monomer models are distinct from known full-length PKSs (PikAIII, MAS-like) and mFAS structures. The AT domain in NcsB is linear to the KS domain, as opposed to the AT in PikAIII that is perpendicular to its KS, leading to an arched architecture. In this respect, the arrangement for KS-AT in NcsB is closer to that of mFAS and MAS where the two domains in those proteins are linearly arranged. Compared to the linker domain that is present

between the AT domain and DH domain in MAS-like and mFAS, NcsB does not contain a long linker domain between the AT domain and the TH domain. As such, NcsB is unable to adopt a cross hatchet conformation as the TH domain is unable to loop above the KS-AT domain.

In the individual NcsB monomers, the catalytic domains are oriented in such a way that the active sites are facing inwards. The closed compact conformation of each monomer provides a reaction chamber for the iterative elongation of the polyketide chain without premature oxidation of the lengthening product. Cross-linking mass spectrometry results indicate that the domains that contain lysine residues do not cross-link, which supports our domain arrangement of the 3D homology structures.

The ACP docking sites on the catalytic domains are located in the proximity of cavity formed by KS, AT, and KR domains, allowing the cavity to serve as the reaction chamber as observed in other PKSs and FASs. Based on the proposed domain arrangement of NcsB, the ACP domain can reach the active sites in the four other domains without ostensible steric hindrance. In the compact conformation, the ACP domain is located close to KS and AT domains to facilitate iterative lengthening of the polyketide chain as the ACP domain can shuttle between KS and AT rapidly while also access KR for ketoreduction and TH for chain termination. Even when we assume that the catalytic domains remain unmoving in catalysis, the travel distance for the ACP domain is within the estimated length of the KR-ACP linker and phosphopantetheinyl arm.

For PikAIII, the two monomers were proposed to entangle each other. The KS domain from one monomer is located in one lateral half of the arch whereas the other domains from the

same monomer were on the other side. On the other hand, NcsB is capable of being monomeric in solution. This implies that NcsB dimers were able to dissociate into individual monomers in solution under low concentration, which suggests that the two monomers do not need to intertwine when forming a dimer. More importantly, while NcsB is dimeric, the proposed architecture suggests that each monomer works independently from the other in polyketide chain lengthening and termination.

Despite extensive studies on individual PKS domains, studying full-length architectures of PKS is still a challenge. Our analysis of the first full-length partially-reducing iPKS NcsB reveals that it adopts a dimeric conformation with a barbell-like architecture containing compact monomers. The conformation is distinctly different from those of modPKS PikAIII, fully reducing iPKS MAS, and mFAS. The architecture and domain organization may play a role in regulating the activity of the enzyme to produce cyclic polyketide products and these results provide new insights into the structure and function of iterative PKSs, further expanding our understanding of the divergence of PKS enzymes.

Chapter III: Structural characterization of highly reducing iterative polyketide synthase DynE8

3.1 Introduction

Enediyne compounds are structurally complex compounds with strong antitumoral activity. These compounds are unique because of their “warheads” – a carbocyclic core with either 9 or 10 carbons, containing a double carbon bond flanked with triple bonds on both sides. These warheads attack and cleave DNA strands *in vitro* through cycloaromatization to an active radical. While these compounds have remarkable efficacy against cancer cells, they often have to be modified for safer usage¹³⁸. Efforts in characterizing the gene clusters of organisms producing enediyne compounds have revealed that these compounds are synthesized by high-reducing polyketide synthases.

Figure 56. Chemical structure of dynemicin A. The left half is the anthraquinone, and the right half is the enediyne core.

Amongst the enediyne compounds, dynemicin A has seen steady research interest since its discovery. Dynemicin A was first isolated in 1989 from *Micromonospora chersina*, a soil bacteria from Gujarat in India¹³⁹. It is composed of two parts: an anthraquinone with a fused

enediyne warhead (Figure 56). It was proposed to bind to the B-DNA minor groove, where the epoxide warhead activates and cleaves the DNA^{140,141}. Dynemicin A is too potent for direct clinical use, and its safer analogues are being explored in the clinical setting instead¹⁴². While the polyketide dynemicin A was discovered in 1989, it was not until much later that the biosynthetic gene cluster (BCG) responsible for dynemicin A production was discovered⁹². In the gene cluster, iterative polyketide DynE8 was discovered to be responsible for dynemicin A synthesis. While DynE8 was first discovered to synthesize the enediyne core, the protein responsible for synthesizing anthraquinone was unknown. It was only a decade later that DynE8 was identified to have dual roles in synthesizing the carbon skeletons of both halves of dynemicin A – both the anthraquinone and the enediyne warhead¹⁴³.

Figure 57. Domain arrangement of highly-reducing iPKS and DynE8.

DynE8 is a highly reducing iterative polyketide synthase that is composed of six domains: ketosynthase (KS), acyltransferase (AT), acyl carrier protein (ACP), ketoreductase (KR), dehydratase (DH), and phosphopantetheinyl transferase (PPTase) (Figure 57). While DynE8 is a highly-reducing iPKS, it is different from another characterized highly-reducing iPKS

MAS-like PKS (MAS), which contains an additional enoyl-reductase domain (ER). The ACP domain is also located in the center of the protein as opposed to the end of the protein. Furthermore, MAS produces long linear fatty acids, as opposed to DynE8 which produces cyclical products. Thus, we hypothesize that DynE8 adopts a structure distinct from the cross-hatched conformation reported for MAS.

Apart from the DynE8, there are other proteins involved in the synthesis of dynemicin A. The thioesterase protein DynE7 cleaves the product from DynE8 for product release and has been structurally characterized¹⁴⁴. Another protein in the dynemicin biosynthetic gene cluster is DynU15. DynU15 (Uniprot accession number Q84HJ1) is a small protein of 328 amino acids. Using the ProtParam tool, its theoretical weight is calculated to be 35.44 kDa with a theoretical pI of 9.94. There is very little known about the DynU15, but it is crucial in dynemicin synthesis as DynU15 deletion mutants abolished dynemicin A production⁹². To date, there is no information on any domains in DynU15. DynU15 is proposed to be an acetylenase for the triple bond formation in the enediyne part of dynemicin A¹⁴⁵.

Given that DynE8 possesses a domain organization that is distinctively different from the other two iPKS from the published structures of modular PKS PikAIII, MAS-like PKS, and animal fatty acid synthase FAS, there is a need to elucidate its structure and understand the structure-function relationship of a polyketide synthase required for perform multiple reduction of the polyketide chain.

3.2 Aims and significance

To date, there are no full-length or partial structures of DynE8 and DynU15. As such, this study aims to understand the atomic structure of both proteins. This will pave way to

understanding how dynemicin A is synthesized. In addition, the study will aim to characterize the putative acetylnase DynU15 and gain insights into triple bond formation for enediynes.

The objectives of this study are the following:

- Express DynE8, DynU15, and co-express both proteins in *E. coli* expression system for protein purification
- Purify the proteins using chromatographic methods
- Optimize DynE8 sample preparation for electron microscopy
- Use electron microscopy techniques to obtain electron micrographs of DynE8
- Reconstruct 3D models of DynE8 using 3D reconstruction techniques

Understanding the structure-function relationship of DynE8 and its dual role in synthesizing the enediyne warhead and anthroquinone moiety will stimulate interest in rational design approaches in expanding the chemical diversity of enediyne analogues for safer and more targeted clinical applications. There are currently no known domains found in DynU15, and structural characterization of DynU15 will present a novel catalytic fold for acetylenase bond formation.

3.3 Materials and methods

3.3.1 Reagents and chemicals

General chemicals used for buffer constitution, unless otherwise stated, were purchased from Sigma-Aldrich. BL21(DE) *E. coli* strains used for protein expression were obtained from Novagen. The R1.2/1.3 and R2/2 grids for electron microscopy were purchased from Electron Microscopy Sciences. The plasmids for expressing DynE8 and DynU15, pET28b-DynE8 and pCDF-DynU15, were obtained from Dr Liang Zhao-xun's lab in Nanyang Technological University.

3.3.2 Expression and purification of DynE8 and DynU15

The plasmid or plasmids was transformed into *E. coli* BL21(DE3) competent cells. The cells were plated on 2TY plates supplemented with either 50 µg/mL of kanamycin, 100 µg/mL of streptomycin, or both for co-expression studies. A single colony was used to inoculate 50 mL of 2YT medium. The culture was incubated in a shaking incubator overnight at 37°C overnight at 180 rpm. 8 mL of the culture was transferred to 800 mL of 2YT medium with 50 µg/mL of kanamycin, 100 µg/mL of streptomycin, or both. The culture was incubated at 37°C at 200 rpm until OD₆₀₀ reached 0.6 before it was cooled to 16°C and induced with 0.5 mM IPTG. The culture is further incubated at 16°C for another 16-20 hours before harvest.

The cell pellet was resuspended in lysis buffer (50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 20 mM imidazole, and 10% (v/v) glycerol) and lysed using French type pressure cell. After centrifugation at 22,000 rpm for 30 minutes at 4°C, the supernatant was filtered through 0.45 µm membrane filters and loaded onto a HiTrap™ Ni²⁺-NTA column (GE Healthcare). The column then washed with the lysis buffer before elution with elution buffer containing 500 mM imidazole. Fractions containing relevant proteins were concentrated with a spin

concentrator. The concentrated proteins were then purified by size-exclusion chromatography using a HiLoad™ 26/60 Superdex™ 200 column (GE Healthcare) equilibrated with the final buffer (20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 5 mM β -ME).

3.3.3 Gradient Fixation (GraFix)

DynE8 was prepared for EM using the GRAFIX method¹⁴⁶. DynE8 (11 mg/mL) was layered on top of a 10 – 30% sucrose gradient (50 mM HEPES pH 8.0, 150 mM NaCl) with or without an increasing concentration of glutaraldehyde (0 – 0.5% w/v) and subjected to ultracentrifugation using a SW60 Ti rotor (Beckman Coulter) at 50,000 rpm for 16 hours in 4°C. Gradients were analysed by SDS-PAGE and fractions containing DynE8 were buffer exchanged into EM buffer (20 mM Tris pH 8.0, 150 mM NaCl) using a 10k MWCO spin concentrator.

3.3.4 Negative stain EM grid preparation and staining

Carbon-coated copper grids were glow discharged in air for 30-60 sec. and 2 μ L of sample was applied onto the grid, incubated for 30 sec., and blotted away using Whatman No. 1 paper. 1% uranyl acetate was then applied on the grid, incubated for 30 sec., followed by blotting of the excess stain. Data was acquired using a Tecnai T12 transmission electron microscope operating at 120 kV in a defocus range of 1.0 - 2.0 μ m and recorded at 49,000 $\times$ nominal magnification on an FEI Eagle 4k x 4k camera, resulting in a final sampling of 2.11 Å/pixel.

3.3.5 2D and 3D reconstruction of DynE8 based on negative stain EM

Particles were picked semi-automatically with *e2boxer.py* from the EMAN2 software package using a square box size of 140 pixels. A total of 6,795 DynE8^{GA+} and 7,040 DynE^{GA-}

particles were semi-automatically picked from 102 and 70 micrographs, respectively. MSA-based reference-free 2D classification with *e2refine2d.py* was used to sort the DynE8 particles into 50 – 100 initial classes, and an initial model was generated with *e2initialmodel.py* using those initial classes. The initial model was subsequently used for 3D classification and refinement in RELION 1.4.

For particle selection in RELION 1.4, the images were first processed to a suitable file format for RELION 1.4 and maximum likelihood-based reference-free 2D classification was performed for 25 iterations with a 140-pixel mask at 2.11 Å/pixel using RELION 1.4. The generated class averages were visually inspected for incorrectly selected particles (i.e. possible degradation products) and overlapping particles or large aggregates, which were then removed from the dataset. For DynE^{GA+} and DynE^{GA-}, the reference-free 3D initial volumes generated by EMAN2 were used for 3D classification and refinement. The preliminary 3D maps were inspected, and particle sets were picked from maps which agreed with the 2D class averages. A representative preliminary map was then low-pass filtered to 60 Å and used as a starting model for 3D auto-refinement against 4,646 and 4,700 particles to obtain the final models of DynE8^{GA+} and DynE8^{GA-} respectively. Reprojections from the two structures were compared to the reference-free class averages for validity.

3.3.6 Cryo-EM grid preparation and data collection

Prior to sample vitrification, the humidity and temperature in the FEI Vitrobot sample chamber was first equilibrated to 100% and 4°C respectively. Both Quantifoil R1.2/1.3 grids with and without carbon coating were used for sample preparation. The grids were glow discharged using a plasma cleaner for 1 minute and then immediately placed in the Vitrobot blotting chamber for use.

For sample freezing, 4 μL of 0.5 – 2.7 mg/mL DynE8 were loaded onto the grid in the blotting chamber. Grids with carbon coating had 30 seconds of incubation time whereas grids without carbon coating were immediately blotted for 2 seconds before plunge-freezing in liquid ethane. Frozen grids were stored in cryo-grid boxes in liquid nitrogen.

Micrographs were collected using an FEI 200 kV Arctica transmission electron microscope. To prevent sample damage, a low dose condition was used for data collection at $\sim 15\text{-}25\text{e}^-/\text{\AA}^2$. Volta Phase plate was used to improve contrast, and images were taken at defocus range. Electron micrographs were collected automatically on the Arctica using an automated single particle collection software EPU at defocus range. 475 and 919 micrographs were recorded in movie mode as a set of 40 frames for a total dose of 1.36805 electrons per $\text{\AA}$ at a $78,000\times$ magnification, which led to a pixel size of 1.28 $\text{\AA}$.

3.3.7 Cryo-EM data processing

A total of 475 and 919 micrographs were collected for data processing. CTF parameters were estimated using CTFFIND3 software. EMAN2 *e2boxer.py* was used to semi-automatically pick particles from the micrographs, and the box coordinates were exported. RELION 1.4 was then used to pick the particles automatically.

For the first and second batches of DynE8 data, 3,518 and 1,461 particles respectively were manually selected and subjected to 2D classification using EMAN2 and RELION. The 2D classes from RELION were used for template-based automated particle picking and 142,317 and 475,654 particles from the first and second batches of micrographs respectively were selected. The particles were then subjected to three more rounds of 2D classification to remove aggregates and ice particles.

3.4 Results

3.4.1 Expression and purification of DynE8

A streamlined purification protocol was developed to purify DynE8. Nickel affinity chromatography was performed as the first step in the purification of DynE8 (Figure 58). DynE8 eluted at around 40% elution buffer volume, i.e. around 500 mM imidazole concentration.

Figure 58. Nickel affinity chromatography results of DynE8. The left side contains the chromatograph from the purification, along with the SDS-PAGE results of the fractions underlined in red.

The nickel affinity purification of DynE8 yielded fractions of DynE8 as observed from the SDS-PAGE results. The 190 kDa DynE8 is present at around the 200 kDa molecular weight marker. However, there were some fractions which contained high molecular weight contaminants. Those fractions were excluded for further purification steps. DynE8 fractions were yellow in color, indicating that there were bound dynemicin precursors. Fractions that contained DynE8 were collected and concentrated using a 30 kDa MWCO spin concentrator

and prepared for gel filtration chromatography using a Superdex 200 16/60 gel filtration column.

Figure 59. Gel filtration chromatography results of DynE8. The left side contains the chromatograph from the purification, along with the SDS-PAGE results of the fractions underlined in red.

The gel filtration chromatography yielded fractions containing DynE8 (Figure 59), although the SDS-PAGE result suggests that some of the high molecular weight contaminants remain. The fractions containing DynE8 were confirmed by observing the color of the protein fraction. DynE8 was concentrated and flash frozen using liquid nitrogen.

3.4.2 Expression and purification of DynU15

Nickel affinity chromatography was performed to obtain DynU15 (Figure 60). A sharp peak with high UV absorbance eluted at 40% elution buffer (400 mM imidazole). The protein was found only in four fractions, and the fractions were examined using SDS-PAGE (Figure 60, inset).

Figure 60. Nickel affinity chromatography results of DynU15. The left side contains the chromatograph from the purification, along with the SDS-PAGE results of the fractions underlined in red.

The molecular weight of DynU15 is around 35kDa, but the four fractions with UV absorbance did not contain strong bands anywhere near the 40 kDa molecular weight in the marker. However, there were thick bands present above 70 kDa for the fractions. It is possible that the fractions contained the DnaK chaperone, a common protein contaminant found during heterologous expression of proteins in *E. coli*¹⁴⁷.

Repeated experiments revealed similar presence of the 70-kDa contaminant. As such, the fractions were pooled and concentrated for subsequent gel filtration chromatography experiments.

Figure 61. Gel filtration chromatography results of DynU15. The fractions under the three peaks are highlighted in red, dark blue, and purple.

Gel filtration chromatogram for DynU15 showed three major peaks (Figure 61). SDS-PAGE results for corresponding fractions indicate that the first peak contains aggregated proteins (eluted as the first peak is at the void volume of the column (40 ml)).

Peak 2 contained the 70 kDa protein contaminant observed in the nickel affinity chromatography (Figure 61, inset, dark blue underline). As such, the gel filtration chromatography step successfully removed the contaminant.

Peak 3 contained a mixture of several proteins of various molecular sizes including small fractions of DynU15. As DynU15 contained an N-terminal hexa-histidine tag, Western blotting was conducted to test the hypothesis that those bands belong to DynU15 (Figure 62).

Figure 62. DynU15 distribution in gel filtration chromatography fractions. Left: SDS-PAGE from the gel filtration chromatography. Right: Western blot.

The Western blot result shows that there were DynU15 present in fractions from all three peaks (Figure 62). Signals were also found at the high molecular bands for Peak 1 and Peak 2. Only half of the fractions in Peak 3 have a chemiluminescent signal, suggesting that the low molecular weight bands between 20 to 30 kDa might be degraded DynU15. The dark bands found at 70 kDa in Peak 2 do not have corresponding signals in the Western blot as well. While the signal for DynU15 is strong on Western blot due to the sensitivity of the technique, the corresponding bands on the SDS-PAGE gel is not as obvious.

3.4.3 Expression and purification of co-expressed DynE8 and DynU15

One possible way to improve DynU15 purification and yield is to co-express it with DynE8. The hypothesis was that these two proteins might interact, as DynU15 is required to modify the polyketide product tethered to DynE8. As such, co-expression studies were conducted by transforming the two plasmids into *E. coli* expression strains for large-scale protein preparation.

Figure 63. Nickel affinity chromatography results of DynE8 + DynU15 co-expression. Fractions from the peak were examined using SDS-PAGE. Red arrow indicates the position of DynU15.

Co-expression and purification of DynE8 + DynU15 revealed that both proteins were present following nickel affinity chromatography, but DynU15 was present in much smaller quantities (Figure 63, inset). SDS-PAGE of the fractions in the peak also showed multiple contaminant bands that were not present in DynE8 purification alone. The peak for the proteins appeared at 40% elution buffer (400 mM imidazole). The fractions containing the two proteins were pooled together and concentrated for gel filtration chromatography.

Figure 64. Gel filtration chromatography results of DynE8 + DynU15 co-expression. The fractions under the three peaks are highlighted in red, dark blue, and purple. Red arrow indicates the position of DynU15.

Gel filtration chromatography of the concentration fractions from nickel affinity chromatography showed a single major peak at 50 mL mark, along with two smaller peaks at the 70- and 85-mL volumes. The SDS-PAGE results show that the major peak contained both DynE8 and DynU15, along with several contaminant. The second and third peaks both contained non-target proteins.

The fractions containing both DynE8 and DynU15 were pooled and concentrated for long-term storage and subsequent experiments.

3.4.4 Negative stain EM data collection of DynE8

The DynE8 sample from gel filtration chromatography was used for negative-stain electron microscopy.

Figure 65. Representative micrograph of purified DynE8. The different shapes are circled in red.

Micrographs of the purified DynE8 showed mostly aggregation and heterogeneous particle sizes between 8 to 25 nm (Figure 65, red circles). This could be due to the high-molecular weight contaminants seen in the SDS-PAGE (Figure 64) or non-optimal grid preparation. Subsequent attempts to repurify DynE8 to reduce the contaminants and aggregation were unsuccessful. Micrographs were not collected from DynE8 negative stain grids.

3.4.5 Sucrose-based gradient centrifugation

In order to increase the uniformity of DynE8 for EM studies, GraFix was employed. In short, GraFix involves the use of cross-linker glutaraldehyde to cross-link DynE8 for stabilization. Cross-linked samples were subjected to sucrose gradient ultracentrifugation and fractionation. A 10-30% sucrose gradient was used for the study. As control, GRAFIX without GA was also performed in parallel.

Figure 66. SDS-PAGE of post-GraFix sucrose density gradient centrifugation fractions. (A) GraFix at 0.2% GA. (B) GraFix at 0.5% GA.

0.2% GA was initially used to cross-link DynE8. However, fractions from the fractionation of the sucrose gradient revealed that there was no cross-linking (Figure 66A). As such, the concentration of GA was increased to 0.5%. When 0.5% GA was used, “new bands” were observed in the SDS-PAGE (Figure 66B). The cross-linked fractions from the 0.5% GA assays were pooled together, and buffer exchanged using the original gel filtration chromatography buffer to remove the sucrose from the protein solution using successive buffer dilution in a 15 kDa MWCO spin concentrator. Removal of sucrose is needed to avoid

complications during sample vitrification as well. The control fractions without GA were also analysed.

Figure 67. SDS-PAGE of GraFix fractions for DynE8 + U15 co-expression. The fractions are numbered in the order that the fractions appear, i.e. fraction 1 came from the bottom of the tube.

Concentrated samples of DynE8 and DynU15 from the co-expression study was used for GraFix as well (Figure 67). However, the proteins were only found in the bottom gradient fractions (Figure 67, fraction 1) and only the bottom 3 fractions contained both DynE8 and U15. Fractions containing non-aggregated DynE8 (Figure 67, fractions 7-9) do not contain any bands near the 40 kDa marker on the protein ladder. This indicated that DynE8 + DynU15 aggregated during ultracentrifugation and migrated quickly to the bottom of the ultracentrifuge tubes. Subsequent experiments yielded the same results and effort was concentrated on DynE8 alone after that.

3.4.6 Negative stain EM data collection of GraFix DynE8

Negative stain grids were prepared using cross-linked DynE8 obtained from GraFix. The DynE8^{GA+} particles appeared similarly sized when visualized using NS-EM.

Figure 68. Representative micrograph of DynE8^{GA+}. Protein particle is circled in red.

Compared to the micrograph containing the non-GraFix DynE8 (Figure 65), the population of particles contained few aggregated particles, with distinct and globular particles. While there were aggregates present, the aggregates were fewer in quantity and the size of aggregates were smaller compared to non-GraFix negative stain grids.

There were two populations of the particles – one population of particles appeared elongated, and another population appeared relatively round. The particles are between 10 – 15 nm (Figure 68, red circle). A total of 102 micrographs of DynE^{GA+} were collected at a defocus value of around -1.2 to -1.5 μm to for further processing.

Negative stain grids were also prepared using non-cross linked DynE8 samples (DynE8^{GA-}) from the GraFix sucrose gradient.

Figure 69. Representative micrograph of DynE8^{GA-}. Protein particle is circled in red.

While the DynE8^{GA-} particles were similar in size to DynE^{GA+} particles, there was a greater variety of conformations observed. There were also similar elongated and globular particles found in the grids (Figure 69). Surprisingly, the sucrose gradient alone was sufficient in improving the purity of the DynE8 samples. A total of 70 DynE8^{GA-} micrographs were collected at a similar defocus value of around -1.2 to -1.5 μm .

3.4.7 Negative stain EM data processing

Using EMAN2 *e2boxer.py* at a box size of 140 pixels, a total of 6,795 DynE8^{GA+} were semi-automatically picked from the grids. As the particles are hypothesized to be heterogenous to some extent, the selected particles were sorted into 100 2D classes to isolate bad particles in their own classes which can be removed from further processing.

Figure 70. 2D classes of DynE8^{GA+} particles. The selected particles were classified and averaged into 100 classes.

The 2D classification revealed the general averaged shapes of DynE8^{GA+}. Similar to the individual particles which were observed in the negative stain EM grids, the two 2D classes have either an elongated or a triangular globular shape (Figure 70). In the globular classes, there is a dark spot in the center, which indicates the possible presence of a cavity. There is presence of large ovoid classes which are different from the elongated shapes.

Figure 71. 2D classes of DynE8^{GA-} particles. The selected particles were classified and averaged into 100 classes.

The same workflow was adopted to process DynE8^{GA-}. A total of 7,040 DynE8^{GA-} particles were picked from the grids. Similarly, the particles were classified into 100 classes using 2D classification (Figure 71). The class averages of DynE8^{GA-} were similar to the class averages of DynE8^{GA+}, with additional smaller round classes that are around half the size of the elongated class averages. On top of that, there are bent elongated or triangular 2D classes that resembled peanuts. A comparison between the two sets of classes reveals that the large ovoid classes from DynE^{GA+} classes are absent in DynE^{GA-}. After 2D classification, the 2D classes from DynE^{GA+} and DynE^{GA-} were then used to construct the respective initial models for 3D refinement and model building.

Since the data was relatively noisy, a second round of 2D classification was conducted using particles that have undergone Wiener filtering, a technique that reduces noise in image. Since the DynE^{GA+} and DynE^{GA-} particles were denoised, the particles were classified into a smaller number of classes.

Figure 72. 2D class average of Wiener filtered particles. (A) DynE^{GA+} particle classes. (B) DynE^{GA-} particle classes.

Overall, the classes generated from Wiener-filtered particles resembled the elongated shapes found in the original classes in Figure 70 – 71 (Figure 72). The classes appear smoother due to removal of noise, and the elongated triangular shapes are more obvious. Several of the classes contain a globular body with a small tip protruding out of the body. The large round classes from DynE8^{GA+} particles remained (Figure 72B). Subsequently, the Wiener-filtered classes were also used to generate 3D initial models.

3.4.8 DynE8 negative stain EM structure

After 2D classification, the 2D classes were used for the generation of eight 3D initial models of DynE8^{GA+} and DynE8^{GA-} using EMAN2 *e2initialmodel.py*.

Figure 73. Representative initial models generated from the 2D classes of DynE8 particles. (A) DynE8^{GA+} initial model. (B) DynE8^{GA-} initial model.

The 3D initial model for DynE8^{GA+} was globular and elongated with a slight cavity space within the model. Similarly, the 3D initial model for DynE8^{GA-} was elongated and globular. The models were generated using relatively noisy data which may not represent DynE8 in reality. Furthermore, upon manual inspection, the views of the initial models do not agree completely with the 2D classes (Figure 70 – 71), as they are globular and spherical instead of being elongated as the 2D classes suggest.

Figure 74. Representative initial models generated from 2D classes of DynE8 Wiener-filtered particles. (A) DynE8^{GA+} initial model. (B) DynE8^{GA-} initial model.

To avoid bias due to noisy data in the 3D initial model during 3D classification and refinement in RELION, the 2D classes using the Wiener filtered particles were also used to generate initial models (Figure 74). The 3D initial models generated were more distinctively elongated, had a relatively more triangular shape, and contained projections that agreed better with its 2D classes. These Wiener filtered samples were chosen for 3D classification in RELION over the original initial models due to the presence of less noise on its surface, which will reduce the probability of a biased model. DynE8 particles were used for 3D refinement in RELION against the DynE8^{GA+} and DynE8^{GA-} initial models.

Figure 75. 3D reconstruction of DynE8 particles. (A) DynE8^{GA+} 3D model. (B) DynE8^{GA-} 3D model. For clarity, the models were filtered using a Gaussian filter to remove noise.

The reconstructed DynE8^{GA+} model measured 167 Å x 124 Å x 87 Å and adopted an asymmetrical triangular shape with a tip protruding out of the body (Figure 75A). The side profile of the model shows a bent volume, which was consistent with the 2D classes from Figure 70 and 72. The reconstructed DynE^{GA-} model measured 162 Å x 115 Å x 100 Å and was similarly asymmetrical (Figure 75B). However, the side profile of the model is more triangular and contains additional density. This suggests that there are slight conformational differences between the two models.

Figure 76. Fourier shell correlation (FSC) graph of DynE8^{GA+} and DynE^{GA-} structures. Density map resolution calculated for the FSC value of approximately 0.143. Global resolution based on FSC for dimer structure is $1/0.033 = \sim 33$ Å

The final resolutions for the negative stain models of DynE^{GA+} and DynE^{GA-} after 3D refinement were both ~ 33 Å (Figure 76) at an FSC of 0.143.

3.4.9 Cryo-EM screening and data collection

To improve the resolution of the DynE8 models, the samples obtained from protein purification experiments were used for cryo-EM sample preparation.

Figure 77. Representative micrograph of the first cryo-EM screening. The particles with dark contrast are ice contaminants.

In the first screening experiment, 1 mg/mL of freshly prepared DynE8 was used with Quantifoil R1.2/1.3 grids without carbon coating. The free-floating DynE8 solution in the holey grids were flash-frozen and immediately used for visualization. The resulting micrographs showed extensive aggregation and ice contamination (Figure 77). DynE8 particles were not discernible. High contrast ice particles were frequently encountered and the interpretation of the image data was challenging.

Figure 78. Representative micrograph of the second cryo-EM screening.

In the second screening experiment, 0.1 mg/mL of DynE8 was also used with Quantifoil R1.2/1.3 grids with carbon coating. The carbon coat allows the particles to bind onto the carbon coat for easier visualization and a higher chance of the particles being present in the grid. The resulting micrographs were an improvement from micrographs from non-carbon coated grids (Figure 78), but the particles were still aggregated.

From the micrographs, it was observed that using fresh DynE8 with cryo-EM resulted in the same problems as negative stain grids, i.e. DynE8 aggregates. As such, to remove aggregates, DynE^{GA-} was used for subsequent cryo-EM experiments. Before sample freezing, the protein was concentrated, and buffer exchanged into its original gel filtration chromatography buffer to remove sucrose.

Figure 79. SDS-PAGE gel of DynE8^{GA}- fractions after GraFix. Red box: fractions used for cryo-EM experiments.

The challenge of using GraFix is low sample yield due to dilution of the same in the fractions after fractionation (Figure 79). During an average GraFix assay, the original concentration of DynE8 (11 mg/mL) was diluted 500-fold after the sucrose fractionation step (~0.02 mg/mL). While the range of the concentration was suitable for negative-stain EM, the concentration was insufficient for cryo-EM purposes which require 1 – 2 mg/mL. As such, several tubes were prepared for each batch of preparation to meet the required amount and concentration.

Figure 80. Representative micrograph of the cryo-EM micrograph with DynE^{GA}.

For the third screening experiment, 0.1 mg/mL of DynE8^{GA-} with Quantifoil R1.2/1.3 without carbon coating was used for cryo-EM sample preparation. Using DynE8^{GA-} improved the sample purity and resulted in relatively better visualization outcome because the micrograph contained fewer aggregation and the particles were more distinct (Figure 80). There was fewer ice particles present as well in the micrographs as well. However, the micrograph contained a much smaller number of particles per micrograph compared to using fresh DynE8. This required a higher concentration of DynE8^{GA-}. As such, more tubes were used per round of ultracentrifugation, i.e. six tubes per batch instead of one tube, and more sample was loaded in each centrifuge tube to increase the overall yield.

Figure 81. Representative micrograph of cryo-EM micrograph with DynE^{GA+}

Additionally, 0.1 mg/mL of DynE^{GA+} with Quantifoil R1.2/1.3 without carbon coating was used for cryo-EM sample preparation. GraFix-treated DynE8 appeared to be aggregated and had poor sample purity, presumably due to the cross-linking action of glutaraldehyde (Figure 81). Subsequent efforts were focused on DynE8^{GA-} instead.

Using the concentrated samples of DynE8^{GA-}, two rounds of cryo-EM data collection were conducted. In the first round, DynE8^{GA-} at 2.0 mg/mL was used in the data collection with using Quantifoil R2/2 grids without carbon coating. 475 micrographs were collected. In the second round, DynE8^{GA-} at 2.7 mg/mL prepared using the same cryo-EM grids was used and 919 micrographs were collected.

Figure 82. Representative micrograph of the cryo-EM micrograph with higher concentrations of DynE^{GA-}. Red circle: individual DynE8 particles.

Using higher concentrations of DynE8^{GA-} resulted in more particles per micrograph (Figure 82). The particles were identifiable from the micrographs. However, the micrographs were noisy due to small ice particles and contained unspecific black spots.

3.4.10 Cryo-EM data processing

EMAN2 *e2boxer.py* was used to manually select 3,518 particles from the first batch of 475 micrographs. The individual particles selected from the micrographs appeared globular and round (Figure 83). Apart from the globular particles, the particles in the micrographs appeared to have an elongated and had a triangular shape which were similar to the negative stain classes in Figure 72.

Figure 83. Individual particles picked manually from micrographs.

The manually selected particles were then classified into 15 classes using EMAN2 for an average of 200 particles per class average (Figure 84).

Figure 84. 2D classes of particles using EMAN2.

The 2D classes appeared to compose of two distinct shapes, the first shape is an elongated bent oval, similar to a peanut and the second shape is round. There are two different sizes of the round 2D class averages. The particles were then class-averaged in RELION to create the template for automatic particle picking. Since there were multiple 2D classes which were similar, the particles were classified into a smaller number of 2D classes.

Figure 85. 2D classes of cryo-EM^{GA} particles using RELION.

The 2D classes generated using RELION agreed with the classes from EMAN2 (Figure 85).

The classes generated using RELION contained both the elongated and circular shapes.

Using RELION's template-based particle auto-picking program, 142,317 particles were picked. The particles were classified into 500 2D classes to identify bad particles.

Figure 86. An example of the DynE8^{GA}- 2D classes from automated particle picking.

The classes generated were composed mostly of circular shapes that were significantly larger than 160 Å and were composed of mostly ice particles (Figure 86). On top of that, there were shapes that extended beyond the circular mask, and shapes that filled up the entire circular mask. Manual inspection of the particles that belong to those classes indicated that ice of various shapes and sizes were also automatically picked during the template-based particle picking process.

Classes containing shapes in Figure 80 were selected and the bad classes were discarded from the dataset. To further remove bad particles and ice, two more rounds of 2D classification were subsequently performed, and final total of 10,546 particles which represented 7.4% of the original dataset were selected for a final round of 2D classification.

Figure 87. 2D classes after three rounds of particle selection.

However, subsequent examination of the particles in the 2D classes revealed that the classes were still composed mostly of high-contrast ice particles with a relatively small proportion of DynE8^{GA-} (Figure 87). The largest particles were larger than expected at ~30 nm instead of the measured ~16 – 17 nm on its longest dimension. Furthermore, several 2D classes were completely circular which is due to high levels of noise and a lack features to average.

Since the first dataset contained extensive ice contaminants and rendered proper analysis challenging, a second round of data collection was performed. In the second round of data collection, R2/2 grids were used instead and 2.7 mg/mL of DynE8^{GA-} was used. 919 micrographs were collected in total. 1,461 particles were collected manually and classified into 8 classes for examination. Care was taken to avoid picking ice particles.

Figure 88. 2D classification of the second round of DynE8^{GA-} dataset using EMAN2.

2D classification of the second round of particles using EMAN2 show that the DynE8 particles are globular, elongated, and peanut-shaped (Figure 88). The sizes of the particles are more consistent, i.e. 10 – 17 nm range as opposed to the first round of data. Using the 2D classes, automatic picking was performed for a total of 475,654 particles. However, most of the dataset still contained noise and ice particles and three rounds of 2D classification were performed to select the best classes. This resulted in a final number of 52,195 particles, which represented 10.9% of the original dataset.

Figure 89. An example of the 2D classes for DynE^{GA-}.

The dataset was then subjected to a final round of 2D classification (Figure 89). The 2D classes maintained the same elongated peanut shape, triangle, and round shapes. However, on top of the 2D classes in the center, most of the classes contain additional density at the periphery of the masks. Despite best efforts to remove the peripheral noise through changing the mask sizes, the 2D classes still contained the additional signals at the periphery of the 2D mask which indicate there was significant overlap between the DynE8 particles. Current efforts are underway to improve the removal of overlapping particles and reducing the peripheral noise.

3.5 Discussion

As a highly-reducing iPKS that produces enediyne, the domains in DynE8 are arranged differently from the partially-reducing iPKS and non-reducing iPKS, where the sequence encoding the ACP domain is located in the middle of the protein sequence. We hypothesized that DynE8 may adopt a conformation that is different from NcsB and other PKS family proteins, e.g., PikAIII, FAS, and MAS.

While there is interest in elucidating the atomic structure of enediyne-producing iPKS, prior studies involving cryo-EM sample preparation for polyketide synthase were unsuccessful due to sample heterogeneity and aggregation. With a modified sucrose gradient ultracentrifugation step, we have – for the first time – obtained monodispersed particle preparation for DynE8 for negative stain EM. 2D classes of DynE^{GA+} contain unusually large classes, which may indicate the presence of non-native conformations caused by the usage of glutaraldehyde.

Using the improved DynE8 protein preparation method and negative stain EM, we obtained low-resolution structures of DynE8. Due to the low resolutions of the DynE8 models of around 30 Å, the models are unsuitable for model segmentation for homology model placements. As such, we can only observe the general shapes of the DynE8 model. Due to the closed conformation of the EM model, it is possible that DynE8 adopt a closed structure to form a catalytic cavity.

Figure 90. Comparison between DynE8^{GA-}, PikAIII, and NcsB. Red line indicates line of symmetry of the three models.

In a comparison between DynE8^{GA-}, PikAIII, and NcsB, an axis of symmetry is proposed for DynE8^{GA-} (Figure 90). DynE8^{GA-} is triangular and resembles the linear conformation of PikAIII compared to the dumbbell-shaped and compact NcsB. However, the bilateral symmetry is not as obvious in DynE8 compared to PikAIII, and may be due to the presence of flexible domains. NcsB adopts a compact conformation because the location of ACP – the domain that shuttles between KS and AT – is located at the end of the protein. As such, there is a great distance between ACP and the two domains if NcsB adopts a linear conformation and the curled up compact conformation reduces that distance. However, the ACP domain in

DynE8 is located at the center of the protein, immediately after the KS and AT domains. As such, it is possible for DynE8 to adopt a linear conformation that resembles PikAIII more than NcsB since the ACP domain can shuttle products easily to other domains.

Sucrose gradient-based ultracentrifugation was also during sample preparation for cryo-EM. This resulted in a sample improvement for both EM studies of DynE8. However, due to the relatively small size of the protein and presence of ice in the micrographs, template-based auto-picking resulted in the selection of ice particles as well. This resulted in class averages that masked the actual signals from DynE8 particles.

Elucidation of putative acetylenase DynU15 would pave a way into understanding the structure-function relationship in proteins that can form triple carbon bonds. The aim of the study also involved purifying DynU15 for crystallization studies and binding it to DynE8 as a strategy to increase total protein size for cryo-EM experiments. However, there were several challenges in DynU15 purification. Firstly, despite the small size of the protein, only small amounts of DynU15 were purified and the amounts were insufficient for crystallization studies. Secondly, DynU15 aggregated with DynE8 under sucrose ultracentrifugation, which posed a challenge.

Future work will involve improving sample preparation for cryo-EM studies to improve yield and to reduce ice artefacts. More work will be conducted to explore sucrose gradient-based ultracentrifugation as a viable method for iPKS sample preparation for cryo-EM studies. In addition, expression and sample preparation for DynU15 will be improved to obtain samples that do not aggregate under sucrose gradient ultracentrifugation.

Chapter 4: Summary

4.1 Conclusion

The first part of the thesis involved a quantitative study into the relationship between a PKS ketosynthase domain physicochemical properties and the product length. By reviewing and manually curating published iPKS data, regressions and machine learning models with predictive properties were obtained.

The second part of the thesis elucidated the overall architecture and revealed a low-resolution structure of partially-reducing iterative polyketide synthase NcsB using a combination of experimental techniques. A novel elongated architecture, distinct from its homologues, was discovered.

The final part of the thesis attempted to characterize highly-reducing iPKS DynE8 and its polyketide acetylase DynU15. While attempts at characterizing DynU15 were unsuccessful, the use of negative stain and cryo-EM was successfully deployed to uncover the low-resolution mode of DynE8. The structure of DynE8 suggests a possible triangular and closed conformation that is different from the current structure of highly-polyketide iPKS MAS-like PKS.

The work contained in this PhD candidature is a step closer to understanding the structure-function relation in iterative polyketide synthases. Taken together, this body of work furthers our understanding, with the predictive potential of KS domain for product length and the low-resolution architecture of NcsB and DynE8.

4.2 Future work and directions

Polyketide synthase research is increasing in quantity and popularity in the past decade. More studies have also been conducted on understanding the roles of individual domains through heterologous expression studies, domain swapping, and domain engineering. The increasing amount of literature will expand the dataset available for the machine learning model, which will improve its performance and accuracy.

Electron microscopy has proved a promising tool for researchers in elucidating polyketide structures. Our study demonstrated that it is possible to visualize particles on the cryo-EM grid, and optimization of the freezing and sample preparation process will reduce ice contamination and improve signal-to-noise ratio for reconstruction of high-resolution structure.

While there are still many challenges in determining high-resolution structures of iterative polyketide synthases, the unyielding spirit of researchers in PKS groups will one day uncover their mysteries and pave way to a new era of tuneable and rationally designable polyketide synthases to create new therapeutic compounds for the future.

References

- 1 Hopwood, D. A. Cracking the polyketide code. *PLoS Biol* **2**, E35, doi:10.1371/journal.pbio.0020035 (2004).
- 2 Keatinge-Clay, A. T. The structures of type I polyketide synthases. *Nat Prod Rep* **29**, 1050-1073, doi:10.1039/c2np20019h (2012).
- 3 Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S. & Melancon, C. E., 3rd. Expanding our understanding of sequence-function relationships of type II polyketide biosynthetic gene clusters: bioinformatics-guided identification of Frankiamicin A from *Frankia* sp. EAN1pec. *PLoS One* **10**, e0121505, doi:10.1371/journal.pone.0121505 (2015).
- 4 Shimizu, Y., Ogata, H. & Goto, S. Type III Polyketide Synthases: Functional Classification and Phylogenomics. *Chembiochem* **18**, 50-65, doi:10.1002/cbic.201600522 (2017).
- 5 Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res* **45**, W49-W54, doi:10.1093/nar/gkx320 (2017).
- 6 Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat Chem Biol* **15**, 813-821, doi:10.1038/s41589-019-0313-7 (2019).
- 7 Ziemert, N. *et al.* The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **7**, e34064, doi:10.1371/journal.pone.0034064 (2012).
- 8 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).

- 9 Shi-Kunne, X. *et al.* In silico prediction and characterisation of secondary metabolite clusters in the plant pathogenic fungus *Verticillium dahliae*. *FEMS Microbiol Lett* **366**, doi:10.1093/femsle/fnz081 (2019).
- 10 Szu, P. H. *et al.* Analysis of the ketosynthase-chain length factor heterodimer from the fredericamycin polyketide synthase. *Chem Biol* **18**, 1021-1031, doi:10.1016/j.chembiol.2011.07.015 (2011).
- 11 Herbst, D. A. *et al.* The structural organization of substrate loading in iterative polyketide synthases. *Nat Chem Biol* **14**, 474-479, doi:10.1038/s41589-018-0026-3 (2018).
- 12 Yang, X. L. *et al.* Molecular basis of methylation and chain-length programming in a fungal iterative highly reducing polyketide synthase. *Chem Sci* **10**, 8478-8489, doi:10.1039/c9sc03173a (2019).
- 13 Liu, L., Zhang, Z., Shao, C. L. & Wang, C. Y. Analysis of the Sequences, Structures, and Functions of Product-Releasing Enzyme Domains in Fungal Polyketide Synthases. *Front Microbiol* **8**, 1685, doi:10.3389/fmicb.2017.01685 (2017).
- 14 Vagstad, A. L. *et al.* Combinatorial Domain Swaps Provide Insights into the Rules of Fungal Polyketide Synthase Programming and the Rational Synthesis of Non-Native Aromatic Products. *Angewandte Chemie International Edition* **52**, 1718-1721, doi:10.1002/anie.201208550 (2013).
- 15 Liu, T., Sanchez, J. F., Chiang, Y.-M., Oakley, B. R. & Wang, C. C. C. Rational Domain Swaps Reveal Insights about Chain Length Control by Ketosynthase Domains in Fungal Nonreducing Polyketide Synthases. *Organic Letters* **16**, 1676-1679, doi:10.1021/ol5003384 (2014).
- 16 Liu, T., Chiang, Y.-M., Somoza, A. D., Oakley, B. R. & Wang, C. C. C. Engineering of an “Unnatural” Natural Product by Swapping Polyketide Synthase Domains in

- Aspergillus nidulans*. *Journal of the American Chemical Society* **133**, 13314-13316, doi:10.1021/ja205780g (2011).
- 17 Vagstad, A. L., Bumpus, S. B., Belecki, K., Kelleher, N. L. & Townsend, C. A. Interrogation of Global Active Site Occupancy of a Fungal Iterative Polyketide Synthase Reveals Strategies for Maintaining Biosynthetic Fidelity. *Journal of the American Chemical Society* **134**, 6865-6877, doi:10.1021/ja3016389 (2012).
 - 18 Yadav, G., Gokhale, R. S. & Mohanty, D. Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput Biol* **5**, e1000351, doi:10.1371/journal.pcbi.1000351 (2009).
 - 19 Bachmann, B. O. & Ravel, J. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* **458**, 181-217, doi:10.1016/S0076-6879(09)04808-3 (2009).
 - 20 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).
 - 21 Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res* **46**, W363-W367, doi:10.1093/nar/gky473 (2018).
 - 22 Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132, doi:10.1016/0022-2836(82)90515-0 (1982).
 - 23 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
 - 24 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

- 25 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794 (ACM, San Francisco, California, USA, 2016).
- 26 Walt, S. v. d., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13**, 22-30, doi:10.1109/MCSE.2011.37 (2011).
- 27 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90-95, doi:10.1109/MCSE.2007.55 (2007).
- 28 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 6639-6649 (Curran Associates Inc., Montré#233;al, Canada, 2018).
- 29 Ke, G. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149-3157 (Curran Associates Inc., Long Beach, California, USA, 2017).
- 30 Zhang, A. *et al.* Efficient disruption of a polyketide synthase gene (pks1) required for melanin synthesis through Agrobacterium-mediated transformation of *Glarea lozoyensis*. *Mol Genet Genomics* **268**, 645-655, doi:10.1007/s00438-002-0780-4 (2003).
- 31 Ma, S. M. *et al.* Enzymatic Synthesis of Aromatic Polyketides Using PKS4 from *Gibberella fujikuroi*. *Journal of the American Chemical Society* **129**, 10642-10643, doi:10.1021/ja074865p (2007).
- 32 Fujii, I. *et al.* Enzymatic synthesis of 1,3,6,8-tetrahydroxynaphthalene solely from malonyl coenzyme A by a fungal iterative type I polyketide synthase PKS1. *Biochemistry* **39**, 8853-8858, doi:10.1021/bi000644j (2000).

- 33 Sanchez, J. F. *et al.* Molecular genetic analysis of the orsellinic acid/F9775 gene cluster of *Aspergillus nidulans*. *Molecular BioSystems* **6**, 587-593, doi:10.1039/B904541D (2010).
- 34 Wang, S. *et al.* Functional characterization of the biosynthesis of radicicol, an Hsp90 inhibitor resorcylic acid lactone from *Chaetomium chiversii*. *Chem Biol* **15**, 1328-1338, doi:10.1016/j.chembiol.2008.10.006 (2008).
- 35 Zhou, H., Zhan, J., Watanabe, K., Xie, X. & Tang, Y. A polyketide macrolactone synthase from the filamentous fungus *Gibberella zeae*. *Proc Natl Acad Sci U S A* **105**, 6249-6254, doi:10.1073/pnas.0800657105 (2008).
- 36 Zhou, H. *et al.* Enzymatic Synthesis of Resorcylic Acid Lactones by Cooperation of Fungal Iterative Polyketide Synthases Involved in Hypothemycin Biosynthesis. *Journal of the American Chemical Society* **132**, 4530-4531, doi:10.1021/ja100060k (2010).
- 37 Zhou, H., Qiao, K., Gao, Z., Vederas, J. C. & Tang, Y. Insights into Radicicol Biosynthesis via Heterologous Synthesis of Intermediates and Analogs. *Journal of Biological Chemistry* **285**, 41412-41421 (2010).
- 38 Xu, Y. *et al.* Insights into the Biosynthesis of 12-Membered Resorcylic Acid Lactones from Heterologous Production in *Saccharomyces cerevisiae*. *ACS Chemical Biology* **9**, 1119-1127, doi:10.1021/cb500043g (2014).
- 39 Moriwaki, A. *et al.* Insertional mutagenesis and characterization of a polyketide synthase gene (PKS1) required for melanin biosynthesis in *Bipolaris oryzae*. *FEMS Microbiol Lett* **238**, 1-8, doi:10.1016/j.femsle.2004.07.007 (2004).
- 40 Zhang, P. *et al.* A cryptic pigment biosynthetic pathway uncovered by heterologous expression is essential for conidial development in *Pestalotiopsis fici*. *Molecular Microbiology* **105**, 469-483, doi:10.1111/mmi.13711 (2017).

- 41 Tsai, H. F., Chang, Y. C., Washburn, R. G., Wheeler, M. H. & Kwon-Chung, K. J. The developmentally regulated alb1 gene of *Aspergillus fumigatus*: its role in modulation of conidial morphology and virulence. *J Bacteriol* **180**, 3031-3038 (1998).
- 42 Fujii, I., Watanabe, A., Sankawa, U. & Ebizuka, Y. Identification of Claisen cyclase domain in fungal polyketide synthase WA, a naphthopyrone synthase of *Aspergillus nidulans*. *Chem Biol* **8**, 189-197, doi:10.1016/s1074-5521(00)90068-1 (2001).
- 43 Chiang, Y. M. *et al.* Characterization of a polyketide synthase in *Aspergillus niger* whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-gamma-pyrone. *Fungal Genet Biol* **48**, 430-437, doi:10.1016/j.fgb.2010.12.001 (2011).
- 44 Frandsen, R. J. N. *et al.* The biosynthetic pathway for aurofusarin in *Fusarium graminearum* reveals a close link between the naphthoquinones and naphthopyrones. *Molecular Microbiology* **61**, 1069-1080, doi:10.1111/j.1365-2958.2006.05295.x (2006).
- 45 Cary, J. W., Ehrlich, K. C., Beltz, S. B., Harris-Coward, P. & Klich, M. A. Characterization of the *Aspergillus ochraceoroseus* aflatoxin/sterigmatocystin biosynthetic gene cluster. *Mycologia* **101**, 352-362, doi:10.3852/08-173 (2009).
- 46 Newman, A. G., Vagstad, A. L., Belecki, K., Scheerer, J. R. & Townsend, C. A. Analysis of the cercosporin polyketide synthase CTB1 reveals a new fungal thioesterase function. *Chemical Communications* **48**, 11772-11774, doi:10.1039/C2CC36010A (2012).
- 47 Studt, L., Wiemann, P., Kleigrew, K., Humpf, H. U. & Tudzynski, B. Biosynthesis of fusarubins accounts for pigmentation of *Fusarium fujikuroi* perithecia. *Appl Environ Microbiol* **78**, 4468-4480, doi:10.1128/AEM.00823-12 (2012).

- 48 Chettri, P. *et al.* Dothistromin genes at multiple separate loci are regulated by AfIR. *Fungal Genetics and Biology* **51**, 12-20, doi:<https://doi.org/10.1016/j.fgb.2012.11.006> (2013).
- 49 Awakawa, T., Kaji, T., Wakimoto, T. & Abe, I. A heptaketide naphthaldehyde produced by a polyketide synthase from *Nectria haematococca*. *Bioorg Med Chem Lett* **22**, 4338-4340, doi:10.1016/j.bmcl.2012.05.005 (2012).
- 50 Chooi, Y. H., Cacho, R. & Tang, Y. Identification of the viridicatumtoxin and griseofulvin gene clusters from *Penicillium aethiopicum*. *Chem Biol* **17**, 483-494, doi:10.1016/j.chembiol.2010.03.015 (2010).
- 51 Lim, F. Y. *et al.* Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in *Aspergillus fumigatus*. *Appl Environ Microbiol* **78**, 4117-4125, doi:10.1128/AEM.07710-11 (2012).
- 52 Chiang, Y. M. *et al.* Characterization of the *Aspergillus nidulans* monodictyphenone gene cluster. *Appl Environ Microbiol* **76**, 2067-2074, doi:10.1128/AEM.02187-09 (2010).
- 53 Li, Y., Chooi, Y.-H., Sheng, Y., Valentine, J. S. & Tang, Y. Comparative Characterization of Fungal Anthracenone and Naphthacenedione Biosynthetic Pathways Reveals an α -Hydroxylation-Dependent Claisen-like Cyclization Catalyzed by a Dimanganese Thioesterase. *Journal of the American Chemical Society* **133**, 15773-15785, doi:10.1021/ja206906d (2011).
- 54 Awakawa, T. *et al.* Physically discrete beta-lactamase-type thioesterase catalyzes product release in atrochrysone synthesis by iterative type I polyketide synthase. *Chem Biol* **16**, 613-623, doi:10.1016/j.chembiol.2009.04.004 (2009).
- 55 Klejnstrup, M. L. *et al.* Genetics of Polyketide Metabolism in *Aspergillus nidulans*. *Metabolites* **2**, 100-133, doi:10.3390/metabo2010100 (2012).

- 56 Regueira, T. B. *et al.* Molecular basis for mycophenolic acid biosynthesis in *Penicillium brevicompactum*. *Appl Environ Microbiol* **77**, 3035-3043, doi:10.1128/AEM.03015-10 (2011).
- 57 Ahuja, M. *et al.* Illuminating the diversity of aromatic polyketide synthases in *Aspergillus nidulans*. *J Am Chem Soc* **134**, 8212-8221, doi:10.1021/ja3016395 (2012).
- 58 Chiang, Y.-M. *et al.* A Gene Cluster Containing Two Fungal Polyketide Synthases Encodes the Biosynthetic Pathway for a Polyketide, Asperfuranone, in *Aspergillus nidulans*. *Journal of the American Chemical Society* **131**, 2965-2970, doi:10.1021/ja8088185 (2009).
- 59 Zabala, A. O., Xu, W., Chooi, Y. H. & Tang, Y. Characterization of a silent azaphilone gene cluster from *Aspergillus niger* ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. *Chem Biol* **19**, 1049-1059, doi:10.1016/j.chembiol.2012.07.004 (2012).
- 60 Storm, P. A., Herbst, D. A., Maier, T. & Townsend, C. A. Functional and Structural Analysis of Programmed C-Methylation in the Biosynthesis of the Fungal Polyketide Citrinin. *Cell Chem Biol* **24**, 316-325, doi:10.1016/j.chembiol.2017.01.008 (2017).
- 61 Bailey, A. M. *et al.* Characterisation of 3-methylorcinaldehyde synthase (MOS) in *Acremonium strictum*: first observation of a reductive release mechanism during polyketide biosynthesis. *Chemical Communications*, 4053-4055, doi:10.1039/B708614H (2007).
- 62 Lackner, G., Bohnert, M., Wick, J. & Hoffmeister, D. Assembly of melleolide antibiotics involves a polyketide synthase with cross-coupling activity. *Chem Biol* **20**, 1101-1106, doi:10.1016/j.chembiol.2013.07.009 (2013).

- 63 Winter, J. M. *et al.* Identification and Characterization of the Chaetoviridin and Chaetomugilin Gene Cluster in *Chaetomium globosum* Reveal Dual Functions of an Iterative Highly-Reducing Polyketide Synthase. *Journal of the American Chemical Society* **134**, 17900-17903, doi:10.1021/ja3090498 (2012).
- 64 Fahad, A. a. *et al.* Oxidative dearomatisation: the key step of sorbicillinoid biosynthesis. *Chemical Science* **5**, 523-527, doi:10.1039/C3SC52911H (2014).
- 65 Chooi, Y. H., Muria-Gonzalez, M. J., Mead, O. L. & Solomon, P. S. SnPKS19 Encodes the Polyketide Synthase for Alternariol Mycotoxin Biosynthesis in the Wheat Pathogen *Parastagonospora nodorum*. *Appl Environ Microbiol* **81**, 5309-5317, doi:10.1128/AEM.00278-15 (2015).
- 66 Chooi, Y. H. *et al.* Cloning and sequence characterization of a non-reducing polyketide synthase gene from the lichen *Xanthoparmelia semiviridis*. *Mycol Res* **112**, 147-161, doi:10.1016/j.mycres.2007.08.022 (2008).
- 67 Yu, P. W., Chang, Y. C., Liou, R. F., Lee, T. H. & Tzean, S. S. pks63787, a Polyketide Synthase Gene Responsible for the Biosynthesis of Benzenoids in the Medicinal Mushroom *Antrodia cinnamomea*. *J Nat Prod* **79**, 1485-1491, doi:10.1021/acs.jnatprod.5b00798 (2016).
- 68 Abdel-Hameed, M., Bertrand, R. L., Piercey-Normore, M. D. & Sorensen, J. L. Identification of 6-Hydroxymellein Synthase and Accessory Genes in the Lichen *Cladonia uncialis*. *Journal of Natural Products* **79**, 1645-1650, doi:10.1021/acs.jnatprod.6b00257 (2016).
- 69 Kashiwa, N., Ebizuka, Y. & Fujii, I. Identification and Heterologous Expression of the Topopyrone Nonaketide Synthase Gene from *Phoma* sp. *Chem Pharm Bull (Tokyo)* **64**, 947-951, doi:10.1248/cpb.c16-00172 (2016).

- 70 Itoh, T. *et al.* Identification of a Key Prenyltransferase Involved in Biosynthesis of the Most Abundant Fungal Meroterpenoids Derived from 3,5-Dimethylorsellinic Acid. *ChemBioChem* **13**, 1132-1135, doi:10.1002/cbic.201200124 (2012).
- 71 Griffiths, S. A. *et al.* Assignment of a dubious gene cluster to melanin biosynthesis in the tomato fungal pathogen *Cladosporium fulvum*. *PLoS One* **13**, e0209600, doi:10.1371/journal.pone.0209600 (2018).
- 72 Artigot, M. P. *et al.* Molecular cloning and functional characterization of two CYP619 cytochrome P450s involved in biosynthesis of patulin in *Aspergillus clavatus*. *Microbiology* **155**, 1738-1747, doi:<https://doi.org/10.1099/mic.0.024836-0> (2009).
- 73 Liu, W. *et al.* The neocarzinostatin biosynthetic gene cluster from *Streptomyces carzinostaticus* ATCC 15944 involving two iterative type I polyketide synthases. *Chem Biol* **12**, 293-302, doi:10.1016/j.chembiol.2004.12.013 (2005).
- 74 Sun, H. *et al.* Synthesis of (R)-mellein by a partially reducing iterative polyketide synthase. *J Am Chem Soc* **134**, 11924-11927, doi:10.1021/ja304905e (2012).
- 75 Zhao, Q. *et al.* Characterization of the azinomycin B biosynthetic gene cluster revealing a different iterative type I polyketide synthase for naphthoate biosynthesis. *Chem Biol* **15**, 693-705, doi:10.1016/j.chembiol.2008.05.021 (2008).
- 76 Xiao, Y. *et al.* Characterization of tiacumicin B biosynthetic gene cluster affording diversified tiacumicin analogues and revealing a tailoring dihalogenase. *J Am Chem Soc* **133**, 1092-1105, doi:10.1021/ja109445q (2011).
- 77 Ahlert, J. *et al.* The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science* **297**, 1173-1176, doi:10.1126/science.1072105 (2002).

- 78 Ito, T. *et al.* Deciphering pactamycin biosynthesis and engineered production of new pactamycin analogues. *Chembiochem* **10**, 2253-2265, doi:10.1002/cbic.200900339 (2009).
- 79 Van Lanen, S. G., Oh, T. J., Liu, W., Wendt-Pienkowski, E. & Shen, B. Characterization of the maduropeptin biosynthetic gene cluster from *Actinomadura madurae* ATCC 39144 supporting a unifying paradigm for enediyne biosynthesis. *J Am Chem Soc* **129**, 13082-13094, doi:10.1021/ja073275o (2007).
- 80 Daum, M. *et al.* Organisation of the biosynthetic gene cluster and tailoring enzymes in the biosynthesis of the tetracyclic quinone glycoside antibiotic polyketomycin. *Chembiochem* **10**, 1073-1083, doi:10.1002/cbic.200800823 (2009).
- 81 Shao, L. *et al.* Cloning and characterization of a bacterial iterative type I polyketide synthase gene encoding the 6-methylsalicylic acid synthase. *Biochem Biophys Res Commun* **345**, 133-139, doi:10.1016/j.bbrc.2006.04.069 (2006).
- 82 Chooi, Y. H. *et al.* An in planta-expressed polyketide synthase produces (R)-mellein in the wheat pathogen *Parastagonospora nodorum*. *Appl Environ Microbiol* **81**, 177-186, doi:10.1128/AEM.02745-14 (2015).
- 83 Kage, H., Kreutzer, M. F., Wackler, B., Hoffmeister, D. & Nett, M. An iterative type I polyketide synthase initiates the biosynthesis of the antimycoplasma agent micacocidin. *Chem Biol* **20**, 764-771, doi:10.1016/j.chembiol.2013.04.010 (2013).
- 84 Tannous, J. *et al.* Sequencing, physical organization and kinetic expression of the patulin biosynthetic gene cluster from *Penicillium expansum*. *Int J Food Microbiol* **189**, 51-60, doi:10.1016/j.ijfoodmicro.2014.07.028 (2014).
- 85 Spencer, J. B. & Jordan, P. M. Purification and properties of 6-methylsalicylic acid synthase from *Penicillium patulum*. *Biochem J* **288** (Pt 3), 839-846, doi:10.1042/bj2880839 (1992).

- 86 Kim, J. A., Hong, S. G., Cheong, Y. H., Koh, Y. J. & Hur, J. S. A new reducing polyketide synthase gene from the lichen-forming fungus *Cladonia metacorallifera*. *Mycologia* **104**, 362-370, doi:10.3852/11-001 (2012).
- 87 Gaisser, S., Trefzer, A., Stockert, S., Kirschning, A. & Bechthold, A. Cloning of an avilamycin biosynthetic gene cluster from *Streptomyces viridochromogenes* Tu57. *J Bacteriol* **179**, 6271-6278, doi:10.1128/jb.179.20.6271-6278.1997 (1997).
- 88 Karlsson, M. *et al.* Insights on the evolution of mycoparasitism from the genome of *Clonostachys rosea*. *Genome Biol Evol* **7**, 465-480, doi:10.1093/gbe/evu292 (2015).
- 89 Lohman, J. R. *et al.* Cloning and sequencing of the kedarcidin biosynthetic gene cluster from *Streptoalloteichus* sp. ATCC 53650 revealing new insights into biosynthesis of the enediyne family of antitumor antibiotics. *Mol Biosyst* **9**, 478-491, doi:10.1039/c3mb25523a (2013).
- 90 Fujii, I. *et al.* Cloning of the polyketide synthase gene atX from *Aspergillus terreus* and its identification as the 6-methylsalicylic acid synthase gene by heterologous expression. *Mol Gen Genet* **253**, 1-10, doi:10.1007/s004380050289 (1996).
- 91 Bacha, N. *et al.* Cloning and characterization of novel methylsalicylic acid synthase gene involved in the biosynthesis of isoasperlactone and asperlactone in *Aspergillus westerdijikiae*. *Fungal Genet Biol* **46**, 742-749, doi:10.1016/j.fgb.2009.07.002 (2009).
- 92 Gao, Q. & Thorson, J. S. The biosynthetic genes encoding for the production of the dynemicin enediyne core in *Micromonospora chersina* ATCC53710. *FEMS Microbiol Lett* **282**, 105-114, doi:10.1111/j.1574-6968.2008.01112.x (2008).
- 93 Sun, H. *et al.* Products of the iterative polyketide synthases in 9- and 10-membered enediyne biosynthesis. *Chemical Communications*, 7399-7401, doi:10.1039/B916751J (2009).

- 94 Zabala, A. O., Chooi, Y. H., Choi, M. S., Lin, H. C. & Tang, Y. Fungal polyketide synthase product chain-length control by partnering thiohydrolase. *ACS Chem Biol* **9**, 1576-1586, doi:10.1021/cb500284t (2014).
- 95 Butchko, R. A., Plattner, R. D. & Proctor, R. H. FUM13 encodes a short chain dehydrogenase/reductase required for C-3 carbonyl reduction during fumonisin biosynthesis in *Gibberella moniliformis*. *J Agric Food Chem* **51**, 3000-3006, doi:10.1021/jf0262007 (2003).
- 96 Xie, X., Meehan, M. J., Xu, W., Dorrestein, P. C. & Tang, Y. Acyltransferase mediated polyketide release from a fungal megasynthase. *J Am Chem Soc* **131**, 8388-8389, doi:10.1021/ja903203g (2009).
- 97 Ma, S. M. & Tang, Y. Biochemical characterization of the minimal polyketide synthase domains in the lovastatin nonaketide synthase LovB. *FEBS J* **274**, 2854-2864, doi:10.1111/j.1742-4658.2007.05818.x (2007).
- 98 Sakai, K., Kinoshita, H. & Nihira, T. Identification of mokB involved in monacolin K biosynthesis in *Monascus pilosus*. *Biotechnol Lett* **31**, 1911-1916, doi:10.1007/s10529-009-0093-3 (2009).
- 99 Ugai, T. *et al.* Heterologous expression of highly reducing polyketide synthase involved in betaenone biosynthesis. *Chemical Communications* **51**, 1878-1881, doi:10.1039/C4CC09512J (2015).
- 100 Chen, L. *et al.* Engineering of New Pneumocandin Side-Chain Analogues from *Glarea lozoyensis* by Mutasynthesis and Evaluation of Their Antifungal Activity. *ACS Chemical Biology* **11**, 2724-2733, doi:10.1021/acschembio.6b00604 (2016).
- 101 Liu, W. *et al.* Rapid PCR amplification of minimal enediyne polyketide synthase cassettes leads to a predictive familial classification model. *Proceedings of the National Academy of Sciences* **100**, 11959, doi:10.1073/pnas.2034291100 (2003).

- 102 Kasahara, K. *et al.* Solanapyrone Synthase, a Possible Diels–Alderase and Iterative Type I Polyketide Synthase Encoded in a Biosynthetic Gene Cluster from *Alternaria solani*. *ChemBioChem* **11**, 1245–1252, doi:10.1002/cbic.201000173 (2010).
- 103 Fujii, I., Yoshida, N., Shimomaki, S., Oikawa, H. & Ebizuka, Y. An Iterative Type I Polyketide Synthase PKSN Catalyzes Synthesis of the Decaketide Alternapyrone with Regio-Specific Octa-Methylation. *Chemistry & Biology* **12**, 1301–1309, doi:<https://doi.org/10.1016/j.chembiol.2005.09.015> (2005).
- 104 Mao, X.-M. *et al.* Efficient Biosynthesis of Fungal Polyketides Containing the Dioxabicyclo-octane Ring System. *Journal of the American Chemical Society* **137**, 11904–11907, doi:10.1021/jacs.5b07816 (2015).
- 105 Kasahara, K., Fujii, I., Oikawa, H. & Ebizuka, Y. Expression of *Alternaria solani* PKS_F generates a set of complex reduced-type polyketides with different carbon-lengths and cyclization. *Chembiochem* **7**, 920–924, doi:10.1002/cbic.200600034 (2006).
- 106 Saha, D. *et al.* Identification of a polyketide synthase required for alternariol (AOH) and alternariol-9-methyl ether (AME) formation in *Alternaria alternata*. *PLoS One* **7**, e40564, doi:10.1371/journal.pone.0040564 (2012).
- 107 Janevska, S. *et al.* Gibepyrone Biosynthesis in the Rice Pathogen *Fusarium fujikuroi* Is Facilitated by a Small Polyketide Synthase Gene Cluster. *J Biol Chem* **291**, 27403–27420, doi:10.1074/jbc.M116.753053 (2016).
- 108 Cox, R. J. *et al.* Rapid cloning and expression of a fungal polyketide synthase gene involved in squalestatin biosynthesis. *Chem Commun (Camb)*, 2260–2261, doi:10.1039/b411973h (2004).

- 109 Whicher, J. R. *et al.* Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem Biol* **20**, 1340-1351, doi:10.1016/j.chembiol.2013.09.015 (2013).
- 110 Tang, Y., Chen, A. Y., Kim, C. Y., Cane, D. E. & Khosla, C. Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. *Chem Biol* **14**, 931-943, doi:10.1016/j.chembiol.2007.07.012 (2007).
- 111 Gay, Darren C. *et al.* A Close Look at a Ketosynthase from a Trans-Acyltransferase Modular Polyketide Synthase. *Structure* **22**, 444-451, doi:<https://doi.org/10.1016/j.str.2013.12.016> (2014).
- 112 Lohman, J. R. *et al.* Structural and evolutionary relationships of "AT-less" type I polyketide synthase ketosynthases. *Proc Natl Acad Sci U S A* **112**, 12693-12698, doi:10.1073/pnas.1515460112 (2015).
- 113 Gay, D. C. *et al.* The LINKS motif zippers trans-acyltransferase polyketide synthase assembly lines into a biosynthetic megacomplex. *Journal of Structural Biology* **193**, 196-205, doi:<https://doi.org/10.1016/j.jsb.2015.12.011> (2016).
- 114 Parker, R. E. & Parker, R. E. *Introductory statistics for biology*. Vol. 43 (Cambridge University Press, 1991).
- 115 Bhetariya, P. J. *et al.* Phylogenetic and Structural Analysis of Polyketide Synthases in *Aspergilli*. *Evol Bioinform Online* **12**, 109-119, doi:10.4137/EBO.S32694 (2016).
- 116 Moffitt, M. C. & Neilan, B. A. Evolutionary Affiliations Within the Superfamily of Ketosynthases Reflect Complex Pathway Associations. *Journal of Molecular Evolution* **56**, 446-457, doi:10.1007/s00239-002-2415-0 (2003).
- 117 Staunton, J. & Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat Prod Rep* **18**, 380-416, doi:10.1039/a909079g (2001).

- 118 Cox, R. J. Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets. *Org Biomol Chem* **5**, 2010-2026, doi:10.1039/b704420h (2007).
- 119 Dutta, S. *et al.* Structure of a modular polyketide synthase. *Nature* **510**, 512-517, doi:10.1038/nature13423 (2014).
- 120 Herbst, D. A., Jakob, R. P., Zahringer, F. & Maier, T. Mycocerosic acid synthase exemplifies the architecture of reducing polyketide synthases. *Nature* **531**, 533-537, doi:10.1038/nature16993 (2016).
- 121 Asturias, F. J. *et al.* Structure and molecular organization of mammalian fatty acid synthase. *Nat Struct Mol Biol* **12**, 225-232, doi:10.1038/nsmb899 (2005).
- 122 Sthapit, B. *et al.* Neocarzinostatin naphthoate synthase: an unique iterative type I PKS from neocarzinostatin producer *Streptomyces carzinostaticus*. *FEBS Lett* **566**, 201-206, doi:10.1016/j.febslet.2004.04.033 (2004).
- 123 Jean, M., Tomasi, S. & van de Weghe, P. When the nine-membered enediynes play hide and seek. *Organic & Biomolecular Chemistry* **10**, 7453-7456, doi:10.1039/C2OB26033F (2012).
- 124 Maeda, H., Ueda, M., Morinaga, T. & Matsumoto, T. Conjugation of poly(styrene-co-maleic acid) derivatives to the antitumor protein neocarzinostatin: pronounced improvements in pharmacological properties. *J Med Chem* **28**, 455-461, doi:10.1021/jm00382a012 (1985).
- 125 Soehano, I. *et al.* Insights into the programmed ketoreduction of partially reducing polyketide synthases: stereo- and substrate-specificity of the ketoreductase domain. *Organic & Biomolecular Chemistry* **12**, 8542-8549, doi:10.1039/C4OB01777C (2014).

- 126 Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-738, doi:10.1038/nprot.2010.5 (2010).
- 127 Wang, F. *et al.* Structural and functional analysis of the loading acyltransferase from avermectin modular polyketide synthase. *ACS Chem Biol* **10**, 1017-1025, doi:10.1021/cb500873k (2015).
- 128 Gay, D., You, Y. O., Keatinge-Clay, A. & Cane, D. E. Structure and stereospecificity of the dehydratase domain from the terminal module of the rifamycin polyketide synthase. *Biochemistry* **52**, 8916-8928, doi:10.1021/bi400988t (2013).
- 129 Bonnett, S. A. *et al.* Structural and stereochemical analysis of a modular polyketide synthase ketoreductase domain required for the generation of a cis-alkene. *Chem Biol* **20**, 772-783, doi:10.1016/j.chembiol.2013.04.014 (2013).
- 130 Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J Struct Biol* **170**, 427-438, doi:10.1016/j.jsb.2010.03.007 (2010).
- 131 Lattman, E. E. Small angle scattering studies of protein folding. *Current Opinion in Structural Biology* **4**, 87-92, doi:[https://doi.org/10.1016/S0959-440X\(94\)90064-7](https://doi.org/10.1016/S0959-440X(94)90064-7) (1994).
- 132 Durand, D. *et al.* NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. *J Struct Biol* **169**, 45-53, doi:10.1016/j.jsb.2009.08.009 (2010).
- 133 Sieber, S. A. *et al.* Evidence for a monomeric structure of nonribosomal Peptide synthetases. *Chem Biol* **9**, 997-1008, doi:10.1016/s1074-5521(02)00214-4 (2002).

- 134 Gallagher, J. R., Kim, A. J., Gulati, N. M. & Harris, A. K. Negative-Stain Transmission Electron Microscopy of Molecular Complexes for Image Analysis by 2D Class Averaging. *Curr Protoc Microbiol* **54**, e90, doi:10.1002/cpmc.90 (2019).
- 135 O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat Struct Mol Biol* **25**, 1000-1008, doi:10.1038/s41594-018-0147-0 (2018).
- 136 Lima, D. B. *et al.* Characterization of homodimer interfaces with cross-linking mass spectrometry and isotopically labeled proteins. *Nat Protoc* **13**, 431-458, doi:10.1038/nprot.2017.113 (2018).
- 137 Merkley, E. D. *et al.* Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. *Protein Sci* **23**, 747-759, doi:10.1002/pro.2458 (2014).
- 138 Nicolaou, K. C., Chen, J. S. & Dalby, S. M. From nature to the laboratory and into the clinic. *Bioorganic & Medicinal Chemistry* **17**, 2290-2303, doi:<https://doi.org/10.1016/j.bmc.2008.10.089> (2009).
- 139 Best, D., Jean, M. & van de Weghe, P. in *Dynemicin A, Uncialamycin and Analogues* (eds Daniel Best, Mickael Jean, & Pierre van de Weghe) 1-70 (Elsevier, 2016).
- 140 Tuttle, T., Kraka, E., Thiel, W. & Cremer, D. A QM/MM study of the Bergman reaction of dynemicin A in the minor groove of DNA. *J Phys Chem B* **111**, 8321-8328, doi:10.1021/jp072373t (2007).
- 141 Tuttle, T., Kraka, E. & Cremer, D. Docking, triggering, and biological activity of dynemicin A in DNA: a computational study. *J Am Chem Soc* **127**, 9469-9484, doi:10.1021/ja046251f (2005).

- 142 Kraka, E., Tuttle, T. & Cremer, D. Design of a new warhead for the natural enediyne
dynemicin A. An increase of biological activity. *J Phys Chem B* **112**, 2661-2670,
doi:10.1021/jp0773536 (2008).
- 143 Cohen, D. R. & Townsend, C. A. A dual role for a polyketide synthase in dynemicin
enediyne and anthraquinone biosynthesis. *Nat Chem* **10**, 231-236,
doi:10.1038/nchem.2876 (2018).
- 144 Liew, C. W. *et al.* Induced-fit upon Ligand Binding Revealed by Crystal Structures of
the Hot-dog Fold Thioesterase in Dynemicin Biosynthesis. *Journal of Molecular
Biology* **404**, 291-306, doi:<https://doi.org/10.1016/j.jmb.2010.09.041> (2010).
- 145 Liang, Z. X. Complexity and simplicity in the biosynthesis of enediyne natural
products. *Nat Prod Rep* **27**, 499-528, doi:10.1039/b908165h (2010).
- 146 Kastner, B. *et al.* GraFix: sample preparation for single-particle electron
cryomicroscopy. *Nat Methods* **5**, 53-55, doi:10.1038/nmeth1139 (2008).
- 147 Rial, D. V. & Ceccarelli, E. A. Removal of DnaK contamination during fusion protein
purifications. *Protein Expr Purif* **25**, 503-507, doi:10.1016/s1046-5928(02)00024-4
(2002).