University of Nottingham
UK | CHINA | MALAYSIA

Unpicking the Role of Acyl Carrier Proteins from Polyketide Synthases with Mass Spectrometry

by

Jeddidiah S. G. Bellamy-Carter, MSci

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

September 2019

Abstract

Polyketide synthases (PKS) are multi-domain enzymes responsible for the biosynthesis of structurally and functionally diverse natural products. A particular subfamily — *trans*-AT PKSs — have modular architecture and account for much of this diversity. These have attracted much attention as targets for engineering tailored PKSs to produce feed-stocks or synthetically challenging organic compounds. Phylogenetic and subsequent experimental studies have shown that ketosynthases (KS), the key domains for catalysing polyketide elongation, from *trans*-AT PKSs are highly substrate specific and may mediate polyketide structure. Until recently, the acyl carrier proteins (ACP) of *trans*-AT PKSs have been seen as passive carriers of the growing polyketide chain.

This thesis explores the role of ACPs through a series of mass spectrometry (MS) based and computational techniques. In Chapter 3, the interactions of acyl-ACPs from the psymberin PKS with cognate enzymes, including acyltransferase (AT) and acylhydrolase (AH) domains, show that ACPs are directly involved in controlling their acyl processing through protein interactions. In most cases, the observed specificities reflect the ACPs' native substrates and position within the PKS. The first ACP of the psymberin PKS was found to protect its native acetyl moiety and resist malonyl loading. Bioinformatics analysis of *trans*-AT ACPs through hidden Markov model profiling demonstrates the surprising relationship between ACPs and their downstream KSs, hinting at co-evolution and a mutual role in polyketide determination.

Native MS has been explored, in Chapter 4, as a prospective technique for studying PKS protein interactions in the gas-phase. In particular, collision-induced unfolding (CIU) shows great promise for understanding the structure of PKS proteins. A variety of PKS proteins — including whole modules — have been successfully analysed by native MS, however, determining acyl-state becomes impractical for larger ions.

Carbene footprinting is an emerging MS technique for the study of protein interactions, including those of PKSs. However, the data are often complex, making high-throughput analyses difficult. Chapter 5 describes the development of PepFoot, an open-source and user-interactive software, to streamline analysis and interpretation of footprinting data, paving the way for in-depth analysis and screening of PKS protein interactions. Benchmarking against experimental data-sets, showed highly comparable results to those published and additional insight — in a fraction of the time.

Acknowledgements

First off, I would like to thank Professor Neil Oldham for his support, guidance, and not least, the opportunity to work on this project. And a special thanks to Dr José Afonso, for starting me out on this project as an MSci student.

I would also like to thank Professor Panos Soultanas for providing a space for molecular biology work in CBS. I would like to thank Dr Matthew Jenner (University of Warwick) for his assistance and informative discussions. Much of the work herein would not have been possible without technical assistance both from outside and within the University, especially from Dave Litchfield who went above and beyond with mechanical and electrical repairs.

My time in the Oldham group would not have been complete without the friendly and supportive environment fostered by its members, in particular: Dr Lucio Manzi, Dr Vanderson Barbosa Bernado, Dr Mariana Breda, Dr Dan Scott, Asia Al-jabiry, Bruna da Silva Granja, Nkazi Tshuma and the many project students to have come through our lab. The same is true of the Panos, Searle and Thomas groups in CBS, with special mention to Faadil Fawzy, Manthos Pitoulias and Dr Vasilis Paschalis.

* * *

To my parents for their lifetime of support and encouragement, and to Abi for her immense love and reassurance — I would like to dedicate this thesis.

Contents

List of Figures

List of Tables

Listings

Abbreviations

ADP	Adenosine diphosphate
AmAc	Ammonium acetate
AmBic	Ammonium bicarbonate
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
CoA	Coenzyme A
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiotheritol
eqv.	Equivalent
FA	Formic acid
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropylthio-β-galactoside
HMM	Hidden Markov model
LB	Luria-Bertani broth
LIC	Ligation independent cloning
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced form)
NMR	Nuclear magnetic resonance
PCR	Polymerase chain reaction
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SLIC	Sequence and ligation independent cloning
SNAC	Small *N*-acetyl cysteamine
TDBA	4-(3-Trifluoromethyl)-3*H*-diazirin-3-yl)benzoic acid
TFA	Trifluoroacetic acid
2xYT	2×Yeast-tryptone broth

Polyketide Synthases

ACP	Acyl Carrier Protein
A-domain	Adenylation domain
AH	Acylhydrolase domain
AL	Acyl ligase
AT	Acyltransferase domain
B-domain	Branching domain
C-domain	Condensation domain
Cy	Cyclase domain
DH	Dehydratase domain
ECH	Enoyl-CoA hydratase domain

ER	Enoylreductase domain
FAS	Fatty acid synthase
GNAT	GCN5-related domain *N*-acetyltransferase domain
HMGS	HMG-CoA synthase
KR	Ketoreductase domain
KS	Ketosynthase domain
MT	Methyltransferase domain
NRPS	Non-ribosomal peptide synthase
OMT	*O*-methyl-transferase domain
Pant	Pantetheine
PKS	Polyketide synthase
Ppant	4′-Phosphopantetheine
PPTase	4′-Phosphopantetheinyl transferase
PS	Pyran synthase domain
TE	Thioesterase

Mass Spectrometry

ATD	Arrival time distribution
CCS	Collision
CID	Collision-induced dissociation
CIU	Collision-induced unfolding
CRM	Charged-residue model
ECD	Electron capture dissociation
ESI	Electrospray ionisation
FID	Free induction decay
FPOP	Fast photochemical oxidation of proteins
FT-ICR	Fourier-transform ion cyclotron resonance
HDX	Hydrogen–deuterium exchange
HPLC	High-performance liquid chromatography
HRF	Hydroxy radical footprinting
IEM	Ion evaporation model
IMS	Ion mobility spectrometry
LC	Liquid chromatography
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
m/z	Mass-to-charge ratio
nESI	Nano-electrospray ionisation
nMS	Native mass spectrometry
Q-ToF	Quadropule–Time-of-flight hybrid
RF	Radio frequency
ToF	Time of flight
TWIMS	Travelling-wave ion mobility spectrometry

Amino Acids

Ala	`A`	Alanine
Asn	`N`	Asparagine
Asp	`D`	Aspartic acid
Arg	`R`	Arginine
Cys	`C`	Cysteine
Gln	`Q`	Glutamine

Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine

1 Introduction

1.1 Mass Spectrometry

First developed by Thomson[1] and Aston[2], mass spectrometry (MS) is a powerful analytical technique by which gas-phase ions are separated according to their mass-to-charge ratio (m/z), rather than their mass alone. In the century since the first mass spectrometer (a parabola spectrograph), a variety of MS techniques have been developed to analyse everything from chemical nerve agents[3] to intact viral oligomers.[4] A mass spectrometer consists of three principal components: a source of ionisation to provide the gas-phase ions, a mass analyser to separate them by m/z — which may be combined with other techniques to interrogate analyte ions, and a detector to quantify separated ions.

1.1.1 Ionisation Sources

Ionisation sources (often called ion sources) are required to generate gas-phase ions from solid, liquid and gaseous analytes. An important consideration in ionisation is the internal energy transferred to generate ions; hard ionisation techniques are very energetic and can typically lead to extensive fragmentation of ions, while soft ionisation techniques impart little energy and produce intact ions.[5] The choice of ion source largely depends on the analyte in question. Electron ionisation (EI) and chemical ionisation (CI) are only applicable to gas-phase analytes, and as such are used for volatile and

thermally stable small molecules. Electrospray ionisation (ESI), atmospheric pressure chemical ionisation (ACPI) and atmospheric pressure photoionisation (APPI) are used for analytes in solution whereby the solution is nebulised into droplets at atmospheric pressure before introduction into a mass spectrometer under vacuum. Matrix-assisted laser desorption ionisation (MALDI) and secondary ion mass spectrometry (SIMS) are typically used for solid-state samples. The only ionisation source described in detail herein is ESI, since it is the sole method employed in this work.

1.1.1.1 Electrospray Ionisation

Electrospray ionisation (ESI) is one of the most useful techniques for production of gas-phase ions from biological samples, as the soft ionisation technique can preserve both covalent and non-covalent interactions.[6] ESI was first developed experimentally by Zeleny[7] in 1914 with the application of an electric field to the liquid meniscus in a glass capillary, then in 1917 he published the first time-lapse images of cone-jet formation.[8] This work was further expanded by Taylor between 1964 and 1969 where the cone-jet formation was characterised and modelled, the distinctive formation is thus called a 'Taylor Cone'.[9–11]

The first published coupling of ESI with MS was by Dole *et al.*[12] in 1968, in which high-molecular-weight polystyrene chains were sprayed from benzene and acetone. It was not until 1989, however, that Fenn *et al.*[13] adapted the technique to analyse multiply charged biomolecule samples; the developments made by Fenn eventually led to him being jointly awarded the 2002 Nobel Prize in Chemistry.

ESI involves applying a strong electric field (~10^6 V/m) over an analyte of interest in a volatile solvent (such as aqueous ammonium acetate (AmAc) or acetonitrile (MeCN)) in a capillary tube. The electric field is produced by an electrical potential (1.5–4 kV) between the capillary and a sampling device; this causes solvent ions of opposite charge to the potential to accumulate at the tip of the solvent. As the ions accumulate, the surface tension of the solvent is overcome by the Coulomb repulsion of the ions and the tip begins to deform into the Taylor cone with the repulsion causing droplets to expel from the tip, forming a spray. These droplets then begin to desolvate as the solvent evaporates thus reducing the size of the droplet and increasing the charge density until again the Coulomb repulsion exceeds the surface tension of the droplet.

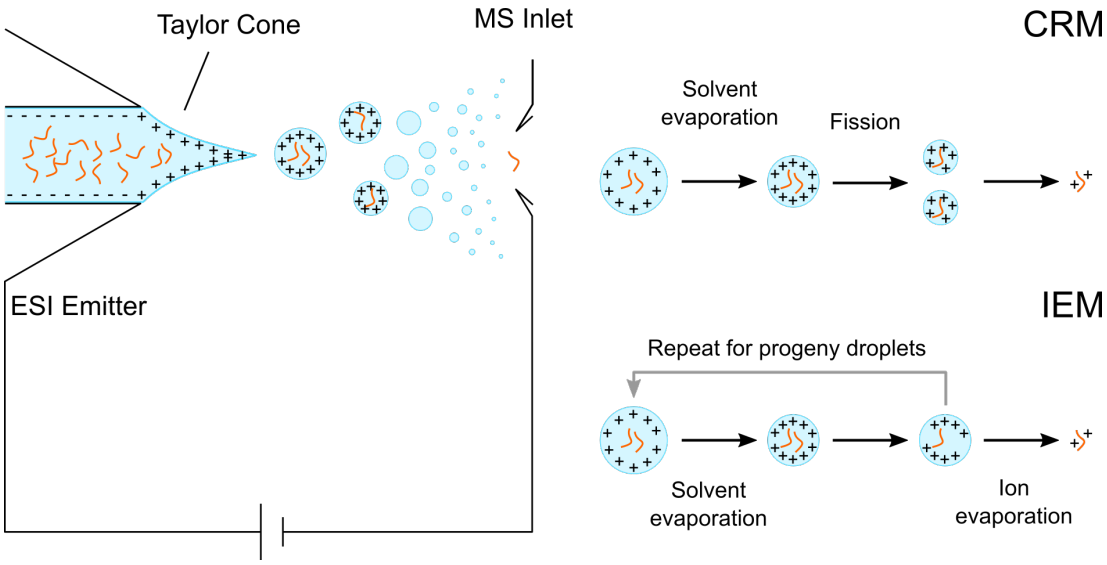


Figure 1.1 | Diagram of ESI process in positive ion mode *(left)*. CRM and IEM droplet evolution models *(right)*.

This point, known as the Rayleigh[14] limit, is defined in Equation 1.1 by the droplet charge q_R, the permittivity of the environment ε_0, surface tension of the solvent γ and the diameter of the droplet D, where z_R is the number of charges in the droplet and e is the elementary unit of charge.

$$q_R = z_R e = \pi \sqrt{8\varepsilon_0 \gamma D^3} \tag{1.1}$$

Work by Gomez and Tang[15] revealed that under the strong electric field the droplets themselves deform to give a Taylor cone, which allows expulsion of daughter droplets prior to reaching the Rayleigh limit mechanically. The daughter droplets are typically ~2 % of the volume and ~15 % of the charge of the precursor droplet, giving 7-fold charge density and promoting further Coulomb fission. The latter stages of this process, where the gas-phase ions are produced from the droplets, is still subject to debate. The two prevailing models are the charged-residue model (CRM)[12] and the ion evaporation model (IEM).[16] CRM suggests that the Coulomb fission process perpetuates until no further solvent can evaporate and all charges within the droplet are transferred onto the analyte. By contrast, IEM suggests that charged analyte ions are ejected from the droplet when a certain radius is reached prior to the Rayleigh limit. The appropriate model depends upon the analyte in question, small molecules appear to favour IEM whereas larger analytes such as proteins favour CRM.[17]

1.1.1.2 Electrospray Ionisation of Proteins

The mass spectra obtained for proteins from ESI contain a distribution of ions corresponding to the multiple protonation states of the protein upon entering the gas-phase, often represented as $[M + nH]^{n+}$ when assigning spectral peaks. Example spectra for denatured and native myoglobin are shown in Figure 1.2. The charge state distribution is dependent upon the structure of the protein as it desolvates. The maximum charge state for any ion can be approximated from the Rayleigh limit (Equation 1.1), where the diameter of the droplet is limited to the size of the analyte with larger analytes capable of acquiring a higher number of charges. Indeed, De La Mora[17] found that the maximum charge state of native globular proteins is approximate to that of spherical droplets with the same radii and defined a relationship to approximate the average charge state z_{av} of a native globular protein, given its mass M (Equation 1.2). These results support the CRM model for large proteins (>6 kDa). Conversely, denatured proteins were observed to have much greater charge states than the Rayleigh charge limit z_R, which was rationalised by droplet elongation at a diameter equivalent to the maximum distance between two polar sites when denatured.

$$z_{av} \approx 0.0778\sqrt{M} \tag{1.2}$$

The mass M of a protein can be determined from an ESI mass spectrum, considering adjacent peaks (m_1 and m_2) to be the same protein with a difference of one proton m_p and thus charge state (z_1 and z_2, where $z_2 = z_1 - 1$),

$$m_1 z_1 = M + z_1 m_p \quad \text{and} \quad m_2 z_2 = M + (z_1 - 1)m_p \tag{1.3}$$

These can be rearranged to determine z_1 and thus M

$$z_1 = \frac{m_2 - m_p}{m_2 - m_1} \quad \text{and} \quad M = z_1(m_1 - m_p) \tag{1.4}$$

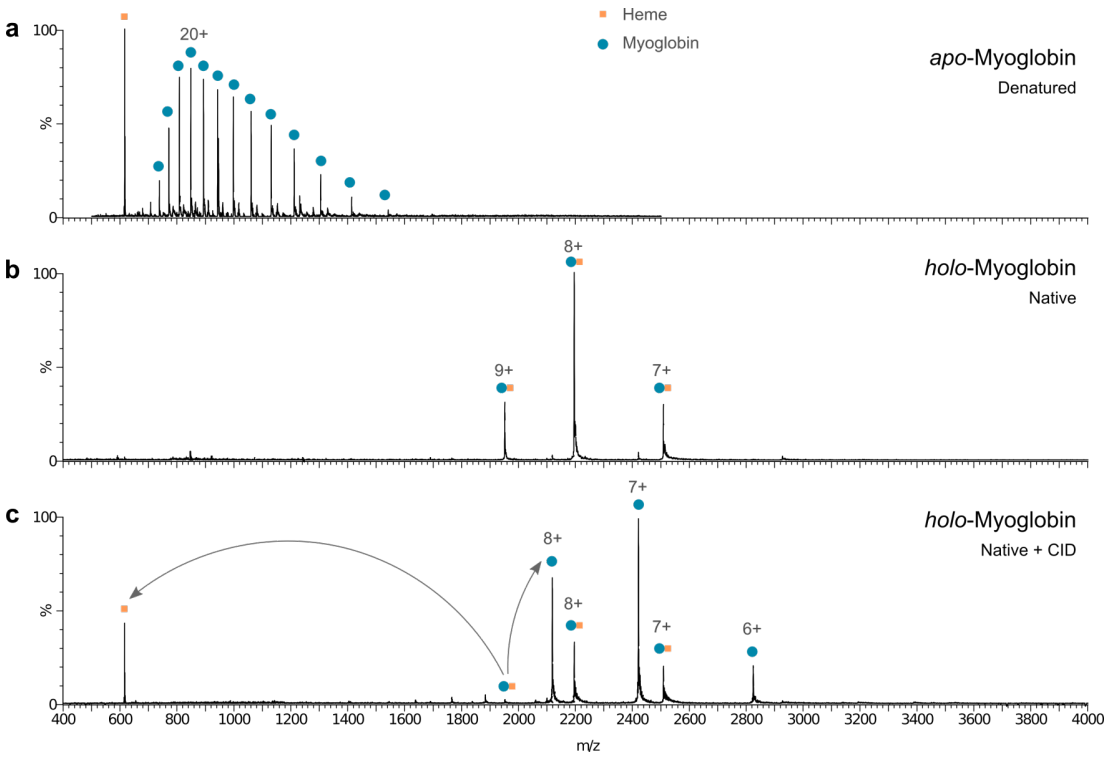


Figure 1.2 | ESI spectra of myoglobin, protein charge states indicated with blue circle, heme indicated with orange square. (**A**) Denatured myoglobin showing high charge states of protein and release of non-covalent ligand heme, sprayed from 50 % MeCN, 0.1 % HCOOH. (**B**) Native myoglobin showing low charge states and complex with heme, sprayed from 25 mM NH_4OAc. (**C**) As (B), with collisional activation to release heme and *apo*-myoglobin.

1.1.1.3 Nano-Electrospray Ionisation

Conventional ESI uses 0.5 mm diameter capillaries with flow rates of several µL/min. The development of emitter tips with ∼1 µm diameters by Wilm and Mann[18] allowed for the use of much smaller sample volumes and flow rates on the order of 10 nL/min, so-called nano electrospray ionisation (nESI). These emitters are typically static and are fabricated from borosilicate capillary tubes coated in gold[18] or back-fitted with platinum wire[19] and lower potentials of 0.5–1.5 kV applied. The advantages of nESI are not limited to sample volumes, however. It has been demonstrated that nESI is more tolerant to salts[20] and provides more stable protein complexes[21] than ESI, which can be attributed to the smaller initial droplet size and reduced energy required for desolvation.

1.1.2 Mass Analysers

Gas-phase ions are separated according to their m/z by a mass analyser. The variety of analysers available take advantage of different principles to separate ions. These principles are kinetic energy (electric sector), momentum (magnetic sector), trajectory stability (quadrupole), velocity (time-of-flight) and resonance frequency (ion trap, ion cyclotron and orbitrap).[5] Quadrupole, time-of-flight and ion cyclotron resonance analysers are discussed in more depth herein.

1.1.2.1 Quadrupole

The quadrupole mass analyser was first described by Paul and Steinwedel[22] in 1953. The device consists of, as the name implies, four perfectly parallel rods of circular or hyperbolic cross-section, and uses stable ion trajectories in an oscillating electric field to separate ions by m/z. The oscillating electric field is generated by applying an alternating potential Φ_0 to rods such that adjacent rods have opposite polarity. Positive ions travelling between the rods will be attracted towards the negative rods and repulsed by the positive rods; providing the rods change polarity before the ion discharges, the ion will change direction. The applied potential is a superposition of an alternating radio-frequency (RF) field, with zero-to-peak amplitude V and angular frequency ω, on a constant field, with direct potential U, is defined as:

$$\Phi_0 = U - V\cos\omega t \tag{1.5}$$

$$\Phi = \frac{\Phi_0(x^2 - y^2)}{r_0^2} = \frac{(U - V\cos\omega t)(x^2 - y^2)}{r_0^2} \tag{1.6}$$

The potential within the field at point (x, y) is defined by Φ in Equation 1.6, where r_0 is the internal radius of the quadrupole. By considering the forces applied to the ions by the electric fields (shown in Equation 1.7, where m and z are the mass and charge of an ion respectively and e is the elementary unit of charge), second-order differential equations for the movement of these ions can be derived, Equation 1.8.

$$F_x = m\frac{d^2x}{dt^2} = -ze\frac{\partial\Phi}{\partial x} \tag{1.7a}$$

$$F_y = m\frac{d^2y}{dt^2} = -ze\frac{\partial\Phi}{\partial y} \tag{1.7b}$$

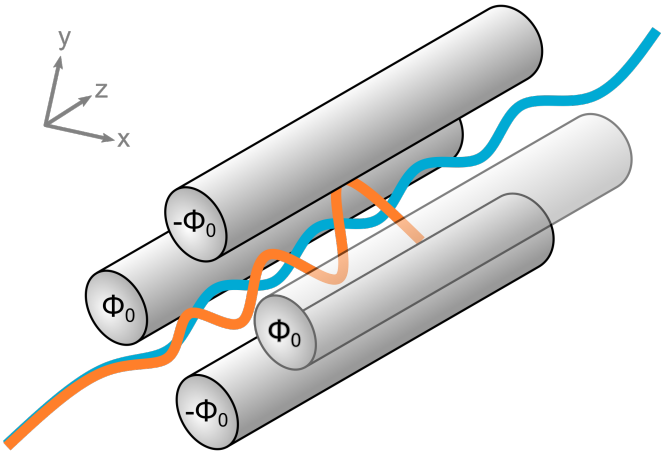


Figure 1.3 | Schematic of a quadrupole mass analyser. The trajectories of two ions are shown in blue and orange, with a stable and unstable trajectory respectively.

$$\frac{d^2x}{dt^2} + \frac{2ze}{mr_0^2}(U - V\cos\omega t)x = 0 \tag{1.8a}$$

$$\frac{d^2y}{dt^2} - \frac{2ze}{mr_0^2}(U - V\cos\omega t)y = 0 \tag{1.8b}$$

Comparing these differentials with the Mathieu equation (Equation 1.9), dimensionless parameters that define regions of stability, a_u and q_u, can be derived.[5,23]

$$\frac{d^2u}{d\xi^2} + (a_u - 2q_u\cos 2\xi)u = 0 \tag{1.9}$$

Substituting either x or y for u and defining $\xi = \omega t/2$, yields the expressions:

$$a_u = a_x = -a_y = U\frac{8ze}{m\omega^2 r_0^2} \qquad \text{and} \qquad q_u = q_x = -q_y = V\frac{4ze}{m\omega^2 r_0^2} \tag{1.10}$$

Which may be rearranged to give

$$U = a_u\frac{m}{z}\frac{\omega^2 r_0^2}{8e} \qquad \text{and} \qquad V = q_u\frac{m}{z}\frac{\omega^2 r_0^2}{4e} \tag{1.11}$$

For a given quadrupole and RF, r_0 and ω are constant. These expressions establish a relationship between m/z and the values of U and V. Furthermore, the direct relationship between (a_u, q_u) and (U, V) respectively allows the stability of an ion with m/z to be related to these values as shown in Figure 1.4. The bounded triangular areas represent overlap between stable trajectories in the x and y plane for a given

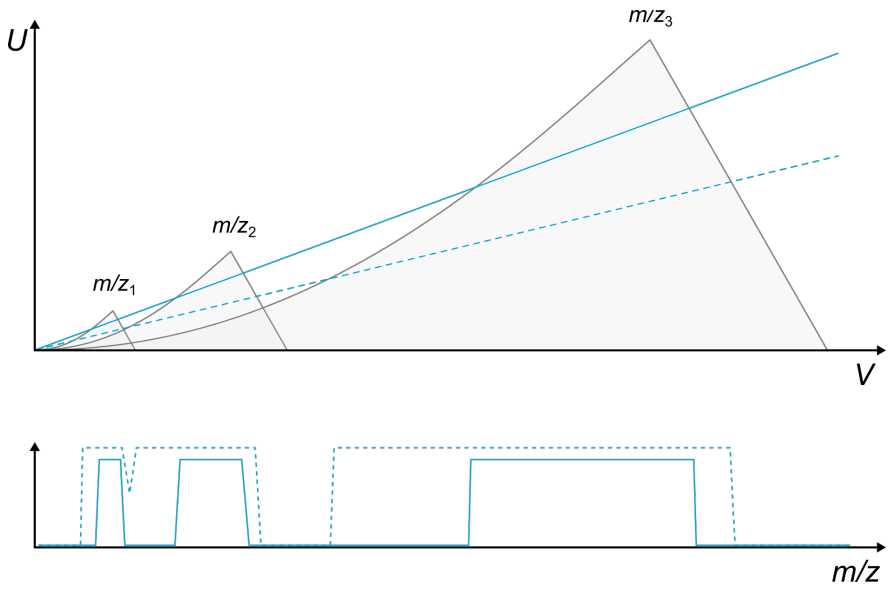


Figure 1.4 | Areas of stability for ions of m/z: m/z_1, m/z_2, and m/z_3 ($m/z_1 < m/z_2 < m/z_3$) with respect to U and V (*top*). Two scan lines with different U/V are shown. The apparent m/z for which these ions can reach a detector show that increasing U/V reduces simultaneous transmission (*bottom*).

m/z; these areas are proportional to m/z. Considering a scan line of constant U/V ratio; at any point on this line, m/z overlapping this point will have stable trajectories and reach the detector. By increasing the gradient of U/V, fewer m/z will be simultaneously transmitted and therefore the resolution increases.

$$M_{max} = \frac{7 \times 10^6 V}{f^2 r_0^2} \tag{1.12}$$

Operating in RF-only mode ($U = 0$) gives no resolution of ions by m/z, but instead ions with m/z stability area below V are discharged, acting as a high-pass filter. The transmission of ions decreases with their m/z, as the focussing of ions towards the centre of the quadrupole is inversely proportional to their m/z and proportional to V^2. The theoretical upper-mass limit M_{max} of a quadrupole can be described by Equation 1.12,[24,25] where f is the frequency of the RF ($\omega/2\pi$). Thus by reducing the RF frequency it is possible to extend the upper-mass limit of a quadrupole to be more suitable for large macromolecules, at the cost of mass resolution in the lower m/z range. In order for a quadrupole to efficiently transmit ions over a wide dynamic range, the value of V must be ramped; this is often termed the *quad profile* and is used to tune for optimum transmission of target ions in hybrid setups.

1.1.2.2 Time-of-Flight

The time-of-flight (ToF) analyser concept was proposed by Stephens[26] in 1946. Ions are separated according to their velocity in a field-free flight tube after initial acceleration by an electric field. Ions with the same nominal kinetic energy have a velocity proportional to their mass, with smaller ions travelling faster than larger ions. A linear ToF is composed of a source region and a detector separated by a field-free flight tube under high vacuum. A high electric potential V_a is applied within the source to accelerate ions into the flight-tube, transferring a potential energy E_p to each ion, as given by

$$E_p = zeV_a = E_k = \frac{mv^2}{2} \tag{1.13}$$

where z is the charge of an ion, e is the elementary unit of charge and m is the mass of the ion. This potential energy is converted into kinetic energy E_k, allowing the velocity v of ions to be expressed in terms of their mass and charge, as shown by

$$v = \sqrt{\frac{2zeV_a}{m}} \tag{1.14}$$

Since the ion travels at constant velocity after entering the flight tube, the time t taken for an ion to reach a detector after travelling the length L of the flight tube is inversely proportional to v

$$t = \frac{L}{v} \tag{1.15}$$

Substituting Equation 1.14 for v and rearranging gives

$$t^2 = \frac{m}{z}\frac{L^2}{2eV_a} \tag{1.16}$$

For a given ToF and electric potential, L and V_a are constant, allowing m/z to be calculated directly from t^2. There is no theoretical upper mass limit for a ToF, making it an ideal analyser for large biomolecules that can be in the MDa range. The resolution of a ToF is proportional to the flight-path, increasing the effective distance travelled by ions increases the resolution between them. While simply increasing the length of the flight tube would achieve this, the non-uniformity of kinetic energy among ions

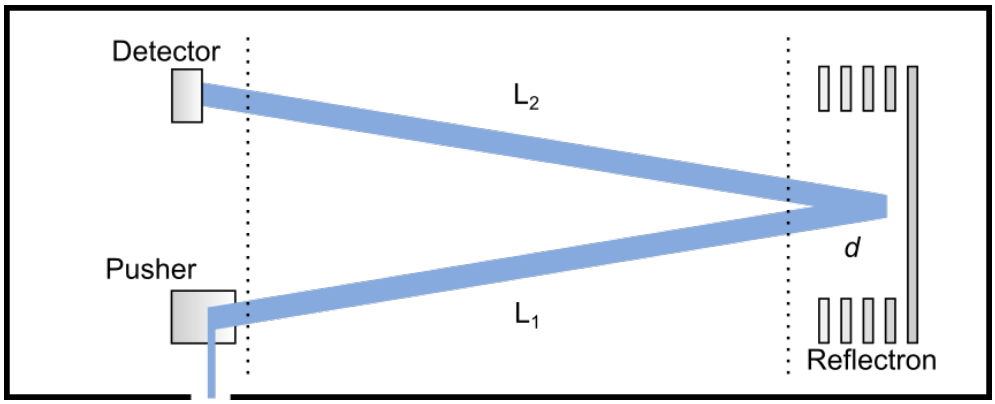


Figure 1.5 | Schematic of a reflectron ToF device. Ion beam indicated as blue line. L_1, L_2 and d are the lengths and depth traversed by ions.

entering the flight tube is chiefly responsible for poor resolution, as ions have varying velocities and thus a distribution of arrival times at the detector.

Relflectron

An electrostatic reflector (reflectron) was first proposed by Mamyrin *et al.*[27] in 1973. It acts as an ion mirror by providing a homogeneous electric field E_R to reverse the direction of the ion beam. The reflectron is placed at the opposite end of the flight tube to the source and detector. An ion of charge q and velocity v will penetrate the field to a depth d such that

$$d = \frac{E_k}{qE_R} = \frac{V_a}{E_R} \tag{1.17}$$

at which the velocity of the ion is zero and will begin to accelerate out of the reflectron with a final velocity equal to v. The mean velocity of an ion within the reflectron to d is $v/2$. The time spent within the reflectron t_R can therefore be inferred as

$$t_R = 2(\frac{d}{v/2}) = \frac{4d}{v} \tag{1.18}$$

Ions with greater E_k will penetrate to a greater depth and spend longer in the reflectron. The total flight time out of the reflectron is proportional to $L_1 + L_2$ where L_1 and L_2 are the distances travelled before and after the reflectron at velocity v respectively. The total flight time t can be expressed as

$$t = \frac{L_1 + L_2 + 4d}{v} \tag{1.19}$$

Which may be substituted and rearranged as Equation 1.16 to give

$$t^2 = \frac{m}{z}\frac{(L_1 + L_2 + 4d)^2}{2eV_a} \tag{1.20}$$

By compensating for the kinetic energy dispersion for ions of the same m/z, the reflectron is able to reduce arrival time distributions and greatly increase resolution.

1.1.2.3 Fourier Transform - Ion Cyclotron Resonance

The Fourier transform ion cyclotron resonance (FT-ICR) mass analyser was first developed by Comisarow and Marshall[28]. Ions are separated according to their ion cyclotron frequency in a homogeneous magnetic field. The cyclotron is formed of a cell under high vacuum surrounded by a superconducting magnet — the cell consists of a set of excitation and detection plates. Unlike quadrupole and ToF analysers, ions are detected by passing near the detection plates, rather than hitting a detector. Ions confined within the cell will orbit perpendicular to the static magnetic field B with stable circular trajectories defined by centrifugal F_C and Lorentz F_B forces:

$$F_C = \frac{mv^2}{r} = F_B = zevB \tag{1.21}$$

where v and r are the velocity and radius of the ion respectively. These can be rearranged and converted to angular cyclotron frequency ω_c,

$$\omega_c = \frac{v}{r} = \frac{zeB}{m} \tag{1.22}$$

Equation 1.22 shows that at constant r and B, the frequency at which an ion orbits in the cell is inversely proportional to its m/z. Unperturbed in a vacuum, ω_c remains constant with ions orbiting at a radius determined by their velocity. A sufficiently high velocity applied to an ion will expel it from the cell. Applying an RF field across the excitation plates with the same frequency as ω_c will allow resonance absorption and increase the kinetic energy of an ion and thus its velocity and radius.

After initial injection, ions do not have sufficient velocity and therefore a wide enough radius to be detected. Rapidly scanning a large frequency range simultaneously excites all of the ions in the cell to a detectable radius. As the ions relax from this orbit

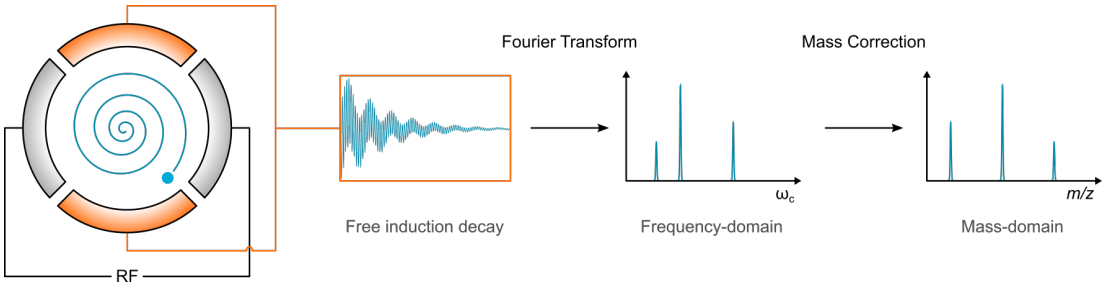


Figure 1.6 | Schematic of FT-ICR. After excitation by an RF applied to excitation plates (grey), an ion (blue) orbits around the cell towards the centre axis with a frequency equal to ω_c, which is recorded by detection plates (orange) as a FID.

they are detected by the detection plates, producing a complex wave that is the sum of sine waves with ω_c for all ions, called a free induction decay (FID) or transient (see Figure 1.6). A Fourier transform is applied to the FID to generate a 'frequency-domain' spectrum, this is in turn converted into the 'mass-domain' using Equation 1.22.

1.1.3 Tandem Mass Spectrometry

Tandem mass spectrometry — often abbreviated MS/MS — is the coupling of multiple mass analysis stages, typically with an activation step between them. When more than two stages of mass analysis are involved the experiment is referred to as MS^n, where n refers to the number of mass analysis steps. Tandem MS experiments can be carried out in space or time, where the analysis is performed by two different analysers in series (e.g. Q-ToF) or the same analyser (e.g. FT-ICR). The activation step allows for generation of new chemical species, typically fragment product ions, enabling identification of precursor ions.

1.1.3.1 Hybrid Instruments

Hybrid instruments are the combinations of different mass analysers in the same instrument. One of the most common hybrid geometries is the QqToF (often abbreviated to QToF), a quadrupole mass filter (Q) coupled to a ToF with an intermediate collision cell (q). In this configuration the quadrupole can transmit a wide range of ions in RF-only mode or a narrow specific m/z in resolving mode, where the selected ions can be fragmented in the collision cell followed by scanning of the full mass range with the ToF. A related hybrid Q-ToF with integrated ion mobility, the Synapt, is shown in Figure 1.7.

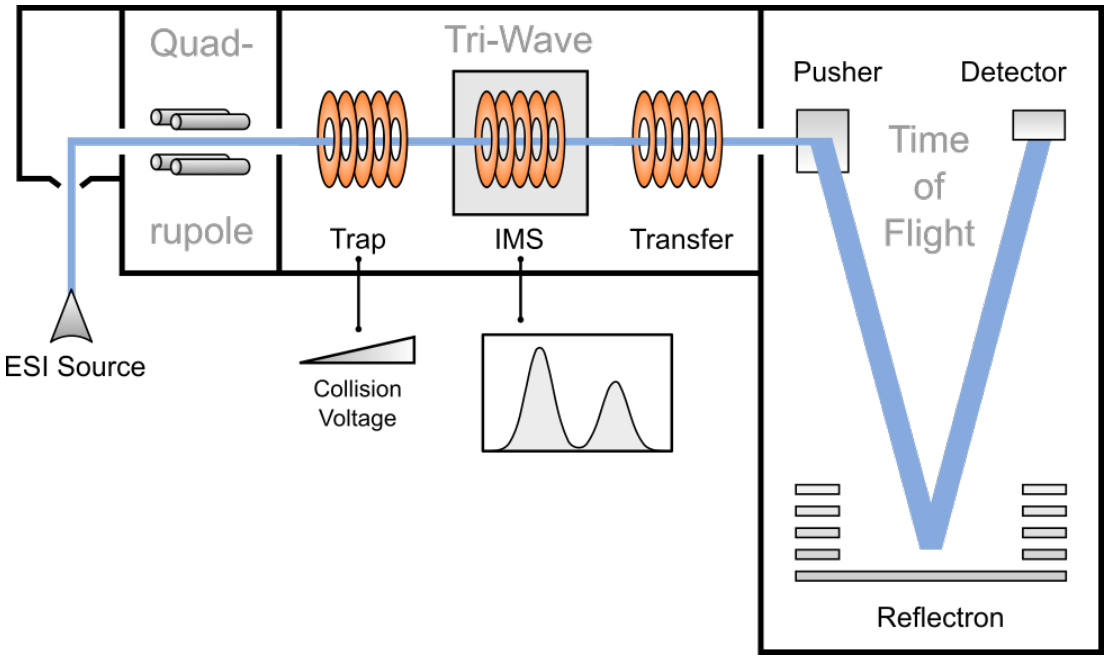


Figure 1.7 | A simplified schematic diagram of the Synapt HDMS instrument. Ion beam is indicated in blue with the stacked ring ion guides for the Tri-Wave indicated in orange.

1.1.3.2 Activation Techniques

A variety of activation methods are available for tandem mass spectrometry experiments, with collision induced dissociation (CID) being the most prominent and available to most commercial instruments. Other, more specialised, techniques such as photodissociation (PD), electron capture dissociation (ECD) and electron transfer dissociation (ETD) are more instrumentally restrictive. The general process of energy transfer by these methods can be described as *ergodic* or *non-ergodic*, in the former, the energy is distributed across vibrational modes of the ion before dissociation occurs whereas in the latter, the rate of dissociation is faster than a typical bond vibration. Ergodic ion-activation methods cause preferential cleavage of the weakest bonds in an ion whereas non-ergodic methods cause cleavage at the site of energy transfer.[29]

CID occurs when analyte ions are accelerated into a chamber of inert gas molecules such as N_2 or Ar. Collisions cause transfer of kinetic energy into internal energy, with the maximum energy transferred during collision in the centre-of-mass reference frame (E_{com}) being described by

$$E_{com} = E_{lab}\frac{M_g}{M_i + M_g} \tag{1.23}$$

where E_{lab} is the ions kinetic energy in the laboratory reference frame ($z \times V_a$), M_g and M_i are the masses of the neutral gas and target ion respectively. Consequently, increasing the ion kinetic energy or the neutral gas mass will result in a greater transfer of energy whilst increasing target ion mass decreases the available energy. As E_{lab} is proportional to the charge of an ion and the accelerating potential applied to it, high charge ions require less acceleration to achieve equivalent dissociation.

1.1.3.3 Peptide Fragmentation

Tandem mass spectrometry experiments can be used to obtain additional information about the composition of peptides, as characteristic product ions are generated by fragmentation at amide bonds.[30] As shown in Figure 1.8, there are three ion pairs generated along the amide bond with types a, b and c occurring on the N-terminal side of the cleavage, and corresponding x, y and z for the C-terminal side for Cα–C, C–N and N–Cα bonds respectively. The position of cleavage along the peptide is indicated with a subscript (e.g. b_n), where n is the number of amino acids in the fragment ion. As the m/z of these fragments are typically unique, the mass shift between two adjacent ions in series (e.g. b_3 and b_4) is the condensed mass of the amino acid at that position.

Ergodic methods, such as CID, typically fragment to give b and y ions, as this is the kinetically weakest bond in the amide. The strained C–N bond in proline containing peptides leads to preferential cleavage and generation of dominant y-ions,[31] while fragments containing basic residues are typically more intense.[32] Side-chain specific bonds may also be broken during CID, amino acids RKNQ can lose ammonia (–17 Da) and STED can lose water (–18 Da); labile post translational modifications such as phosphorylation may also be lost.

1.1.4 Native Mass Spectrometry

Native mass spectrometry is defined here as any mass spectrometry technique that allows interrogation of the native structures and interactions of biomolecules. Such techniques include the more formal native MS (nMS), where ions are maintained in a native-like state in the gas-phase, as well as integrative techniques such as cross-linking and protein footprinting where the 'memory' of solution-phase structure is analysed by mass spectrometry.

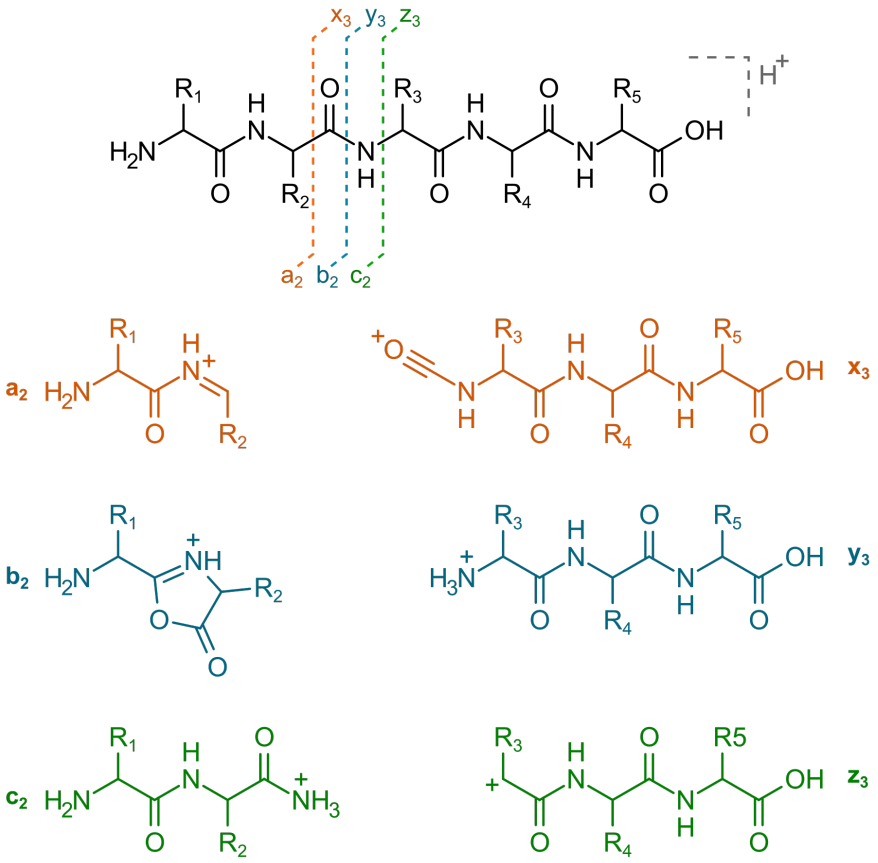


Figure 1.8 | Scheme of peptide backbone fragmentation. A representative pentapeptide is shown with corresponding fragment ions.

1.1.4.1 Native-like Ions in the Gas Phase

The soft ionisation provided by ESI and its derivatives allows for non-denaturing conditions to be applied to ions when entering the gas-phase, so-called 'native' ions.[33] While there is much debate about the extent of solution-like structure retention in the gas-phase, there is considerable evidence for gaseous proteins retaining some interactions and conformations.[34] This is typically achieved through the use of non-denaturing volatile solvents, such as aqueous ammonium acetate (AmAc), and conditions that reduce collisional heating of ions prior to detection.[35] A particular problem for nMS is the presence of non-specific salt adducts (e.g. Na^+ and K^+), these adducts can cause broadening of highly charges peaks and can perturb gas-phase protein-ligand complexes.[36] These adducts can be suppressed using nESI,[20,37,38] collisional cleaning,[39–41] exposure to solvent vapour during desolvation,[42] supercharging additives,[43,44] and even amino acids[45] in the analyte solution. Compared with traditional methods of investigating protein interactions, mass spectrometry

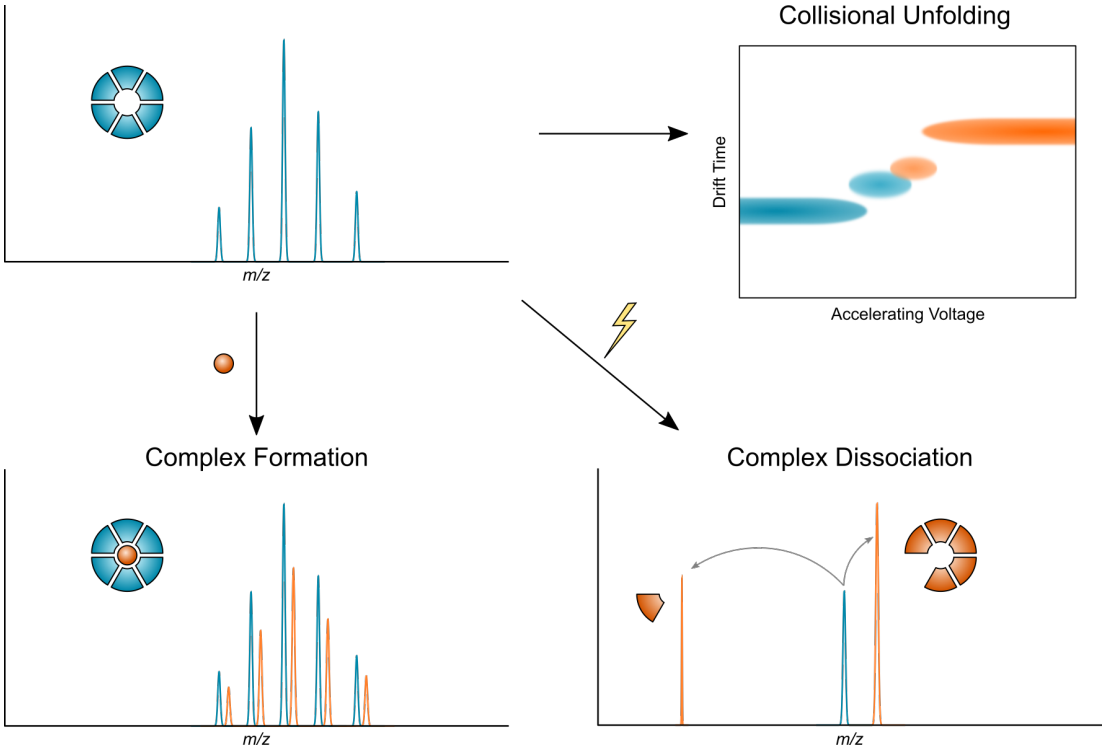


Figure 1.9 | **Native mass spectrometry overview.** nMS can be used to probe protein structure and interactions in the gas-phase. Interaction stoichiometry can be probed with varying concentrations of binding partners to reveal complex formation (*bottom-left*). Protein complexes can be dissociated under gentle collisional activation to reveal affinity (*bottom-right*). Protein unfolding in the gas-phase can be produced under collisional activation and measured with ion mobility in CIU experiments (*top-right*).

requires small sample volumes and provides the dimension of mass analysis. Complex protein stoichiometries can be determined with mass precision[6,21,46,47] and complexes can be dissociated under collisional activation to determine binding affinities,[48,49] see Figure 1.9.

The pre-eminent instrumentation for nMS experiments is the Q-ToF geometry, where the quadrupole is used as a mass filter with occasional mass selection and the wide mass range and sensitivity of the ToF allows for transmission and analysis of high mass ions. The typical quadrupole used in early commercial Q-ToF instruments had an upper mass limit of 4000 m/z, which would correspond to the average m/z of a 100 kDa globular protein according to Equation 1.2. To examine larger proteins and protein complexes, modified instruments with reduced RF-frequency quadrupoles[24] and higher source pressures[50] were developed. Using these modified instruments and nESI, the stoichiometries of large protein complexes[46] including intact viral capsids,[4] is possible.

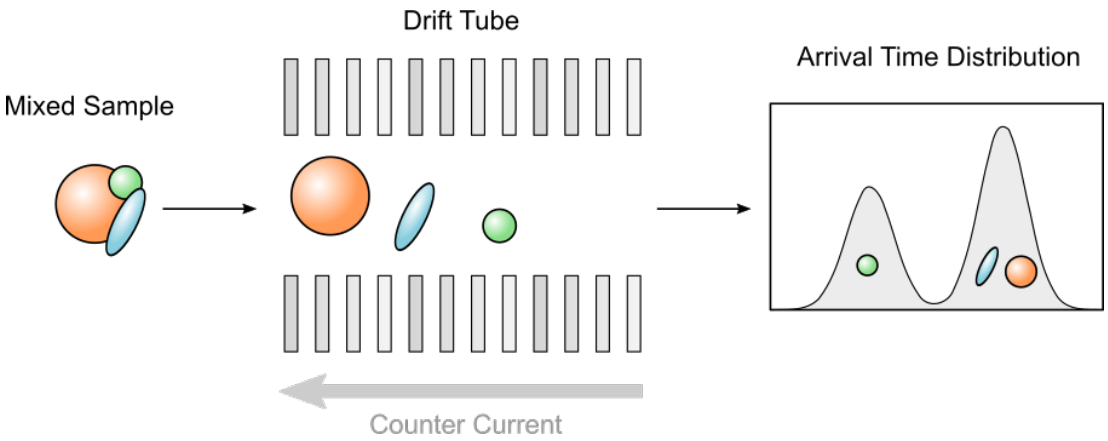


Figure 1.10 | Diagram of a basic drift tube IMS experiment with a mix of ions then separated by ion mobility. Where the orange circle and blue ellipse have similar Ω_0

1.1.4.2 Ion Mobility

Ion mobility spectrometry (IMS) is a technique for separating gas-phase ions by their electrical mobility K in a carrier gas.[51] Considering a drift tube filled with carrier gas (typically N_2 or He) and a static electric field E, the ions average drift velocity v_D can be converted into its rotationally averaged collision cross-section Ω_0 with the Mason-Schamp[52] equation

$$v_D = KE, \qquad \Omega_0 = \frac{ze}{K_0}\frac{3}{16N}\sqrt{\frac{2\pi}{\mu k_B T}} \tag{1.24}$$

where K_0 is the reduced mobility (K corrected to standard temperature and pressure), N is the number density of the gas, μ is the reduced mass of the analyte and carrier gas $\left(\frac{m_1 m_2}{m_1+m_1}\right)$, z is the charge state of the ion, e is the elementary charge, k_B is the Boltzmann constant and T is the temperature. As analyte ions–carrier gas collisions are normally-distributed, the arrival time distribution (ATD) is Gaussian, the centroid of which is Ω_0. The mobility of an ion is inversely proportional to Ω_0, ions with larger collisional cross-sections will undergo more collisions and have reduced velocity compared with ions of the same m/z but more compact cross-sections and thus have later ATDs. This allows for the folding state of protein ions to be determined, with more folded conformations exhibiting shorter drift times, as demonstrated by Clemmer *et al.*[53] for cytochrome c.

Travelling Wave Ion Mobility

Conventional drift tube ion mobility instruments can result in radial scattering of ions, reducing the sensitivity of MS measurements. Travelling wave IMS (TWIMS) provides non-linear mobility separation using a constantly changing electric field.[51] A stacked ring ion guide (SRIG) radially confines ions with an RF field, where adjacent rings have opposite phases; a transient DC pulse is superimposed across the pair of electrodes and sequentially propagates along the SRIG, generating a travelling wave to propel ions against the gas. This has been most successfully incorporated into the Synapt series of mass spectrometers from Waters (Manchester), shown in Figure 1.7.[54] The Synapt instruments have a Q-IM-ToF geometry which allows for mass isolation, ion mobility and mass analysis where the mobility domain information can be treated much like a chromatographic step. In this instrument the central region, known as the 'TriWave', is composed of three SRIGs: the trap, IMS and transfer cells. The trap and transfer SRIGs are under vacuum and can act as collision cells with an optional travelling wave applied across them, while the IMS SRIG is enclosed and filled with gas (N_2 or He). Due to the dynamic electric field, the direct relationship between K and Ω in Equation 1.24 no longer holds and drift times must be calibrated to literature values.[55]

Collision Induced Unfolding

Collisional activation of ions can effect partially unfolded conformations that are stable for milliseconds, these conformations can be separated by IMS.[56] Collision induced unfolding (CIU)[41,57,58] is an extension of this; ions are activated with a varying potential prior to IMS, allowing the resulting ATD to be plotted as a function of this potential in so-called 'CIU fingerprint' plots (a simulated example is shown in Figure 1.9). This technique is most useful when comparing the same protein with varying ligands[57,59–62] or modifications, such that the stabilisation effects of complexes can be determined, but may also be used for highly homologous proteins.[63,64]

1.1.4.3 Protein Footprinting

Protein footprinting is a set of techniques that sample structure and dynamics of proteins in solution and analyse these using mass spectrometry. The most prominent examples of these techniques are hydrogen–deuterium exchange (HDX), hydroxyl radical footprinting (HRF) and carbene footprinting. In comparison to standard biophysical techniques in frequent use such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM), MS-based methods are rapid and require small samples quantities.[65] Protein footprinting introduces a mass shift to exposed regions of proteins, typically this is followed by proteolytic digestion and liquid chromatography coupled MS (LC–MS) where the memory of this 'footprint' is retained as the extent of labelling. Conformational changes or binding will result in differential solvent exposure, and therefore propensity for labelling. Differential experiments, in the absence and presence of an interacting partner, are used to identify regions of proteins that show changes in the extent of labelling.

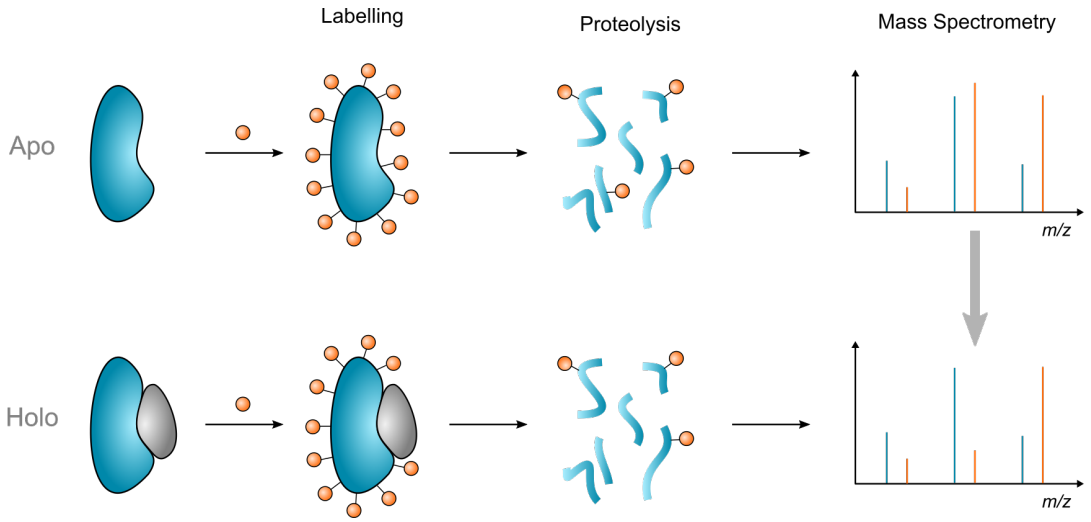


Figure 1.11 | Simplified schematic of a protein footprinting experiment. A protein (blue) is labelled with a footprinting reagent (orange), followed by proteolysis and mass spectrometry to give both unlabelled (blue) and labelled peptides (orange). This can also be performed on a '*holo*' protein, where a binding partner is added to the reaction and gives masking which is reflected in the abundance of labelled peptides.

HDX, one of the earliest MS footprinting techniques, probes the solvent exposure of a protein's surface through deuterium uptake.[66] This relatively non-specific technique provides high coverage of dynamics and binding interactions, however, the common LC–MS set ups are rife for back-exchange to hydrogen[67] and MS/MS experiments

can be obscured by H/D scrambling[68] during collisional activation. The mass shifts introduced by HDX are 1 Da for each exchange and are seen as an overall shift of the isotope distribution; but typically multiple exchanges occur for a given peptide, therefore the shift in overall distribution is observed and typically averaged. The exchange to deuterium does not change the mobility of peptides for LC–MS experiments and therefore all variants of exchanged peptides are observed at once.

Covalent Modifications

Covalent labels are much less labile than HDX, with the modifications retained well through processing steps. Due to the larger mass shifts from these labels, the modification of a peptide is typically expressed as fractional modification f_m for a singly labelled peptide,

$$f_m = \frac{PA_{mod}}{PA_{unmod} + PA_{mod}} \tag{1.25}$$

where PA_{unmod} and PA_{mod} are the chromatographic peak areas for the unmodified and modified peptides respectively. To obtain residue-level information, singly labelled peptides are selected for MS/MS and the fractional modification of fragment ions $f(y_i)$ determined as Equation 1.25. The absolute level of labelling for a given amino acid residue i is determined as the difference between adjacent fragment ions,

$$labels/residue = f_m[f(y_i) - f(y_{i-1})] \tag{1.26}$$

HRF provides a covalent modification in the form of oxidation of residues by the presence of a hydroxyl radical $^{\bullet}OH$, generated from X-ray irradiation of water.[69,70] Fast photochemical oxidation of proteins (FPOP) is a variant of HRF that provides microsecond exposure times for rapid sampling of native conformations, in this case the radicals are generated from hydrogen peroxide.[71,72] The reactivity of $^{\bullet}OH$ with amino acid side-chains varies over three orders of magnitude,[70] leaving many mostly aliphatic amino acids under-sampled. A similar method that produces a trifluoromethyl radical ($^{\bullet}CF_3$) has been reported to give excellent coverage of residues.[73]

Carbene footprinting is an alternative to the HRF approach, utilizing a highly reactive carbene intermediate to insert into any neighbouring X–H bond (where X can be C, N, O or S) in a protein.[74–76] The rate of insertion for carbenes is on the

order of nanoseconds,[77] providing even more rapid sampling of conformations than FPOP. Typically diazirine reagents are used to generate the carbene *in situ* with a near-ultraviolet laser. Diazirine gas[74] and photoleucine[76,78] were initially used but more recent studies employ trifluoromethylaryldiazirines due to their solubility, stability and efficacy.[79–81] The reactive carbene itself does not have significant preferences for residues, however the nature of the aryl group provides a high affinity for hydrophobic environments.

1.2 Polyketide Synthases

Polyketide synthases (PKS) are the enzymes responsible for the biosynthesis of polyketides from simple organic precursors through a series of Claisen condensations and reductive processing.[82–85] PKSs are megasynthases, closely related to fatty acid synthases (FAS) and non-ribosomal peptide synthases (NRPS).[86,87] Polyketides are a large family of natural products — including macrolides, polyphenols, enediynes, etc[88] — with diverse bioactivities ranging from antibiotics such as erythromycin[89] (**1**) to statins such as lovastatin[84] (**2**) and neurological treatments such as bryostatin[90] (**3**).

All PKSs rely on Claisen condensation for the elongation of polyketide chains (see Figure 1.12). This is catalysed by a ketosynthase enzyme (KS) with a thiotemplate containing an elongating unit.[91] In many PKSs, the thiotemplate is in the form of a phosphopantetheinyl (Ppant) modification to an acyl carrier protein (ACP). The ACP is loaded with the appropriate extender unit by an acyltransferase (AT) from an acyl-CoA (e.g. malonyl-CoA), forming a thioester bond. The KS activates this thiotemplate through decarboxylation, the resulting enolate undergoes Claisen-like condensation towards the intermediate polyketide that is bound as a thioester to a catalytic cysteine on the KS, resulting in a β-ketothioester intermediate attached to the ACP.

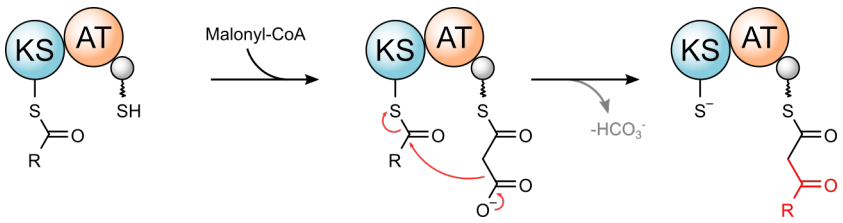


Figure 1.12 | Simplified scheme of polyketide elongation. A minimal PKS module featuring a ketosynthase (KS), acyltransferase (AT) and an acyl carrier protein (ACP, gray circle), the Ppant arm is shown as a wavy line. AT coordinates acyltransfer of malonyl-group from malonyl-CoA to the ACP (*left*). Claisen condensation mediated by decarboxylation of malonyl-ACP and attack of an acyl-KS liberating carbonate (*middle*). Elongated β-ketoacyl intermediate bound to ACP (*right*).

1.2.1 Classification and Architecture

Despite the great diversity of polyketides, their structures are predominantly derived from stepwise condensation of simple acyl-CoAs mediated by KS enzymes. PKSs are classified into three types (I, II and III) by domain and subunit organisation, shown in Figure 1.13.[84,92] Type I are found in bacteria and fungi, type II are found exclusively

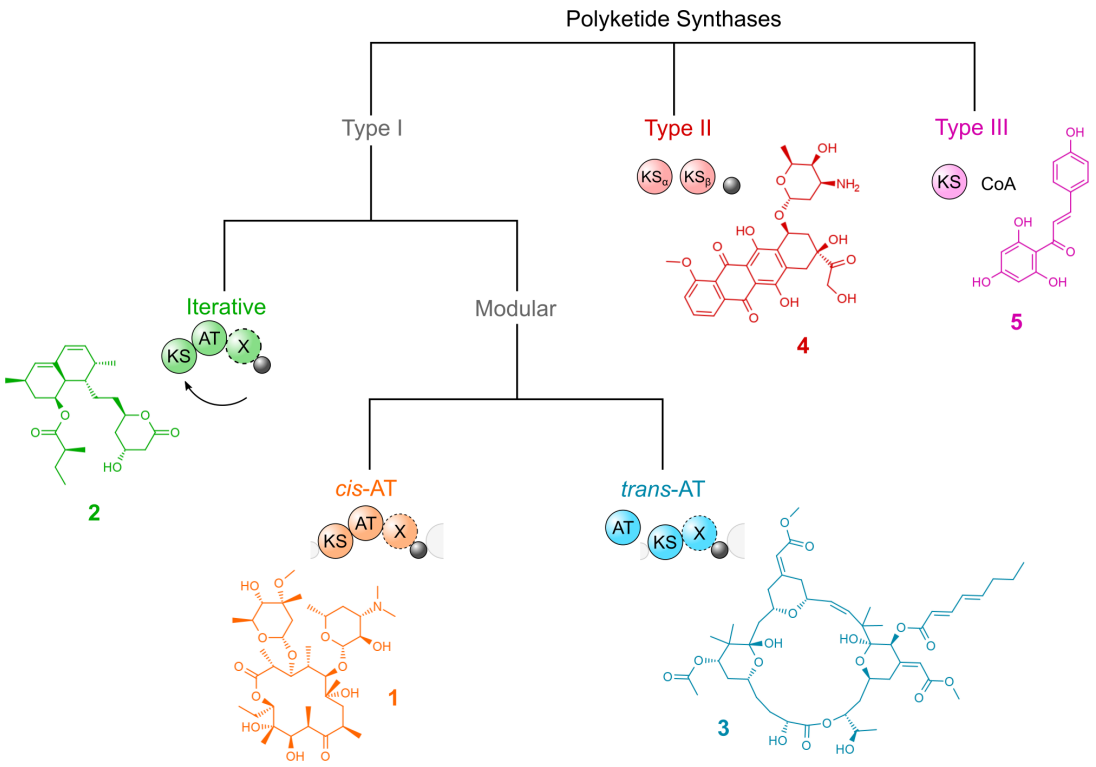


Figure 1.13 | Flowchart showing the classification of PKSs. Examplar modules are shown for each final branch along with a representative polyketide from that class of PKS – erythromycin A (**1**), lovastatin (**2**), bryostatin 1 (**3**), doxorubicin (**4**) and nargingenin chalcone (**5**). Codes in circles indicate domain: KS, ketosynthase; AT, acyltransferase; X, optional reducing domains. ACPs are shown as small dark grey circles.

in bacteria and type III are found predominantly in plants. The direct correspondence between PKS architecture and the resulting polyketide is known as the *principle of co-linearity*.[93] This principle not only allows for the structures of polyketides to be determined from gene mapping,[88,94] but also for genetic manipulations for the creation of novel polyketides.[95–97]

1.2.1.1 Type I PKSs

These PKSs are multi-domain enzymes organised into modules composed of KS, ACP, AT and optional reductive processing domains — ketoreductase (KR), dehydratase (DH) and enoylreductase (ER).[87] Type I PKSs are subdivided into architectures: *iterative*, which feature a single set of domains that iterate to build the polyketide chain and *modular*, large assemblies of independent modules. Iterative type I PKSs, much like the type II and III PKSs, typically generate aromatic polyketides.[88]

Modular Type-I PKSs

These modular PKSs are responsible for the most diverse polyketides, with a considerable number of modular PKSs also featuring at least one NRPS module which introduces a peptide bond into the polyketide intermediate. Modular PKSs offer the means to direct biosynthesis of polyketides by rational composition of modules.[98,99] The elongating polyketide chain traverses the PKS in an assembly line fashion, where each module typically contributes an elongation and some reductive processing. The variety in modules allows for the greater variety within the polyketide than is available from the iterative PKSs.[82,83] Due to the number of modules required for most polyketides, modular PKSs tend toward the MDa molecular weight range and are often comprised of two or more proteins containing several modules each. The delivery of extending units to the ACP is mediated by the AT, the location of this AT serves as a further subdivision for modular type I PKSs. The canonical modular type I PKSs features an in-module AT adjacent to the KS, these are termed *cis*-AT PKS; while their counterparts, *trans*-AT PKS, utilise free-standing ATs that interact with all modules.[87]

1.2.1.2 Type II PKS

Type II PKSs are the most prominent PKSs found within bacteria. The mechanism of polyketide production is iterative, however the domains for these PKSs are expressed from different but clustered genes. A 'minimal PKS', comprised of heterodimeric KS units termed KS_α and KS_β and an ACP. Optional KR, cyclase (CYC) and aromatases (ARO) dictate the reduced structure of the polyketide. The KS_α is the active condensing enzyme while the KS_β lacks the catalytic cysteine and acts to control the length of the polyketide chain and occasionally acts to generate an acetyl starter unit from malonyl-CoA.[100]

1.2.1.3 Type III PKS

Type III PKSs, unlike the other types, lack ACPs for tethering the growing polyketide chain and instead perform all reactions directly onto CoAs. Furthermore, these PKSs are multifunctional, with a single enzyme selecting the starter unit, elongating the chain and forming the final polyketide by cyclisation. Many starter units are accepted

by these enzymes and the final size of the polyketide is determined by the size of the active site.[88,101]

1.2.1.4 Non-Ribosomal Peptide Synthases

Non-ribosomal peptide synthases (NRPS) are modular megasynthases, much like the modular type I PKSs. Unlike the modular PKSs however, NRPSs use amino acids as the units for peptide synthesis. With more than 500 non-proteinogenic amino acids available in nature, NRPSs can produce peptides with much greater complexity than available to proteins.[102] Like PKSs, NRPSs utilise carrier protein tethered thioesters to carry out elongation, these peptidyl carrier proteins (PCP) are highly homologous to ACPs from PKSs and FASs. PCPs are loaded with aminoacyl-AMP mixed anhydrides by an adenylation (A) domain, this is achieved by activating amino acids with ATP allowing acyl transfer from the mixed anhydride. Elongation of a peptide chain is catalysed by a condensation (C) domain, where the amino group attacks the thioester of the previous module forming an amide bond.[103,104] Additional domains may be incorporated into modules for cyclisation (Cy), oxidation (Ox) and epimerisation (E). Like PKSs, the final module contains a thioesterase domain (TE) to hydrolise or cyclise the peptide and generate the mature non-ribosomal product.

1.2.2 *trans*-AT Polyketide Synthases

Initially thought to only account for a small fraction of modular PKSs, *trans*-AT PKSs are abundant. Indeed, a recent study found that *trans*-AT accounted for 25 % of non-redundant modular PKS and PKS-NRPS hybrid pathways.[105] In contrast to *cis*-AT PKSs, *trans*-AT PKSs exhibit highly aberrant architecture and do not follow the canonical principle of co-linearity.[87] This is due to a greater number of variant domains and, in many cases, transient interactions with processing enzymes. The diversity of *trans*-AT PKSs largely stems from the incredible number of module variations, more than 50 currently identified, compared to *cis*-AT PKSs which only have eight variants (KS-AT-ACP, KS-AT-KR-ACP, KS-AT-DH-KR-ACP, KS-AT-DH-ER-KR-ACP and their four counterparts with methyl transferase (MT) domains). The module variants of *trans*-AT PKSs are due to unusual domain orders, novel domains, repeated domains, modules split across two proteins, transient acting enzymes and apparently superfluous

Clade	Substrate	X–Cys Residue	Example
I	α/β-branch	A	OnnKS3
II	β-OH	A	PedKS9
III	KS^0	A/Q/M	BaeKS14
IV	β-OH	A	PedKS8
V	Single/double bond	A	BaeKS2
VI	Starter units	A	PedKS1
VII	Double bond	A	ChiKS15
VIII	β-OH	M	OnnKS6
IX	Double bond	M	MmpKS4
X	KS^0	A	BaeKS13
XI	Double bond after KS_{XIV}	V	BaeKS8
XII	Double bond	A	DszKS5
XIII	Double bond	A	DifKS1
XIV	β-OH, KS^0	A	BaeKS3
XV	Double bond in module without DH	A	DifKS10
XVI	Peptide	N	BaeKS1

Table 1.1 | KS clades identified by Nguyen *et al.*[106] showing the moiety introduced, most highly represented amino acid at position **X**–Cys[107] and representative KS for the clade. PKS cluster codifiers: Bae, bacillaene; Chi, chivosazol; Dif, difficidin; Dzs, disorazol; Mmp, mupirocin; Onn, onnamide; Ped, pederin. The clade of a KS may be denoted using subscript e.g. KS_{XVI}.

domains.[82,83] Examples of some non-canonical modules architectures found in *trans*-AT PKSs are shown in Figure 1.14.

1.2.2.1 Assigning *trans*-AT PKSs by KS Specificity

As the canonical co-linearity does not apply well to *trans*-AT PKSs, an alternative predictive model was proposed by Nguyen *et al.*[106], based on phylogenetic studies of *trans*-AT KS domains. In this model, *trans*-AT KSs that are evolutionarily related accept substrates with similar moieties in the α- to γ- region from the thioester, with 16 clades identified, see Table 1.1.

Performing phylogenetic analysis on unknown *trans*-AT PKSs with this model set allows the predominant KS clade, and therefore polyketide intermediate, to be determined. Correlation between module positioning and predicted clade allows for prediction of full polyketide products. This has been successfully applied for the characterisation of several polyketide synthases.[108,109] TransATor, a web application has been developed for predicting *trans*-AT polyketides on this basis.[110] The substrate specificity of *trans*-AT KSs can, in part, be rationalised with clade assignment on the basis of **X**–Cys residues — amino acid residues opposite the catalytic cysteine in the KS

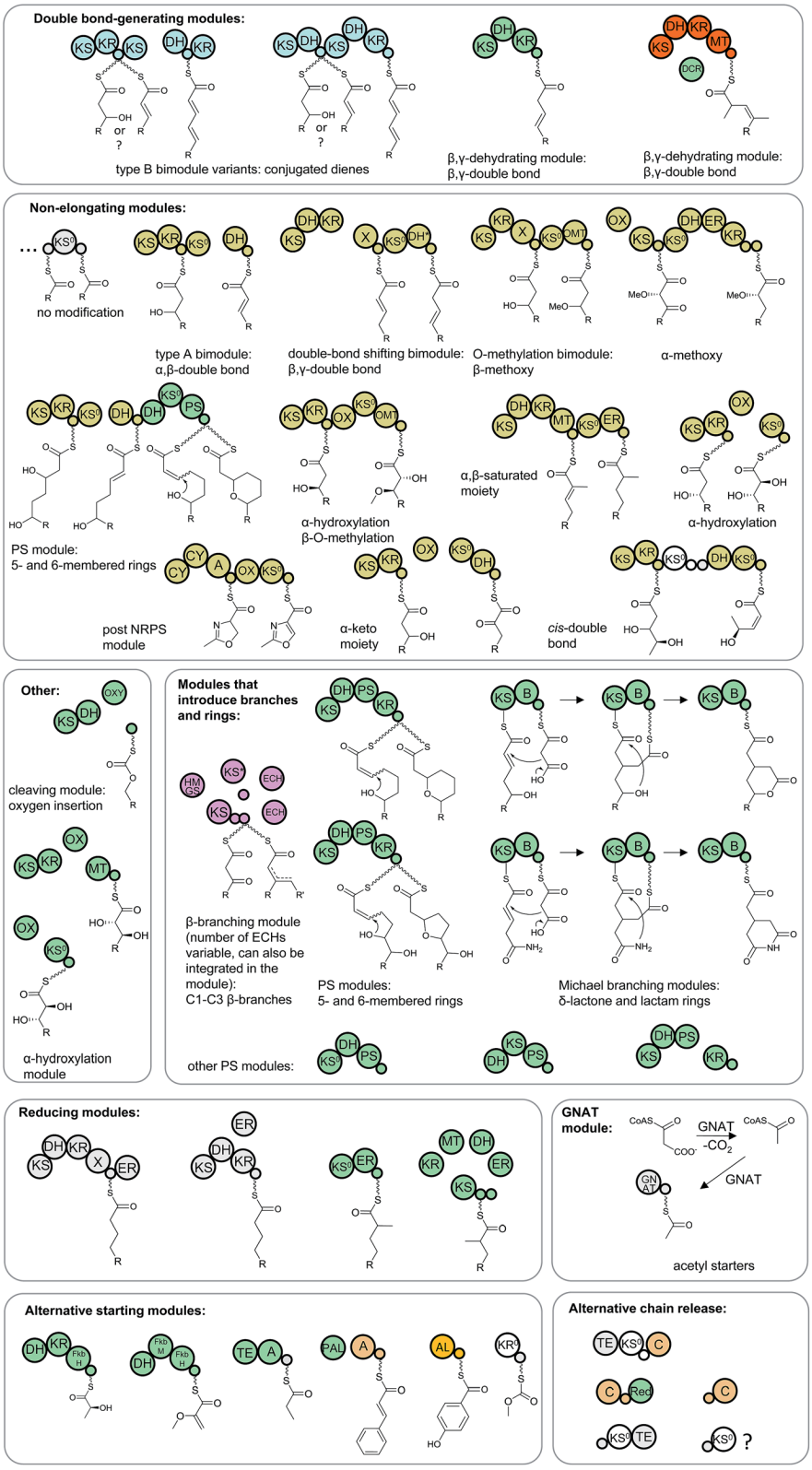


Figure 1.14 | Non-canonical modules from *trans*-AT PKSs and proposed functions. Polyketide products are shown attached to ACPs. Connected circles are in the same protein. KS^0, non-elongating KS; DH* dehydratase-like domain involved in double bond migration; OXY, oxygenase; PS, pyran synthase; PAL, phenylalanine ammonia lyase; AL, acyl-ligase, FkbH, hydroxylase; FkbM, methyltransferase; Red, reductase; ECH, enoyl-CoA hydratase; B, branching domain; OMT, O-methyltransferase. The following colour coding is used: grey, canonical PKS module architecture; orange, NRPS domains; white, putatively inactive domains lacking active site residues; pink, β-branching module; gold, bimodule involving a non-elongating KS^0; blue, type B bimodule; red, iterative module; green, other non-canonical module. Reproduced from reference 83.

active site.[107,111,112] Interestingly, *cis*-AT KSs are found to clade together based upon pathway and organism. This suggests evolution through duplication of whole modules within single phyla.[113,114] Whereas pathway-specific clades are not observed for *trans*-AT KSs and multiple phyla are found within single clades, suggesting permissive horizontal gene transfer and recombination.[106]

1.2.3 Structure and Mechanism

1.2.3.1 Acyl Carrier Protein

Acyl carrier proteins (ACP) are small ~80 residue tetra-helical bundles that facilitate the transfer of polyketide intermediates from the KS through various processing domains and ultimately onto the next module. ACPs utilise a thiotemplate mechanism, where the polyketide is covalently attached to the ACP via a 4'-phosphopantetheine (Ppant) prosthetic group. This Ppant arm is itself a flexible 18 Å post-translational modification to a conserved serine residue (DxGLD**S**L).[82–84,87] The flexibility and length of the Ppant allows for the polyketide intermediate to reach the active sites of processing domains within a module.[86,87]

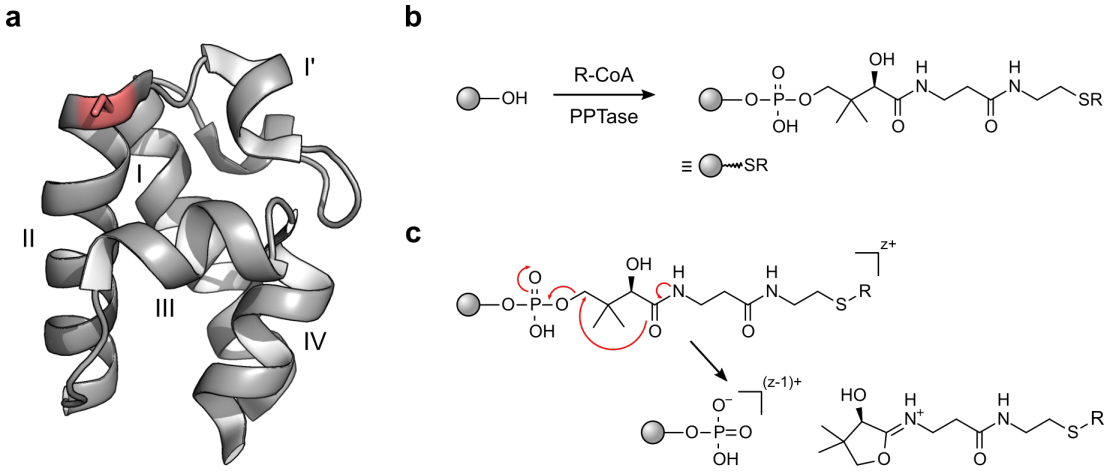


Figure 1.15 | (**A**) Rendering of Curacin ACP (PDB: 2LIU) with conserved serine highlighted (brown/red) and helices labelled. (**B**) Scheme illustrating Ppant loading by a PPTase and the abbreviated symbol for *holo*-ACP. (**C**) Scheme of Ppant ejection showing the two major product ions and proposed mechanism.[115]

Despite low sequence similarity, carrier proteins from several megasynthase families such as FAS, type II PKS and NRPS share the same fold. The conserved helices are numbered (I–IV), as shown in Figure 1.15A, helices I, II and IV are typically rigid with strong hydrophobic interactions between the helices, while the shorter

helix III is observed to be highly dynamic under NMR studies.[116,117] The proximity of this helix to the thioester bond in some structures[118] suggests that helix III might be involved in acyl-ACP interactions. The conserved serine is located on the N-terminal region of helix II, also known as the 'recognition helix' due to conserved negative charge which is believed to facilitate universal interaction with basic patches on KS and processing enzymes.[87,119,120] A peculiar feature of modular type I PKS ACPs is the existence of a conserved short hydrophobic helix at the N-terminal region, termed helix 0, which does not appear to contact the helical bundle in available structures.[87,117] The exclusive presence of helix 0 in modular PKSs suggests some role in downstream module docking. Module-swapping experiments for the erythromycin PKS cluster (DEBS) showed optimal polyketide transfer between ACPs and their native KS.[121]

When an '*apo*-ACP' is modified with a Ppant arm it is referred to as a '*holo*-ACP' and when a polyketide is bound is it described as an 'acyl-ACP'. The *in vivo* loading of a Ppant arm onto *apo*-ACP is catalysed by phosphopantetheinyl transferases (PPTase) using CoA as the Ppant donor (Figure 1.15B).[119,122] A particularly useful feature of the Ppant is the labile phoshoester bond, which can be cleaved during tandem mass spectrometry experiments to yield characteristic fragment ions (Figure 1.15C).[115] This technique termed 'Ppant ejection' allows the acyl group on an ACP to be determined with mass accuracy. Ppant ejection has been used to probe the state of acyl-ACPs during reactions with other PKS proteins, in particular KSs and GNATs.[97,112,123–126]

Solution NMR studies of ACP2 from the erythromycin PKS cluster (DEBS) have shown that the boundaries for the ACP domain in modular PKSs are ~40 residues on each side of the Ppant attachment site.[117] This study docked the ACP to its cognate KS to find that the recognition helix may indeed be involved in ACP–domain interactions. More recently an ACP involved in β-branching from the mupirocin PKS cluster was investigated for interactions with *trans*-acting hydroxymethylglutaryl-CoA synthase (HMGS) protein by NMR, and subsequent bioinformatic analysis, to identify a conserved motif for β-branching ACPs.[127] One of the key questions regarding ACPs is the nature of their interaction with acyl-Ppant arms. FAS ACPs such as the *E. coli* ACPP are known to have a hydrophobic pocket nestled between the three main helices[118] which provides protection from hydrolysis. PKS ACPs are not so clear-cut, however, chemical shift perturbation studies of erythromycin ACP2 show no differences between

holo and various lengths (C_3 to C_6) of PKS intermediates,[128] and similar studies performed on ACP9 from the highly reducing *cis*-AT PKS for mycolactone showed that non-polar chains (C_4 to C_8) docked into a groove in the surface but did not penetrate the helix bundle.[129] This landscape of acyl-Ppant protection demonstrates that all ACPs cannot be painted with the same brush, in fact each ACP may favourably protect its native substrates.

1.2.3.2 Ketosynthase

Ketosynthase (KS) domains are the work-horses of the PKS assembly line. These ~430-residue proteins are responsible for forming carbon–carbon bonds through Claisen condensation.[87] The KS is very highly conversed across PKSs and FASs, allowing the design of nucleotide probes for phylogenetic analysis and identification of PKS clusters.[106] KSs possess a thiolase fold, a set of alternating α-helices and β-sheets in a 2α/5β/2α/5β/2α arrangement (Figure 1.16).[91] The strong interface regions make KSs highly dimeric and contribute significantly to the dimeric nature of type I PKSs. Elongating KSs possess a Cys-His-His catalytic triad located in the TA**C**SSS, EA**H**GTG and KSNIG**H**T motifs.[87] This catalytic cysteine is found on the N-terminal elbow of an α-helix where the positive dipole of the helix serves to decrease the its pK_a by ~1.5 units,[85] making the cysteine a thiolate under physiological conditions thus promoting transthioesterification of the polyketide intermediate from an upstream ACP.

Catalysis at the KS is broken down into three steps (see Figure 1.16): 1. *Transthioesterification* of the acyl-group from the acyl-ACP of the preceding module by the cysteine thiolate, 2. *Decarboxylation* of malonyl-ACP from the current module mediated by proton abstraction of water by the EA**H**GTG histidine to generate an enolate, 3. *Elongation* of the enolate-ACP by Claisen condensation towards the acyl-KS thioester to yield a β-ketothioester on the ACP for further processing.

Non-canonical KS

PKSs can also employ catalytically inactive KSs for specific purposes within biosynthesis, the two most prominent of which are the loading KSs and the non-elongating KSs. Termed KS^Q, loading KSs are often found in the loading modules of *cis*-AT PKSs. In these variants, the catalytic cysteine is replaced with a glutamine

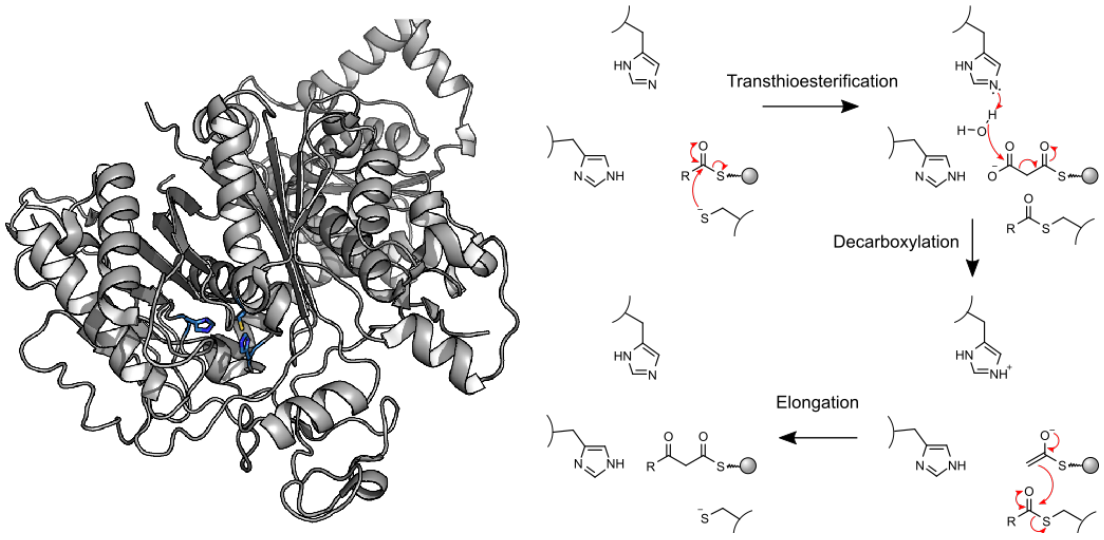


Figure 1.16 | Crystal structure rendering of KS (*left*, PDB: 4NA1) with catalytic triad side-chains shown as blue sticks. Catalytic mechanism for elongation (*right*). ACPs are shown as grey circles, electron flow arrows are shown in red.

residue, leaving only decarboxylative activity intact.[130] This variant can be simulated with active KSs by the alkylation of the cysteine with iodoacetamide.[131]

Perhaps more intriguing are the non-elongating KS, termed KS^0. These KS^0 lack the HGTG histidine necessary for decarboxylation and are believed to shuttle intermediates to downstream modules and processing domains without extending the polyketide chain, with several non-elongating module architectures known (see Figure 1.14). Notably, KS^0 domains occur frequently at the interface between docking PKS proteins.[82,87] KS^0 are spread across three defined clades (see Table 1.1): KS^0_{III} occur on the N-terminus of a PKS protein for docking to an upstream ACP; KS^0_X occur within modules that do not feature an active DH, typically a DH-like domain (DH^*) catalyses double-bond isomerisation;[83] and KS^0_{XIV} which typically occur as a split module, with an active DH at the N-terminus of an docking PKS protein.[83,106]

1.2.3.3 Acyl Transfer Domains

Acyltransferase (AT) domains are highly conserved across FASs and PKSs. The ~300 residue proteins consist of two domains, a catalytic ~240 residue hydrolase subdomain and a ~60 residue ferredoxin-like subdomain.[87] A ~140 residue 'KS/AT' adapter region C-terminal to the KS is present in all type I PKSs and potentially serves as a docking domain for *trans*-acting AT domains in *trans*-AT PKSs.[132,133] ATs catalyse the transfer of acyl groups from acyl-CoAs to the Ppant chain of *holo*-ACPs; most often malonyl-

or (2*S*)-methylmalonyl-CoA. Transfer occurs by a ping-pong mechanism similar to that of a serine protease via a Ser-His dyad, located in the GH**S**xG and (H/Y)A(F/S)**H** motifs,[134] whereby the CoA acyl group forms an ester intermediate with the catalytic serine before transfer to *holo*-ACP and release of acyl-ACP. Two key active site residues are known to affect the initial acylation step: a phenylalanine residue adjacent to the catalytic histidine (HAF**H**) selects for malonyl extender units through steric hindrance with the α-substituted group of the CoA,[134,135] and a conserved arginine in the active site that stabilises the malonyl carboxylate through a salt-bridge.[132,136]

Acylhydrolase

A recently discovered sub-clade of ATs found in *trans*-AT PKSs act as proofreading enzymes by hydrolysing stalled acyl-ACP intermediates, these AT are termed acyl hydrolases (AH).[137] AHs have a point mutation at the conserved carboxylate-stabilising arginine to glutamine, destabilising the malonyl-CoA substrate and switching activity to hydrolysis of acyl-ACPs. Site-directed mutagenesis (Arg→Gln) of a standard AT from the pederin PKS showed that this mutation was indeed responsible for removing canonical AT activity and hydrolysis of acyl-ACPs to *holo*-ACPs, however, no mechanism for acyl-AH hydrolysis was observed.[124] This study also suggested that AHs preferentially hydrolyse acetyl-ACPs, the most common form of stalled ACP.

Loading Domains

While ATs are involved in starter modules as well as extending modules, there are alternative loading modules for *trans*-AT PKSs. Among these, are GCN5-related *N*-acetyltransferase (GNAT), acyl ligase (AL), FkbH and others.[88] These domains are typically only present in the starter module of *trans*-AT PKSs and are responsible for the inclusion of non-standard units at the beginning of a polyketide.[82,83] Typically at the N-terminal region of *trans*-AT PKSs, GNATs are found to resemble GCN5-related *N*-acetyltransferase enzymes which acetylate histones and are thus hypothesised to incorporate acetyl starter units into the assembly line; with few exceptions, GNAT are found only within *trans*-AT PKSs. GNATs decarboxylate malonyl-CoA and transfer the resulting acetyl to the adjacent ACP thus performing the joint role of KS-AT.[123] X-ray crystallography of the GNAT from the curacin PKS showed that two pockets

are present in PKS GNATs, one pocket was shown to exclusively bind malonyl-CoA indicating that the other tunnel is the ACP Ppant insertion site.

1.2.3.4 Reducing Domains

Found in FASs and PKSs; ketoreductase (KR), dehydratase (DH) and enoylreductase (ER) domains serve to reduce ACP bound β-ketoacyl intermediates to β-hydroxyl through α,β-double bond to fully saturated intermediates respectively. While FASs use all three of these domains in series to produce fully saturated fatty acids, PKSs have a variety of configurations to produce the myriad of polyketides known.[87]

Ketoreductase

The ~430 residue KR domain utilises NADPH to stereoselectively reduce the β-keto group of the β-ketoacyl intermediate formed by Claisen condensation. The domain is comprised of a structural subdomain and a catalytic subdomain, both of which possess a conserved Rossmann-like fold.[138,139] The catalytic core of KRs feature a conserved tyrosine (in a **Y**AAAN motif) and serine to activate the β-keto group, stabilised by a neighbouring lysine. The pro-4*S* hydride of the NADPH attacks the carbonyl of the β-keto group and the oxygen deprotonates the tyrosine to generate the β-hydroxyl polyketide intermediate.[138] KRs may be classified as one of three type: A-type generate a L-hydroxyl, B-type a D-hydroxyl and C-type are reductase inert.[140]

Dehydratase

The DH domain is ~280 residues in length and dimeric across the twofold axis of the PKS. DHs catalyse the dehydration of β-hydroxyl intermediates by E1cB elimination to generate an α,β-unsaturated group. Dehydration is catalysed by a conserved aspartate residue and histidine in the HPALL**D** and **H**xxxGxxxxP motifs. The aspartate donates a proton to the β-hydroxyl group while the histidine abstracts an α-proton effecting the *syn* elimination of water and the α,β-double bond.[89] Some DH containing modules can also generate β,γ double bonds in module. Alternatively, α,β double bonds can by isomerised to β,γ by 'shift modules' that contain a DH-like domain, DH* or EI (enoyl-isomerase); such modules also contain a KS^0 domain.[141–143]

Enyolreductase

The ~310 residue domain is composed of a nucleotide-binding domain and a substrate-binding domain. ERs stereoselectively reduce *trans*-α,β-double bonds generated by DHs.[138] The reduction is hypothesised to involve the pro-4*R*-hydride of NADPH attacking the β-carbon to generate an enolate intermediate. The α-carbon of the enolate intermediate may then accept a proton from a catalytic tyrosine or lysine, for L- or D-α-substituent respectively.[144] ERs may be in-*cis* with the module, in which case they are inserted between the two subdomains of the KR, or they may act in-*trans* in many type I PKSs.[145]

1.2.3.5 Pederins

Pederins are a group of closely related polyketides featuring cytotoxic properties and a distinct methylene functional group. The members of this group to date are pederin,[146,147] onnamide,[148] mycalamides,[83] diaphorin,[149] nosperin[109] and psymberin[108] shown in Figure 1.17. All of these are believed to originate from unculturable bacterial endosymbionts. The host organisms of these endosymbionts include beetles, sea sponges, lichen and yeast; making this group of PKS one of the most diverse known.

Psymberin is the polyketide derived from the sea sponge *Psammocinia* aff. *bulbosa*. It was first discovered in 2004, afforded from a bioguided assay of extracts from the sponge, its potency and structural similarity to pederin prompted work into identifying the PKS/NRPS responsible for its synthesis.[150] This was later achieved in 2009 where the *psy* gene cluster was isolated and characterised with the polymerase chain reaction (PCR) methods described by Fisch *et al.*[108]. As shown in Figure 1.17, psymberin is produced by two PKSs: PsyA containing three modules and PsyD containing 10 modules, the latter being the largest single PKS protein to date. PsyA is initiated by a GNAT loading module and features a β-branching module with two ECH domains which generates the unusual branched β,γ' unsaturated moiety and the final module contains an uncharacterised structural domain and a KR; ACP3 of this module docks with $KS3^0$ of PsyD.

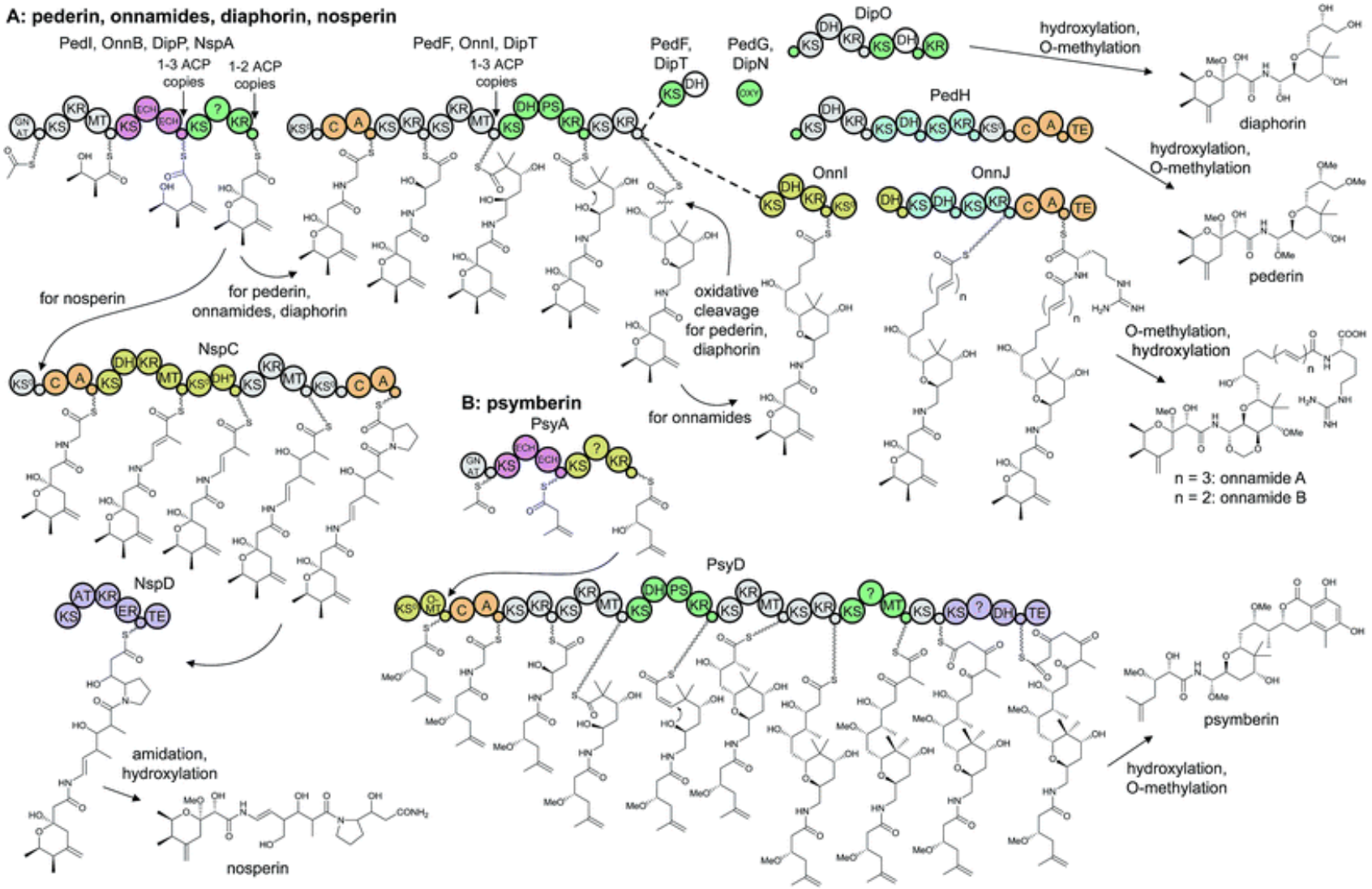


Figure 1.17 | Overview of the pederin-type PKS clusters and the mature polyketides. (**A**) Pederin, onnamide, diaphorin and nosperin PKS/NRPS. (**B**) Psymberin PKS/NRPS. Modules that introduce non-canonical functionality are coloured: β-branching (purple), ring formation (green), peptide (orange). Reproduced from reference 83.

1.2.4 *In Vitro* Study of Polyketide Synthases

In order to fully characterise PKSs it is convenient to produce recombinant PKS proteins. Due to the size of many of these *trans*-AT proteins, it is necessary to produce excised domains or, in some cases, whole modules.[151,152] Once recombinant proteins are available, there are several options for interrogating their interactions. One particularly useful development are simple *N*-acetylcysteamine (SNAC) thioesters, which mimic Ppant and can be used to readily produce a variety of acyl-groups for study where the acyl-CoAs are either unavailable or prohibitively expensive.[99] SNACs have been used to great success, especially for probing the substrate specificity of KSs.[111,112] While in many cases the removal of protein–protein interactions can simplify interpretation, the interactions of ACPs may contribute to the function of many PKSs. In such cases it is common place to prepare an acyl-ACP from a pool of acyl-CoA and a PPTase.[122] As the cost and availability of acyl-CoAs is limiting, researchers are turning towards the generation of *in situ* acyl-CoAs with readily synthesised acyl-pantetheines and a set of CoA synthesis enzymes.[99,153] The reactions of PKS domains and modules are typically monitored offline in one of several ways: (1) HPLC or radio-thin layer chromatography of released polyketide products,[154,155] requiring the presence of TE domains; (2) Ppant ejection of acyl-ACPs, intact single-domain proteins[112,115,156] or peptides from proteolysed modules;[97] or (3) trapping of ACP–enzyme interactions through reactive 'crypto'-pantetheine derivatives.[99,157–159]

1.3 Aims

Previous bioinformatic and experimental studies have shown that KSs from *trans*-AT PKSs are highly substrate specific and may act as 'gate-keepers' for polyketide biosynthesis. Until recently, ACPs from *trans*-AT PKSs have been seen as passive carriers for the growing polyketide due to their size and homology. In this thesis, the role of ACPs has been explored through a combination of MS studies – both denatured and native – and computational techniques. Software tools have been developed to allow protein footprinting to be utilised for interrogating protein–protein interactions, paving the way for in-depth and high-throughput analysis of PKS systems.

2 Materials and Methods

2.1 Buffers and Reagents

All buffers and solutions were prepared with **ultrapure water (mqH_2O)** (18.2 MΩ/cm), obtained from a **Milli-Q Integral Water Purification System** (Millipore). **Coenzyme A (>93 %, CoA)** and variants (**acetyl**, **butyryl**, **octanoyl** and **malonyl**, all >90 %) were purchased from Sigma-Aldrich and prepared as 20 mM stocks in mqH_2O before storage at –20 °C.

2.1.1 Biological Buffers

Reagents used for preparation of standard biological buffers are shown in Table 2.1. Biological buffers were adjusted to the desired pH with HCl, monitored with an **8100 pH meter** (ETI). Kan, Car and IPTG were prepared as filter sterilised 1000-fold stocks in mqH_2O. For the preparation of SDS-PAGE gels APS was prepared as a fresh 10 % stock, while TEMED and ProtoGel were added directly. SDS-PAGE gels were stained with **GelCode Blue Safe protein stain** (ThermoFisher). Typical biological buffers used were:

IMAC Binding buffer 20 mM Tris-HCl, 300 mM NaCl, 30 mM imidazole, pH 7.6

IMAC Eluting buffer 20 mM Tris-HCl, 50–300 mM NaCl, 300 mM imidazole, pH 7.6

IMAC Stripping buffer 20 mM Tris-HCl, 500 mM NaCl, 50 mM EDTA, pH 7.6

SAX Low buffer 20 mM Tris-HCl, 5 % glycerol, pH 8.0

Reagent	Abbr.	Source
Trizma base	Tris	Sigma-Aldrich
Sodium chloride	NaCl	Sigma-Aldrich
Imidazole		Sigma-Aldrich
Ethylenediaminetetraacetic acid	EDTA	Sigma-Aldrich
Sodium dodecyl sulfate	SDS	Sigma-Aldrich
Dithiothreitol	DTT	Sigma-Aldrich
Glycine		Sigma-Aldrich
Glycerol		Sigma-Aldrich
Kanamycin	Kan	Sigma-Aldrich
Carbenicillin	Car	Sigma-Aldrich
Isopropylthio-β-galactoside	IPTG	Sigma-Aldrich
LB Agar (Miller)		Sigma-Aldrich
LB media (Miller)		Sigma-Aldrich
2xYT media		Sigma-Aldrich
HCl		Fisher Scientific
Ammonium persulfate	APS	Sigma-Aldrich
Tetramethylethylenediamine	TEMED	Sigma-Aldrich
ProtoGel (30 %, 37.5:1)		National Diagnostics

Table 2.1 | Reagents for preparation of biological buffers. Abbreviations as used elsewhere are given. All regents were of the highest grade provided by the source.

SAX High buffer 20 mM Tris-HCl, 1 M NaCl, 5 % glycerol, pH 8.0

Storage buffer 20 mM Tris-HCl, 100 mM NaCl, 10 % glycerol, pH 7.6

TAE buffer 40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.3

SDS Running buffer 25 mM Tris-HCl, 192 mM glycine, 30 % SDS, pH 7.6

SDS 15 % Resolving solution 375 mM Tris-HCl, 0.1 % SDS, 50 % ProtoGel (30 %), 0.04 % APS, 0.1 % TEMED, pH 8.8

SDS 3 % Stacking solution 250 mM Tris-HCl, 0.1 % SDS, 10 % ProtoGel (30 %), 0.1 % APS, 0.25 % TEMED, pH 6.8

2.1.2 Mass Spectrometry Solutions

Reagents and solvents used for mass spectrometry are given in Table 2.2. A solution of **sodium iodide (2 µg/µL)** and **caesium iodide (50 ng/µL)** in aqueous isopropanol (50 %) (Waters) was used for mass calibration of all ToF instruments. Myoglobin, ubiquitin and alcohol dehydrogenase were used for calibration of ion mobility conditions as 5 µM solutions in 10 mM AmAc or 50 % MeCN, 0.1 % FA. Solutions used for ZipTip desalting (Section 2.3.2) and LC–MS (Section 2.3.8.1) were:

Reagents	Abbr.	Grade	Source
Acetonitrile	MeCN	HPLC	Fisher Scientific
Methanol	MeOH	Analytical	Fisher Scientific
Trifluoroacetic acid	TFA	Analytical	Sigma-Aldrich
Formic acid	FA	Analytical	Fluka
Ammonium acetate	AmAc	Reagent	Fisher Scientific
Ammonium bicarbonate	AmBic	Reagent	Sigma-Aldrich
Iodoacetamide			Sigma-Aldrich
Myoglobin (equine)			Sigma-Aldrich
Ubiquitin (bovine)			Sigma-Aldrich
Alcohol dehydrogenase (yeast)			Sigma-Aldrich
Trypsin		Mass Spectrometry	Promega
AspN		Mass Spectrometry	Promega

Table 2.2 | Reagents for preparation of mass spectrometry solutions. Abbreviations as used elsewhere are given.

Equilibration solution 5 % MeOH, 0.1 % FA

Elution solution 80 % MeCN, 0.1 % FA

Mobile phase A 5 % MeCN, 0.1 % FA

Mobile phase B 95 % MeCN, 0.1 % FA

2.2 Cloning and Expression

All polymerase chain reactions (PCR) were performed using a **MultiGene OptiMax thermocycler** (LabNet). Small volume (<2 mL) centrifugation was performed on a **Microfuge 16 centrifuge** (Beckman), cell culture centrifugation was performed on a **Avanti J-26 XP centrifuge** (Beckman Coulter) and protein clarification was performed on a **Avanti J-26S XPI centrifuge** (Beckman Coulter). Cell lysis was performed using a **Soniprep 150 ultrasonic disintegrator** (MSE) with a 9.5 mm probe. Protein and DNA quantification was performed using a **NanoDrop 2000 spectrophotometer** (Thermo Scientific). SDS-PAGE was performed using a **Mini-Protean Tetra Cell gel system** (Bio-Rad) with images taken using a **G:Box gel imager** (Syngene). Antibiotics Kan and Car were used at final concentrations of 30 and 50 µg/mL respectively.

2.2.1 Expression Vectors

All recombinant proteins were expressed with a His-tag for purification with immobilised metal affinity chromatography (IMAC). The details of their expression vectors are given in Table 2.3.

Protein	Location	Vector	Tag	M_r	pI	ε
PsyACP1	624–717	pET-28	N-His_6	12716	7.2	5500
PsyACP2	1784–1865	pET-28	N-His_6	11642	6.5	8480
PsyACP3	3221–3297	pET-28	N-His_6	10728	5.4	8480
PsyACP4	945–1028	pET-28	N-His_6	11283	5.9	0
PsyGNAT	412–622	pET-28	N-His_6	25696	6.5	26930
PsyGNAT-ACP1	412–717	pET-28	N-His_6	36201	6.5	32430
PsyAR-GNAT-ACP1	1–717	pET-28	N-His_6	81407	6.5	107830
EcACP		pET-29	C-His_6	13104	4.6	1490
PedACP4		pET-28	N-His_6	14124	9.8	6990
PedC		pET-24	C-His_6	39084	6.1	42400
PedD		pHis8	N-His_8	42929	5.7	37360
$PedD^{R97Q}$		pHis8	N-His_8	42900	5.6	37360
PsyKS1		pHis8	N-His_8	65016	6.2	66350
PsyACP-KS1		pET-28	N-His_6	75774	6.2	71850
Svp		pQE-70	C-His_6	26685	5.5	34490

Table 2.3 | Details for the recombinant proteins used. M_r, molecular weight; pI, isoelectric point; and ε, molar extinction coefficient were calculated using ExPASy ProtParam (https://web.expasy.org/protparam/). Sequences for these proteins are found in Section A.1. The preparation of constructs highlighted in grey are described herein, for these proteins the corresponding locations in the full-length protein (PsyA or PsyD) are given.

The plasmid for a C-terminally His-tagged *E. coli* FAS ACP (EcACP, Addgene plasmid #75016)[157] was a gift from Prof. Michael Burkart (UCSD). The plasmid for PedACP4 was a gift from Dr Matthew Jenner (University of Warwick). Plasmids for PedC (pederin cluster AH) and PedD (pederin AT), as well as PsyKS1 (psymberin KS1) and Svp (PPTase from *Streptomyces verticillus*)[160] were provided by Prof. Jörn Piel (ETH, Zurich). The pPSKF1[108] fosmid containing the cDNA for PsyA (ADA82581) and the N-terminal fraction of PsyD (ADA82585) was provided by Dr Anna Vagstad (Piel Lab, ETH Zurich). The mutant $PedD^{R97Q}$ plasmid and the PsyACP-KS1 didomain plasmid were prepared by Dr José Afonso and Dr Matthew Jenner. All remaining constructs were amplified from pPSKF1 and cloned into **pET-28b** (Novagen) as described in Section 2.2.2. Primers were designed to occur between annotated domain boundaries in regions with no secondary structure, as predicted by Phyre2[161] for the full protein sequence. They were flanked with additional DNA (~17 bp) homologous to the pET-28

vector such that the PKS DNA was inserted between the *Nde*I and *Not*I restriction sites (Figure 2.1), with the exception of PsyACP4, which was inserted before the *Nde*I and after the *Xho*I sites. Reverse primers included an ochre stop codon (TAA) after the PKS gene to remove the C-terminal His-tag from the recombinant protein. These constructs all possess a thrombin-cleavable N-His_6-tag.

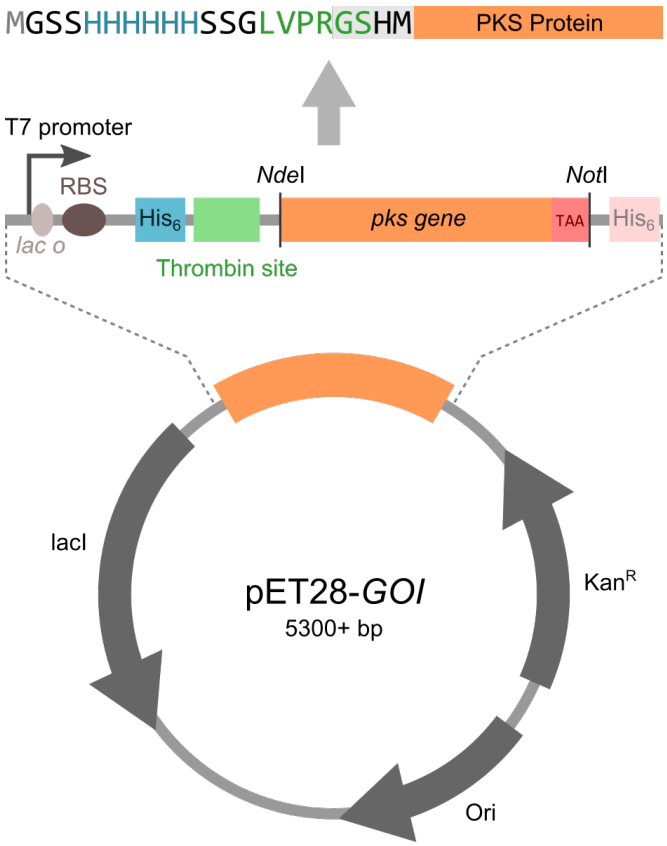


Figure 2.1 | Plasmid map for cloning into pET-28. A PKS gene of interest (*pks gene*) is typically inserted directly between the *Nde*I and *Not*I restriction sites with an ochre stop codon (TAA) immediately before the *Not*I. Expression is under the control of the T7 promoter and the *lac* operon (*lac o*), producing the expected recombinant protein (*top*). RBS, ribosomal binding site; His_6, His-tag; Kan^R, kanamycin resistance cassette; Ori, origin of replication.

2.2.2 Construct Amplification

Genes of interest were amplified by PCR from plasmid or fosmid DNA with either **Phusion** or **Q5 High Fidelity polymerases** (New England Biolabs) on a 50 µL scale according to the manufacturer's protocol. Briefly, pPSKF1 fosmid DNA (50 ng) was mixed with polymerase (1 U), dNTPs (200 µM), forward (For) and reverse (Rev) primers (see Table 2.4, 0.5 µM each), and optional dimethylsulfoxide (DMSO, <3 % v/v) on

ice. The PCRs were then performed in a preheated thermocycler using the program described in Table 2.5.

Protein	→	Primer (5'→3')	T_m
PsyACP1	For	cgcgcggcagccatatgCGTCAGCGGCGAGATGG	68
	†Rev	tgctcgagtgcggccgcttaCTTCGTCGACGGAGCCAAA	
PsyACP2	For	cgcgcggcagccatatgGATGTCTCATCTGCCTTGTATAAGC	66
	Rev	tgctcgagtgcggccgcttaCTTACGGACCGATTCTTCCTG	
PsyACP3	For	cgcgcggcagccatatgACGTCGAGCGGGGAACTTG	71
	Rev	tgctcgagtgcggccgcttaAACGCATACCGCTTCGAGCTG	
PsyACP4	For	gcggcctggtgccgcgcggcagcACTTCGTCGCCAAAGGG	67
	Rev	gtggtggtggtggtgatgttaAGGTATACACGCTTCGATGTGG	
PsyGNAT	‡For	cgcgcggcagccatatgGATGAAACAGGCGCCTGC	69
	Rev	tgctcgagtgcggccgcttaCGCCAGGCCATAGTGGAC	
PsyGNAT-ACP1	‡For	cgcgcggcagccatatgGATGAAACAGGCGCCTGC	68
	†Rev	tgctcgagtgcggccgcttaCTTCGTCGACGGAGCCAAA	
PsyAR-GNAT-ACP1	For	cgcgcggcagccatatgATGCTGAACATCCCTTTTAGCCAT	68
	†Rev	tgctcgagtgcggccgcttaCTTCGTCGACGGAGCCAAA	

Table 2.4 | Primers used for cloning of *psy* genes. Plasmid overlapping regions in lowercase with restriction sites underlined. Equivalent primers are marked with a † or ‡. T_m are given for the primer pair.

Step	Temp (°C)	Time (s)
Initial Denaturation	98	180
25–30×	98	10
	T_m	30
	72	15/kb
Final Extension	72	120

Table 2.5 | Base PCR program for amplifying PKS genes of interest. T_m is the melting temperature of the primer pair as indicated in Table 2.4.

Successful amplification was confirmed by TAE-agarose (1 %) gel electrophoresis. PCR products were then purified with the **GeneJet PCR Clean-Up kit** (Thermo Scientific) into mqH_2O. The vector backbone was linearised by restriction digest (with *Nde*I and *Not*I) or PCR (with primers p28LIC_N and p28LIC_C, Table 2.6) followed by purification with the GeneJet PCR Clean-Up kit. Amplified genes were assembled with linearised vector (3:1 molar ratio) using either the **InFusion HD Cloning kit** (Clontech) or one-step sequence and ligation independent cloning (SLIC),[162] described in Section 2.2.3.

Name	→	Primer (5'→3')
p28LIC_N	Rev	GCCGCGCGGCACCAGGCCG
p28LIC_C	For	CACCACCACCACCACTGAGATCCGGC
T7P	For	TAATACGACTCACTATAGGG
T7T	Rev	TATGCTAGTTATTGCTCAGCGG

Table 2.6 | Primers used for general purposes. p28LIC primers are for linearisation of pET-28 plasmids, PCR performed with Q5 polymerase with $T_m = 72$ (Table 2.5). T7 primers are for amplification of DNA between T7 promoter and T7 terminator regions.

2.2.3 Sequence and Ligation Independent Cloning

Typically, a purified gene of interest was combined with linearised vector in a 3:1 molar ratio and made up to 8.3 µL with mqH_2O, to this **NEB 2.1 buffer** (1 µL), **DTT** (0.5 µL, 100 mM), and **T4 DNA Polymerase** (0.2 µL, New England Biolabs) were added and the mixture incubated at room temperature for 3 min followed by 10 min on ice. XL-1 Blue or TOP10F' chemically competent cells (50 µL) were transformed (Section 2.2.4) with 5 µL of reaction mixture and spread onto antibiotic-containing LB agar plates. Presence of gene of interest was verified by colony PCR (with primers T7P and T7T, Table 2.6) followed by overnight cultures of positive colonies and purification of DNA with the **GeneJet Plasmid Extraction kit** (Thermo Scientific). Purified DNA was confirmed by sequencing with T7P and T7T primers.

2.2.4 Transformation

Chemically competent cells were prepared with the calcium chloride method.[163] XL-1 Blue or TOP10F' were used for storage and sequencing of DNA while BL21 (DE3) were used for expression of recombinant protein. For a typical transformation, 1 µL of purified DNA was added to 50 µL of competent cells and incubated on ice for 30 min, the cells were then heat shocked at 42 °C for 30 s and placed on ice for 2 min, 200 µL of SOC media (New England Biolabs) was added and cells recovered for 1 h (37 °C, 200 rpm), 50 µL were then spread onto LB agar containing selective antibiotic and incubated overnight at 37 °C. Glycerol stocks of recombinant strains were prepared from a 5 mL overnight culture into 500 µL of LB+25 % glycerol.

2.2.5 Expression

Strains of BL21 (DE3) containing recombinant plasmid were grown at 37 °C in overnight starter cultures containing selective antibiotic from glycerol stocks. A 1:200 dilution of starter culture was used to inoculate 1 L of LB or 600 mL of 2xYT media containing selective antibiotic. Cultures were grown in an orbital shaker at 37 °C and 200 rpm until an OD_{595} of 0.6–0.8 was reached (~3 h), the cultures were then induced with 500 µM IPTG and further incubated for 18 h at 16 °C. Cells were harvested by centrifugation (3500 *g*, 10 °C) for 15 min. The resulting cell pellets (3–8 g) were stored at −80 °C until needed.

Recombinant proteins PedC, PedD, $PedD^{R97Q}$, PsyACP-KS1 and Svp were expressed and purified by Dr José Afonso and Dr Matthew Jenner. PksM3 (bacillaene PKS module 3 from *Bacillus subtilis*) was expressed and purified by Dr Matthew Jenner.

2.2.6 Purification

Cell pellets were thawed on ice and resuspended in 10 mL of IMAC binding buffer (20 mM Tris-HCl, 300 mM NaCl, 30 mM imidazole, pH 7.6). The cell suspension was sonicated 8–15×30 s on ice and clarified by centrifugation (35 000 *g*, 4 °C) for 30 min. The clarified supernatant was then passed through a **0.22 µm sterile filter** (Sartorius) prior to column purification. This supernatant was loaded onto a nickel affinity chromatography column — **HisTrap HP 1 mL** (GE Healthcare) operated by syringe or **HisGraviTrap 1 mL** (GE Healthcare) operated by gravity — pre-equilibrated with binding buffer and contaminating proteins removed by washing with 10 mL binding buffer. The column was then eluted stepwise with binding buffer containing increasing concentrations of imidazole: 3 mL each of 50, 100, 200 and 300 mM, collecting 1 mL fractions. A 15 % SDS-PAGE gel was performed to confirm the presence of the protein of interest.

Additional strong anion exchange (SAX) purification was performed to improve purity when necessary. Briefly, samples were diluted ten-fold into SAX low salt buffer (20 mM Tris-HCl, 5 % glycerol, pH 8.0) before loading onto a **HiTrap Q 1 mL column** (GE Healthcare). The column was then washed with 5 mL low salt buffer and eluted

stepwise with 3 mL each of buffer containing 50, 100, 200, 300, 400, 500 and 1000 mM NaCl, collecting 1 mL fractions.

Fractions containing protein of interest were pooled and exchanged into storage buffer (20 mM Tris-HCl, 100 mM NaCl, 10 % glycerol, pH 7.6) using either a **HiTrap Desalting column 5 mL** (GE Healthcare) or a **Vivaspin centrifugal concentrator** (Sartorius) with the appropriate MWCO (5 kDa for ACPs and 10 kDa for other proteins). Protein samples were split into aliquots (50–100 µL) and flash-frozen with liquid N_2 before storage at −80 °C. Proteins were quantified using absorbance at 280 nm with a NanoDrop 2000 for proteins with a sufficient molar extinction coefficient as shown in Table 2.3, otherwise a **Pierce BCA Protein Assay kit** (Thermo) was used according to the manufacturer's micro-assay instructions with a NanoDrop 2000.

2.3 Mass Spectrometry

Denatured MS measurements were performed on a variety of QToF instruments, depending on availability — typically on a **QToF-2** (Micromass) or a **Synapt HDMS** (Waters). All nMS measurements were performed on either a Synapt HDMS or on a high-mass modified QToF-2 (**HM-QToF**, Micromass) with a 32 k quadrupole rather than the standard 4 k. All IMS–MS experiments were performed on a Synapt HDMS. These instruments were operated with the **MassLynx** (v4.0 and v4.1, Waters) data system with a standard z-spray source and a **100 µL gas-tight glass syringe** (Hamilton) or a nESI source with emitter tips prepared from thin-wall borosilicate capillaries as in Section 2.3.1. Dry nitrogen (N_2) gas was used in all instruments for desolvation, nanoflow and ion mobility carrier gas. Dry argon (Ar) was used for collision gas in Synapt and QToF-2 instruments while sulfur hexafluoride (SF_6) was used in the HM-QToF instrument. All spectra were acquired in positive ion mode with minimal smoothing and background subtraction applied. All LC–MS measurements were performed on a **LTQ FT Ultra** (Thermo Scientific) coupled to an **UltiMate 3000 nano-LC** (Dionex). These instruments were operated with **Xcalibur** (v2.0, Thermo) and **Chromeleon** (v6.80, Dionex) respectively. Protein digests were cleaned on a **PepMap300 C_{18} trap column** (Thermo) before separation on a **PepMap300 C_{18} nanocolumn** (Thermo) connected directly to a **PicoTip** (New Objective). Reactions

Heat	Velocity	Pull	Time
RAMP+15	—	15	80
RAMP+15†	—	15	80†
RAMP+15	100†	25†	80

Table 2.7 | Parameters for pulling nESI emitter tips on Flaming/Brown P-97 micropipette puller with a 4.5 × 4.5 mm box filament. The RAMP value was determined with the instrument for both capillaries, giving 715 and 725, for 0.80 mm i.d. tips without filament and 0.78 mm i.d. tips with filament respectively. Values highlighted with a † may be changed to slightly alter the taper and final diameter of the tip.

were temperature controlled in a **myBlock Mini Dry Bath** (Benchmark Scientific). Centrifugation was performed in a **5417R refrigerated centrifuge** (Eppendorf).

2.3.1 Preparing nESI Emitter Tips

Emitter tips for nESI were prepared from borosilicate capillaries using a **Flaming/Brown P-97 micropipette puller** (Sutter Instruments) fitted with a 4.5 × 4.5 mm box filament. Tips were pulled to have an approximate final inner diameter of 0.5 mm, so that a spray could be initiated with minimal 'tapping off'. Two dimensions of capillary were used: (**1**) 0.78 mm inner diameter (i.d.) × 1.00 mm outer diameter (o.d.) × 10 cm long with filament (**G100TF-4**, Warner Instruments) and (**2**) 0.80 mm i.d. × 1.00 mm o.d. × 10 cm long (**920-10-10**, Hirschmann), both made from borosilicate 3.3 glass. Filamented capillaries can improve 'wicking' of viscous sample solutions towards the tip, but are more expensive. Typical conditions for pulling tips to the desired shape are shown in Table 2.7. A **platinum wire** (99.9 %, 0.127 mm diameter, Sigma-Aldrich) was back-fitted into the nESI tip prior to sample loading — after, in the case of filamented capillaries — such that the end of the wire was approximately 10 mm away from the tip orifice. Samples were loaded into the emitter tips with **GELoader tips** (Eppendorf).

2.3.2 Desalting for Denatured MS

Protein samples were desalted using **C_{18} or C_4 ZipTips** (Millipore). Briefly, a 10 µL protein sample was spiked with 1 µL of 1 % TFA or FA. A ZipTip was wetted with 3 × 10 µL of *Elution solution* (as defined in Section 2.1.2) then equilibrated with 5 × 10 µL of *Equilibration solution*. The sample was then be loaded onto the ZipTip with 15 × 10 µL

aspirations before washing with 8–20 × 10 µL of *Equilibration solution*, and finally recovered with 5 × 8 µL aspirations into *Elution solution*. Eluate was then typically analysed by ESI-MS as described in Section 2.3.3.

2.3.3 Acquiring Denatured Mass Spectra

After desalting and denaturing (Section 2.3.2), protein samples were injected inline to a syringe flow of *Elution solution* at 8 µL/min. The flow was connected to the ESI capillary, which was held at 2.5 kV. Mass spectra were collected over approximately 2 min in full scan or mass isolated modes. For Ppant ejection of ACPs, ions were selected using the quadrupole at a medium resolution to allow all acyl variants to be selected.

2.3.4 Buffer Exchange for nMS

Standard biological buffers (e.g. Tris, NaCl) are not compatible with MS experiments, therefore proteinaceous samples must be exchanged into MS-compatible buffers/solutions, such as ammonium acetate (AmAc). For all nMS experiments, samples were exchanged into AmAc (25 mM or 200 mM, pH 6.9) using **Zeba spin desalting columns (7 k MWCO, 75 µL or 0.5 mL bed volume)** (Thermo Scientific) following the manufacturer's protocol. Briefly, columns were equilibrated with 4 × 50 µL of AmAc, followed by centrifugation at 1000 *g* for 1 min each. Samples (5–12 µL) were then loaded and centrifuged for 2 min and collected, with minimal dilution or sample loss. For 0.5 mL columns these values are 300 µL, 1500 *g* and 50–130 µL respectively.

2.3.5 Acquiring Native Mass Spectra

After exchange into a suitable solution (Section 2.3.4), protein samples were typically diluted to 10 µM in the same solution. Samples sprayed using a conventional ESI source were supplied via a syringe, with flow at 5 µL/min. The capillary voltage was operated at 2.8 kV for ESI experiments and 1.5 kV for nESI experiments. Sampling cone voltage, collision voltage and quadrupole profiles were optimised on a per-protein basis, see Table 2.8 for typical values.

	ACP	KS	ACP-KS	Module
Sampling Cone (V)	10	60	60	80
Collision (V)	5	10	20	40
$Quad_{m1}$	1000	2000	2000	2500
$Dwell_{m1}$ (%)	15	5	5	10
$Ramp_{m1\rightarrow m2}$ (%)	20	15	15	20
$Quad_{m2}$	2000	4000	5000	6000
$Dwell_{m2}$ (%)	40	50	50	50
$Ramp_{m2\rightarrow m3}$ (%)	20	30	30	20
$Quad_{m3}$	3000	8000	8000	12000†

Table 2.8 | Typical voltages and quad profiles for nMS of PKS proteins. The Collision voltage is the value used for better spectra, for CIU experiments this is lowered. Dwell and Ramp are expressed as the percentage of scan time used for these steps. † Limited to 8000 on Synapt HDMS.

2.3.6 Collision-Induced Unfolding of Proteins

Collision induced unfolding (CIU) experiments were performed on a Synapt HDMS instrument. CIU stability experiments on *holo*/acyl-PsyACP1 were performed with a conventional ESI source while experiments on PsyKS1 and PsyACP-KS1 were performed by nESI. Optimal instrument settings were different for the two sets of experiments and are outlined in Table 2.9. For experiments comparing PsyACP1 to the other proteins, PsyACP1 was sprayed by nESI under identical parameters. For CIU experiments investigating the effect of the Ppant chain, PsyACP1 was treated with **thrombin** (Sigma-Aldrich) to remove the flexible His-tag that might obscure unfolding. ATDs were extracted, either manually in MassLynx or in batch with TWIMExtract,[164] to .csv files. The CIU data were plotted using either CIUSuite2[165,166] or custom Python scripts.

2.3.6.1 Quantifying Stabilisation in CIU

A weighted average arrival time, or centroid time, was calculated for each collision energy and then fitted with a four-parameter logistic sigmoid curve, as described by Equation 2.1, where x is the collision energy, x_0 is the midpoint energy, c is the minimum centroid time, a is the height of the function ($c + a$ is the maximum centroid time) and k is the slope of the function.

$$f(x) = \frac{a}{1 + e^{-k(x-x_0)}} + c \tag{2.1}$$

		acyl-ACP	KS and ACP-KS
Pressures (mbar)	Backing†	3.5	5.1
	Trap†	2.1×10^{-2}	4.0×10^{-2}
	IMS	4.4×10^{-1}	4.3×10^{-1}
	ToF	1.5×10^{-6}	2.4×10^{-6}
Voltages (V)	Trap	5–20	10–160
	Transfer	7	8
	Bias	15	20
	Trap Height	10	15
	Extract Height	5	10
T-wave (V at m/s)	Trap	0.2 at 300	0.2 at 300
	IMS†	7 at 280	10 at 320
	Transfer	3 at 248	3 at 248

Table 2.9 | Optimised instrument (Synapt HDMS) settings for CIU experiments. The Trap voltages were incremented in steps of 0.5 V and 5 V for the two groups of experiments respectively. T-wave settings are for the wave height (V) at wave velocity (m/s) respectively. Settings marked with a † significantly affect the transmission of high m/z ions.

The data were fit using the `curve_fit` function from `scipy.optimize` with a function equivalent to Equation 2.1 and initial estimate values of 90, 5, 3 and 0.5 for x_0, c, a and k respectively.

2.3.6.2 Super-Coarse Grained Prediction of Composite CIU Profiles

By treating the collisional cross-section (CCS) of a protein ion as the area of a sphere ($CCS = \pi r^2$), a super-coarse grained approximation can be made.[167,168] A theoretical approximation of the CCS of PsyACP-KS1 was made by summing the volumes ($\frac{4}{3}\pi r^3$) of the respective spheres for PsyACP1 and PsyKS1 to give a composite sphere with an area approximate to this CCS (CCS_{acp+ks}):

$$CCS_{acp+ks} \approx (r^3_{acp+ks})^{2/3} \quad \text{where} \quad r^3_{acp+ks} \approx r^3_{acp} + r^3_{ks} \tag{2.2}$$

This approximation requires the data to be in E_{com} units, as the energy transferred during collision is dependent upon ion mass (Equation 1.23). The $^{TW}CCS_{N2}$ of all ions were calculated using the method set out by Ruotolo *et al.*[55], with native myoglobin (7+→9+) and alcohol dehydrogenase (24+→26+) as calibrants, giving values of $X = 0.18315$, $m = 10.2902$ and $c = -0.03431$ with a residual error (R^2) of 0.9995. The CCS_{N2} values for these calibrants were taken from the McLean[169] and Bush[170]

databases respectively. For each arrival time distribution the weighted average CCS (CCS_{avg}) was calculated. The CCS_{avg} and E_{com} values for all proteins were interpolated between 0.05 and 1.5 eV with a spacing of 0.01 eV using the `interp1d` function from `scipy.interpolate`, to allow for summing. For each energy the CCS_{acp+ks} was calculated according to Equation 2.2, using the mean CCS_{avg} across charge states for PsyACP1 and PsyKS1.

2.3.7 Native Ppant Ejection

2.3.7.1 Native Ppant Ejection of PsyACP-KS1

Holo-PsyACP-KS1 was prepared as in Section 2.4.1 and exchanged into 25 mM AmAc as described in Section 2.3.4. Samples were diluted 4-fold into 200 mM AmAc immediately prior to spray by nESI (Section 2.3.5). The 17+ charge state (4397 m/z) was isolated with the quadrupole (LM: 4 and HM: 4) and collided with SF_6 at increasing collision voltages (20–160 V, E_{com} = 0.65–5.24 eV)

2.3.7.2 Native and Denatured Ppant Ejection of PsyACP1

Holo-PsyACP1 was prepared as in Section 2.4.1 and exchanged into 25 mM AmAc as described in Section 2.3.4. Samples were diluted 5-fold to give 25 mM AmAc; 25 mM AmAc, 0.1 % FA; and 60 % MeCN, 0.1 % FA. All samples were sprayed by nESI (Section 2.3.5) on a Synapt HDMS with the 7+ charge state (1847 m/z) isolated by the quadrupole (LM: 8 amd HM: 10). Ions were fragmented by CID with 10–80 V (E_{com} = 0.22–1.77 eV) collision voltage. *Apo*-PsyACP1^{7+} (1798.7 m/z) was isolated and fragmented by CID. Fragment ions were identified using ProSite Light[171] (v1.4). Denatured *holo*-PsyACP1 charges states 7+ (1847 m/z) and 13+ (995 m/z) were isolated and fragmented by CID at 75 V (E_{com} = 1.66 eV) and 35 V (E_{com} = 1.44 eV) respectively.

2.3.7.3 Ppant Ejection and Fragmentation of CoA

Coenzyme A was prepared as a 20 µM solution in 10 mM AmAc and sprayed by nESI (Section 2.3.5) under conditions for native PsyACP1. The CoA^+ ion (768.1 m/z) was isolated using a quadrupole (LM: 10, HM: 10) and collided with Ar gas at increasing collision voltages (10–60 V).

2.3.7.4 Supercharging of PsyACP-KS1

PsyACP-KS1 was exchanged into 25 mM AmAc, as in Section 2.3.4. Samples were diluted 4-fold into 200 mM AmAc or 200 mM AmAc with 2.5 % sulfolane immediately prior to spraying by nESI (Section 2.3.5). *Holo*-PsyACP-KS1 was prepared as described in Section 2.4.1 before exchange into AmAc. The 23+ charge state (3312 *m/z*) of supercharged PsyACP-KS1 was isolated and subjected to increasing collision voltage.

2.3.8 In-Gel Digestion of Proteins for LC–MS

Protein samples were prepared for peptide LC–MS by the in-gel digestion protocol described by Shevchenko *et al.* [172]. Briefly, bands on an SDS-PAGE gel were excised with a scalper and cut into 1 mm^3 cubes and transferred to a microcentrifuge tube. The cubes were destained with the addition of 50 mM AmBic, 50 % MeCN (50 µL) and incubated for 10 min; MeCN (450 µL) was then added and incubated for 5 min to dehydrate the cubes. The liquid was discarded, 10 mM DTT in 100 mM AmBic (50 µL) was added and incubated for 30 min at 55 °C, again followed by MeCN (450 µL) to dehydrate the cubes. The liquid was discarded, 55 mM iodoacetamide in 100 mM AmBic (50 µL) was added and incubated for 30 min in the dark, followed by MeCN (450 µL). The liquid was discarded and residual MeCN allowed to evaporate, mass spectrometry grade trypsin (10 ng/µL) or AspN (2 ng/µL) in 100 mM AmBic (50 µL) was added and incubated at 37 °C overnight. The reaction was stopped with 1 µL FA and analysed by LC–MS (Section 2.3.8.1).

2.3.8.1 Peptide LC–MS

Typically, using an UltiMate 3000 nano-LC instrument, 2–5 µL of a digest reaction was injected onto a PepMap C_{18} trap column and washed for 3 min with mobile phase A (5 % MeCN, 0.1 % FA) at a flow rate of 300 µL/min. The trap column was then switched inline with a PepMap C_{18} nano column with a flow rate of 0.3 µL/min, this was followed by a 30 min linear gradient to 55 % mobile phase B (95 % MeCN, 0.1 % FA), a 5 min step at 90 % mobile phase B and column equilibration back to mobile phase A for 20 min. The nano column was connected directly to a PicoTip operating at 1.7 kV

towards an LTQ-FT Ultra. Spectra were acquired in positive ion mode with a nominal resolution of 100 000 (at 400 *m/z*) and full mass range of 400–2000 *m/z*.

For peptide identification, a dynamic exclusion data-dependent method was used. Briefly, the four most intense ions from a full scan spectrum were isolated (8 *m/z* window) and subjected to CID fragmentation (35 % nominal energy). Any singly charged ions were rejected; any ions observed for 3 consecutive scans were also rejected for a period of 6 min. The data were searched with X!Tandem (via SearchGUI[173] v3.3.6 and PeptideShaker[174] v1.16) against a custom database containing the protein of interest and the common Repository of Adventitious Proteins (ftp://ftp.thegpm.org/fasta/cRAP/crap.fasta).

2.3.9 Carbene Footprinting of PsyACP1

PsyACP1 was cleaved with thrombin to remove the flexible His-tag that might obscure labelling. *Holo*-PsyACP1 was generated as in Section 2.4.1. PsyACP1 (25 µM) was incubated with **4-(3-trifluoromethyl)-3*H*-diazirin-3-yl)benzoic acid** (TDBA, 10 mM) (Sigma-Aldrich) for 5 min at room temperature, aliquots (6 µL were transferred to **tapered polypropylene micro-vials (0.3 mL, 11 mm, snap-ring)** (Fisherbrand) and flash-frozen with liquid N_2. Samples were immediately irradiated for 16 s at 349 nm by an **Explorer 349 Nd:YLF laser (1000 Hz repetition frequency, 125 µJ pulse energy)** (Spectra Physics). Irradiated samples were then separated by SDS-PAGE (15 % acrylamide) and prepared as described in Section 2.3.8 using AspN as the protease.

2.3.10 Coenzyme A Derivative HPLC

An **HP 1100 series HPLC system** (Agilent) was used with a **SunFire C_{18} (2.1 × 30 mm 3.5 µm)** (Waters) column. A step-gradient was performed at 0.5 mL/min: 0 min (10 mM AmAc), 1 min (10 mM AmAc), 1.1 min (10 mM AmAc, 3 % MeCN), 6 min (10 mM AmAc, 30 % MeCN), 6.5 min (10 mM AmAc), total time 11 min. Absorbance at 260 nm was recorded with a reference at 330 nm with a Diode Array Detector (G1315B, Agilent).

2.4 Enzyme Assays

All assays were performed in storage buffer (20 mM Tris-HCl, 100 mM NaCl, 10 % glycerol, pH 7.6) at 25 °C unless otherwise stated.

2.4.1 Synthesis of *holo*/acyl-ACP

A sample of *apo*-ACP (50 µM) was incubated with 2 eqv. of the appropriate coenzyme A derivative and Svp (1 µM) in storage buffer containing $MgCl_2$ (10 mM). The reaction was initially monitored by ESI-MS, as described in Section 2.3.3, with the reaction found to be complete within 10 min.

2.4.2 Acyl-ACP Hydrolysis by PedC/PedD^{R97Q}

Acetyl-, butyryl- and octanoyl-ACPs (20 µM), generated as described in Section 2.4.1, were incubated with PedC (5 µM) in storage buffer for up to 1 h. Aliquots (10 µL) were taken at various time points and quenched by the addition of TFA (1 %, 1 µL). These samples were then prepared for ESI-MS (see Section 2.3.2). Control reactions were performed in the absence of PedC to monitor background hydrolysis for up to 24 h. A similar experiment was performed with 5 mM reduced glutathione to monitor thiolysis. PsyACP1 and PsyACP3 were monitored with Ppant ejection on a Synapt HDMS, while PsyACP4 was monitored by full scan on a QToF2.

Acetyl-PsyACP1 (30 µM) was incubated with PedD^{R97Q} (30 µM) for 1 h before quenching with the addition of TFA (1 %, 1 µL) and clean-up for ESI-MS (see Section 2.3.2).

2.4.3 Malonyl-Loading of *holo*-ACPs by PedD

Holo-PsyACP1 and *holo*-PsyACP3 were incubated with PedD (5 µM) and malonyl-CoA (1 mM) in storage buffer for 25 min and 15 min respectively. The reactions were then loaded onto a C_{18} ZipTip and prepared for ESI-MS as in Section 2.3.2. Concurrent control reactions were performed in the absence of PedD to monitor self-malonation of the ACP. The acidification of this reaction with TFA or FA was found to cause precipitation and block the ZipTip, this was therefore avoided in these samples.

2.4.4 Loading of Malonyl-Ppant with Svp

Apo-PsyACP1 and *apo*-PsyACP3 (20 µM) were incubated with malonyl-CoA (200 µM) and Svp (1 µM) in storage buffer containing $MgCl_2$ (10 mM) for 25 min. The reaction was stopped with the addition of TFA (1 %, 1 µL) and prepared for ESI-MS (Section 2.3.2)

2.4.5 GNAT-Mediated Acylation of *holo*-PsyACP1

Holo-PsyACP1 was generated as described in Section 2.4.1. For overall interaction, *holo*-PsyACP1 (20 µM) was incubated with PsyGNAT (10 µM) and malonyl-CoA (400 µM) in storage buffer at 25 °C. Aliquots were removed over 24 h followed by mass spectrometry analysis (Section 2.3.2). Concurrent control reactions were performed without GNAT to monitor background processes. For determining acetyl transfer, *holo*-PsyACP1 (20 µM) was incubated with PsyGNAT (10 µM) and acetyl-CoA (400 µM) in storage buffer at 25 °C.

For ACP-independent decarboxylation, malonyl-CoA (200 µM) was incubated with PsyGNAT (2 µM) in storage buffer with and without a cocktail of metals (2 mM $CaCl_2$, 2 mM $MgCl_2$, 0.5 mM $MnCl_2$ and 0.1 mM FeOAc) at 25 °C. Aliquots were removed over 24 h followed by HPLC analysis (Section 2.3.10). Concurrent control reactions were performed without GNAT to monitor background decarboxylation.

2.5 Computational Methods

All bioinformatic analysis and software development was performed using the Python programming language, on a machine running the Ubuntu 18.04 operating system. Subsequent testing was performed on machines running the Windows 7/10 and OS X 10.11 operating systems where necessary.

2.5.1 Phylogenetic Analysis

The sequences of 315 ACP/PCP and 260 KS domains from 22 *trans*-AT PKS (shown in Table 2.10) were retrieved from the MIBiG online repository[175] (http://mibig.secondarymetabolites.org/) where possible, although a small number of sequences

were retrieved from the NCBI database. Domains were selected to be from *trans*-AT PKS gene clusters either used by Nguyen *et al.*[106] to develop a classification for KS, or containing putative GNAT domains and were cross-checked against recent reviews of *trans*-AT PKS.[82,83]

Pathway	Codifier	Source	MIBiG ID	# of CP
Bacillaene†	Bae	*Bacillus amyloliquefaciens*	BGC0001089	19
Batumin	Bat	*Pseudomonas fluorescens*	BGC0001099	16
Bongkrekic acid	Bon	*Burkholderia gladioli*	BGC0000173	13
Bryostatin	Bry	Symbiont of *Ca.* Endobugula sertula	BGC0000174	18
Chivosazol†	Chi	*Sorangium cellulosum*	BGC0001069	24
Diaphorin	Dip	*Ca.* Profftella armatura	BGC0001092	13
Difficidin†	Dif	*Bacillus amyloliquefaciens*	BGC0000176	18
Disorazol	Dsz	*Sorangium cellulosum*	BGC0001093	11
Enacyloxin	Ena	*Burkholderia ambifaria AMMD*	BGC0001094	12
Kirromycin	Kir	*Streptomyces collinus*	BGC0001070	17
Lankacidin†	Lkc	*Streptomyces rochei*	BGC0001100	5
Leinamycin†	Lmn	*Streptomyces atroolivaceus*	BGC0001101	10
Macrolactin†	Mln	*Jeotgalibacillus marinus*	BGC0001383	15
Mupirocin†	Mmp	*Pseudomonas fluorescens*	BGC0000182	11
Myxovirescin†	Ta	*Myxococcus xanthus*	BGC0001025	17
Nosperin	Nsp	*Nostoc* sp. *Peltigera membranacea*	BGC0001071	14
Onnamide†	Onn	Symbiont of *Theonella swinhoei*	BGC0001105	11
Pederin†	Ped	Symbiont of *Paederus fuscipes*	BGC0001108	17
Psymberin	Psy	Symbiont of *Psammocinia* aff. *bulbosa*	BGC0001110	13
Rhizoxin	Rhi	*Burkholderia rhizoxinica*	BGC0001112	20
Thailandamide†	Tai	*Burkholderia thailandensis*	BGC0000186	20

Table 2.10 | Gene clusters used for phylogenetic analysis of *trans*-AT PKS CPs showing the codifiers used to refer to domains from these clusters and the number of carrier proteins retrieved for each cluster. The sources listed are those for the MIBiG ID of the pathway or the associated NCBI accession. *Ca.*, *Candidatus*; aff., species affinis. †*Clusters used to determine KS clades*[106]

ACP sequences were aligned using **Clustal Omega**[176,177] (v1.2.2, default settings) to determine the Ppant attachment site for each; sequences were then extended to 35 residues flanking this site. The resultant sequence database was aligned using Clustal Omega (maximum of five combined iterations and three HMM iterations) and a maximum likelihood tree generated using PhyML[178] (v3.3, LG substitution model, BioNJ initial tree, NNI topology search, aBayes branch support, random seed = 1491938582). The tree was rooted to the *cis*-AT EryACP4 and visualised as a circular phylogram in FigTree (github.com/rambaut/figtree).

2.5.1.1 Ketosynthase Clade Determination

For any unassigned KS, the clade was determined by performing a BLAST[179] (Gap Cost: Existence = 13, Extension = 1) search against a reference database of assigned KS, the closest homologues were then recorded. The reference database contained all KSs used by Nguyen *et al.*[106] with the addition of EryKS4, a *cis*-AT KS. KSs were conferred the same clade as the closest homologues, if they were identical, and given no assignment if the two closest homologues were of different clades.

2.5.1.2 HMM Profile Analysis

After grouping sequences as above, the groups were individually aligned using Clustal Omega (maximum of five combined iterations and three HMM iterations) and a HMM profile generated for each using `hmmbuild` (default settings) from **HMMER3**.[180–182] The profile was subsequently searched against the ACP database using `hmmsearch` (default settings) from HMMER3. The bit scores from the profile search were then parsed with any absent scores being conferred a value of 10 for comparison giving S_g. Sequence logos for these alignments were produced with webLogo[183] (v3.7.4, no compositional adjustment). Helix regions were inferred from the solution NMR structure of MmpACP7a[127] (PDB: 2L22, 1–76).

Adapting a method by Haines *et al.*[127], the above process was applied to an alignment of all ACPs to generate standard scores for each ACP S_s; the quality of a group was then be inferred by comparing these scores against the standard scores where values of $S_g/S_s > 1$ were considered better matched to the grouped profile than the standard profile. These were plotted against each other, with a line $y = x$ dividing the plot, such that data points above the line represent $S_g/S_s > 1$. This ratio was also used to generate clustered heatmaps of the HMMs using the `clustermap` function from the Python package `seaborn`.

2.5.1.3 Homology Modelling

Homology models were generated using the CPHmodels[184] server (v3.2) which uses profile–profile alignment for remote homology prediction. The quality of the models was verified by assessing Ramachandran plots generated by the RAMPAGE web server

(http://mordred.bioc.cam.ac.uk/~rapper/rampage.php). All modelled structures were rendered in PyMol[185] with ray tracing.

2.5.2 PepFoot

PepFoot was developed in the **Python** (v3.6.6) programming language with **Jupyter Notebook** (v5.0.0) used for prototyping and **Spyder IDE** (v3.2.6) used for full development. The graphical user interface was built in Qt5 using PyQt (v5.11.2) and the **Qt Designer** (v5.9.6) application. In addition to standard packages: `h5py`, `numpy`[186] and `scipy` were used to handle and process MS data; `pyteomics`[187,188] was used to generate theoretical peptide information; `matplotlib`[189] was used to visualise data; and `json` was used to interact with the .pfoot file format. NGL Viewer[190,191] was incorporated as a standalone JavaScript file using the HTML web engine functionality of Qt.

2.5.2.1 Data Processing Benchmarks

The batch processing speeds were determined by taking a fully analysed .pfoot project and sequentially removing assignments. For each number of assignments, the time taken to complete processing of a data file was monitored using the Python `time` module. The time taken to perform the initial read (including generation of the in-memory ragged *m/z* array) and to integrate all assigned peptides were measured separately.

2.5.2.2 Benckmarking against Published Data

Data for benchmarking was acquired from the public **Proteomics Identifications Database** (PRIDE, https://www.ebi.ac.uk/pride).[192,193] The data files were converted to the .mz5 file format using the **MSConvert**[194] (ProteoWizard).

3

Enzymatic Interactions of Acyl-Acyl Carrier Proteins and Bioinformatic Rationale

3.1 Introduction

The ACPs of *trans*-AT PKSs and ACPs of modular type I PKSs in general, are often overlooked as benign carriers for the growing polyketide, with little more control over their interactions than a highly charged helix II to facilitate KS docking. It is becoming increasingly apparent, however, that there is more to these small four-helix bundles than meets the eye. Studies on type-II FAS ACPs show that they form a hydrophobic pocket to protect the acyl-Ppant from hydrolysis[118] with PKS ACPs showing aberrant Ppant sequestration[128,129] and chimeric modules show a preference for the native ACP–KS pairing.[121,195] To explore further any potential specificity imbued by the ACPs themselves, a set of enzymatic assays were developed — with a primary focus on the interactions of simple saturated acyl-ACPs with stand-alone AH and AT proteins. Initial studies by Jenner *et al.*[124] on a prototypical AH from the pederin polyketide synthase (PedC), showed a preference for hydrolysis of small acyl groups from ACPs — seemingly providing 'house keeping' activity for stalled modules. PedC has also been shown to be tolerant to a variety of acyl-SNACs.[137] The neighbouring AT from this PKS, PedD, has been shown to readily malonate a set of typical ACPs,[124,137] with specificity for α-unsubstituted malonyl-CoA. By mutating the active site Arg to Gln,

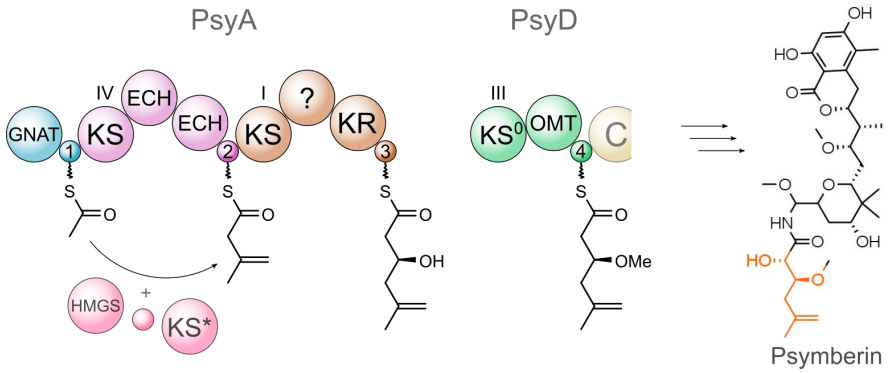


Figure 3.1 | Psymberin PKS modules 1–4 from PsyA and PsyD proteins.[108] The final state of the polyketide intermediate is shown for each module on the ACP. ACPs are numbered as used herein, KS clades are shown above each KS. The complete structure of psymberin is shown with the fragment contributed by these modules highlighted in orange.

PedD switches over to deacylating activity from acyl-ACPs, but crucially lacked the ability to hydrolyse the acyl-PedD^{R97Q} enzyme intermediate.[124] Only a single ACP (PsyACP3) was used in these studies, leaving open the possibility of ACP specific interactions regarding acyl-chain hydrolysis.

In elongating modules, ACPs are malonated prior to the elongation step, this is mediated through a *trans*-acting AT in the case of *trans*-AT PKSs. These *trans*-AT have been shown to have specificity for the different α-substituted malonyl-CoAs, but it is unclear whether the interactions between ACPs and *trans*-AT are generic. At least one ACP from a non-elongating module (PsyACP4) has been shown to be incapable of malonylation in the presence of a *trans*-AT,[137] while two others (RizACPs 12 and 18) have been shown to malonate with their native *trans*-AT.[196] As these modules have no decarboxylative activity, the presence of these malonyl units poses a problem to the flux of the biosynthesis. The authors proposed that there may be accessibility constraints in the full modular context preventing this malonation, or a demalonating enzyme acting in a way similar to AHs.

The main set of ACPs investigated in this chapter are from the psymberin PKS, the first four ACPs were selected to cover a diverse set of acyl-groups and interacting enzymes. Shown in Figure 3.1, PsyACP1 is from a GNAT loading module with a native acetyl starter unit, PsyACP2 is a β-branching ACP that has a β,γ'-double bond product, PsyACP3 terminates the PsyA PKS protein and docks to a downstream non-elongating KS_{III}, and PsyACP4 accepts the polyketide intermediate from PsyACP3 and allows *O*-methylation of the β-hydroxyl before passing on to a NRPS module.

3.2 Expression and Purification of ACPs

All ACPs were recombinantly expressed in *E. coli* as His-tag fusion proteins, followed by purification with IMAC as described in Section 2.2. PsyACP1, PsyACP3, PsyACP4, EcACP and PedACP4 were successfully expressed and purified (spectra shown in Figure 3.2). PsyACP2 was expressed in inclusion bodies, but successfully purified

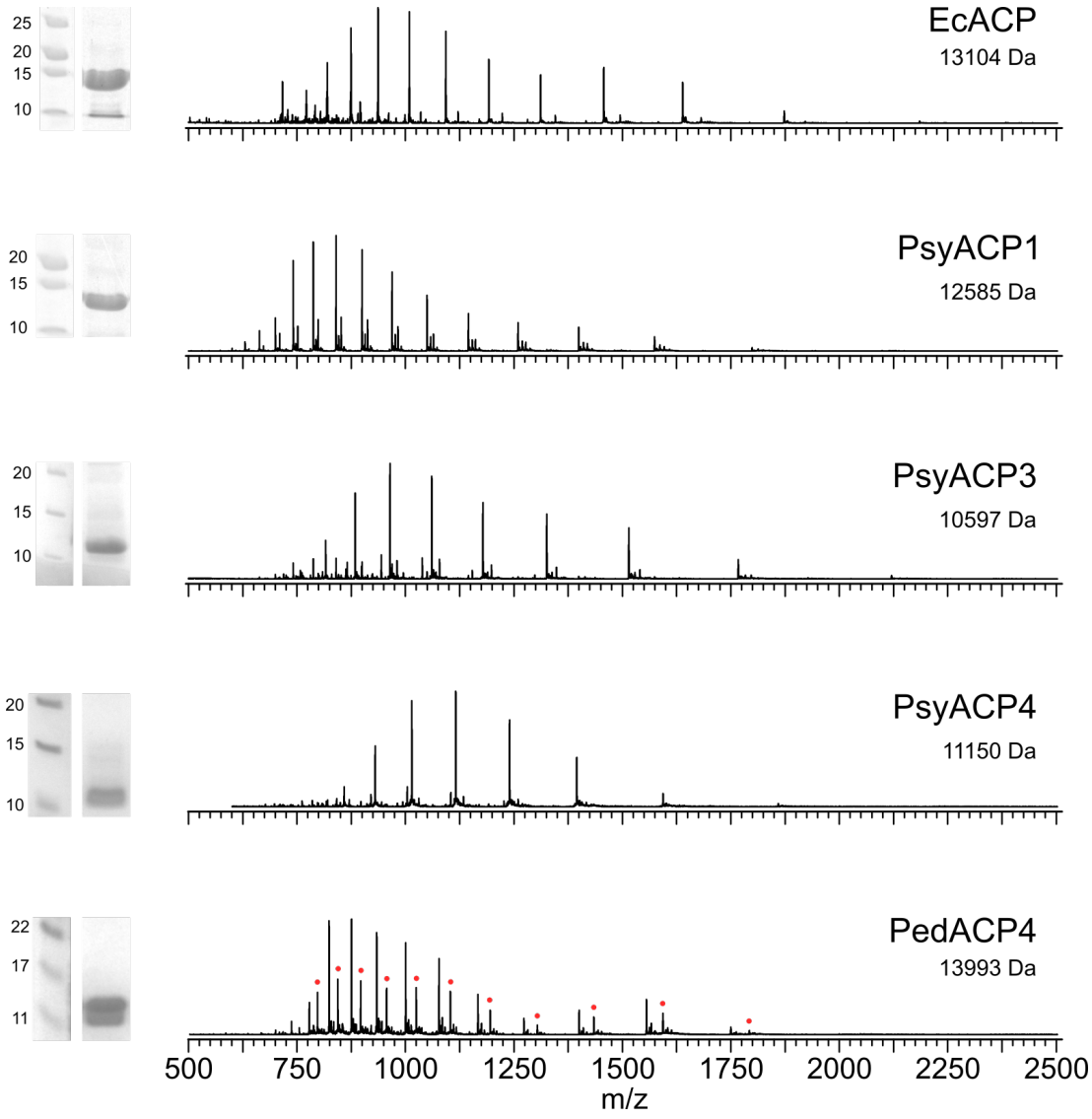


Figure 3.2 | SDS-PAGE of purified ACPs with molecular weight markers indicated (*left*). Stacked full scan denatured ESI-MS of successfully expressed ACPs prepared with a C_{18} ZipTip into 80 % MeCN, 0.1 % FA (*right*). *In vivo holo*-PedACP4 peaks are indicated with a red dot •.

by IMAC after solubilisation with 4 M urea. Removal of imidazole — into several exchange buffers by centrifugation, dialysis or column chromatography — resulted in precipitation of protein. All recombinant proteins except EcACP, the type II FAS ACP from *E. coli*, had the N-terminal Met removed *in vivo*, giving an intact mass 131 Da

less than the theoretical value. PedACP4 was found to have 40 % conversion to *holo*-ACP and was thus not used for further study, as acyl-PedACP4 production would be incomplete. All attempts at loading EcACP with Ppant using the PPTase Svp failed and thus was not used for further study.

3.3 Enzymatic Hydrolysis of Acyl-ACPs

Acyl-ACPs (PsyACP1, PsyACP3 and PsyACP4) were generated from acyl-CoAs (acetyl, butyryl and octanoyl) as described in Section 2.4.1. In all cases, ACPs were completely loaded within 10 min, allowing for minimal background hydrolysis prior to time-course experiments. The acyl-ACPs were incubated with PedC, the AH from the pederin PKS. PedC was co-expressed with GroEL for solubility and purified by Dr José Afonso and Dr Matthew Jenner following a previously described methodology.[137]

Ppant ejection[115] was used for monitoring the hydrolysis of acyl-PsyACP1 and acyl-PsyACP3; exemplar ejection spectra are shown for *holo*-, acetyl-, butyryl and octanoyl-PsyACP1 in Figure 3.3. Monitoring Ppant ejection ions, rather than precursor ions, allows for more direct comparison of reaction progress and, in the case of simple β-keto reduction, sufficient discrimination between acyl variants. By first monitoring the intrinsic hydrolysis of acyl-Ppant over 24 h, it is clear that small acyl-groups are more readily lost for both PsyACP1 and PsyACP3 with acyl-PsyACP3 exhibiting a greater extent of hydrolysis, shown in Figure 3.4 and Figure 3.5.

In the presence of PedC however, PsyACP1 shows protection of the smaller acetyl-group compared to the octanoyl, see Figure 3.6. This is not the case for PsyACP3, which behaves as previously[124] shown: small groups are preferentially hydrolysed by the enzyme, see Figure 3.7. Again, PsyACP3 is far more readily hydrolysed than PsyACP1, showing greater hydrolysis of all acyl-groups after 10 min than is seen for PsyACP1 after 60 min. It is clear that PsyACP1 does not intrinsically protect the smaller chains from hydrolysis, but that the PsyACP1–PedC interaction is in some way different from the PsyACP3–PedC interaction. From a biological perspective, it would be detrimental to psymberin biosynthesis if acetyl-PsyACP1 was rapidly hydrolysed. Psymberin biosynthesis is initiated with an acetyl unit,[108] provided by decarboxylation

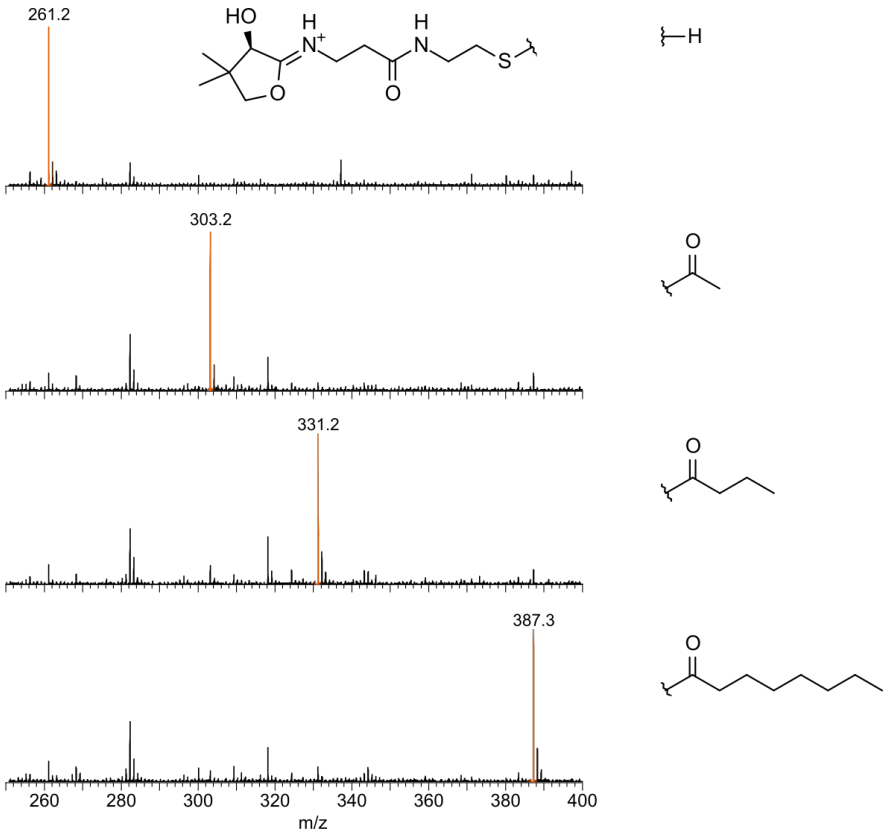


Figure 3.3 | Ppant ejection spectra for *holo*-, acetyl-, butyryl- and octanoyl-PsyACP1 after incubation of *apo*-PsyACP1 with 2 eqv. acyl-CoA and Svp for 10 min. $Pant^{+}$ ions are highlighted in orange: *holo* = 261.2, acetyl = 303.2, butyryl = 331.2 and octanoyl = 387.3 *m/z*.

of malonyl-CoA by the loading module GNAT. Likewise, PsyACP3 would be stalled by an acetyl-Ppant, and thus requires facile cleaning to allow biosynthesis to continue.

To see whether the background deacylation of PsyACP1 was truly independent of the ACP, a stronger nucleophile than water — a thiol such as glutathione — can be used to increase the rate of reaction. Doing so achieves similar results after 2 h to hydrolysis after 24 h (see Figure 3.8), with acetyl-PsyACP1 thiolysing to a greater extent than the octanoyl moiety.

To probe these interactions further, an ACP from a non-elongating module was used, PsyACP4. In the presence of PedC, neither acetyl- nor octanoyl-PsyACP4 were hydrolysed above the background level, see Figure 3.9. Such ACPs would not be expected to be readily malonated, in fact PsyACP4 has been previously shown not to malonate with PedD.[137] The interactions of ACPs, and therefore their sequences, are more complex than would be first imagined, showing some degree of control over their acyl-state mediated by AHs. Of equal interest are the interactions of ACPs with ATs.

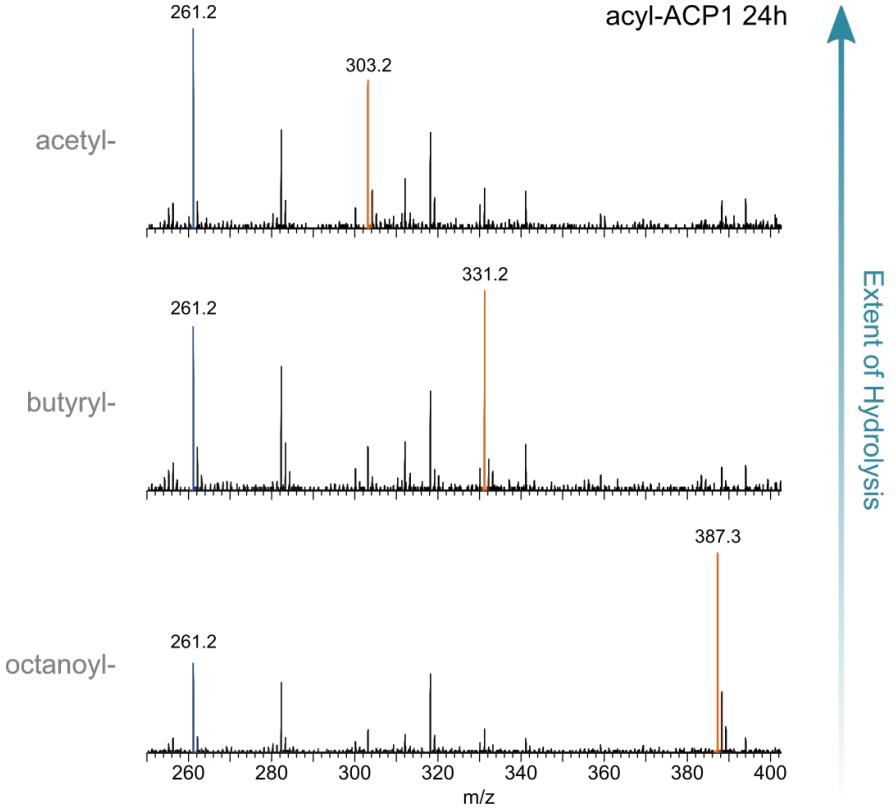


Figure 3.4 | Ppant ejection of acetyl- (*top*), butyryl- (*middle*) and octanoyl- (*bottom*) PsyACP1 incubated for 24 h in the absence of PedC; showing greater hydrolysis of small acyl groups. $Pant^{+}$ ions are coloured blue for *holo*- and orange for acyl-$Pant^{+}$.

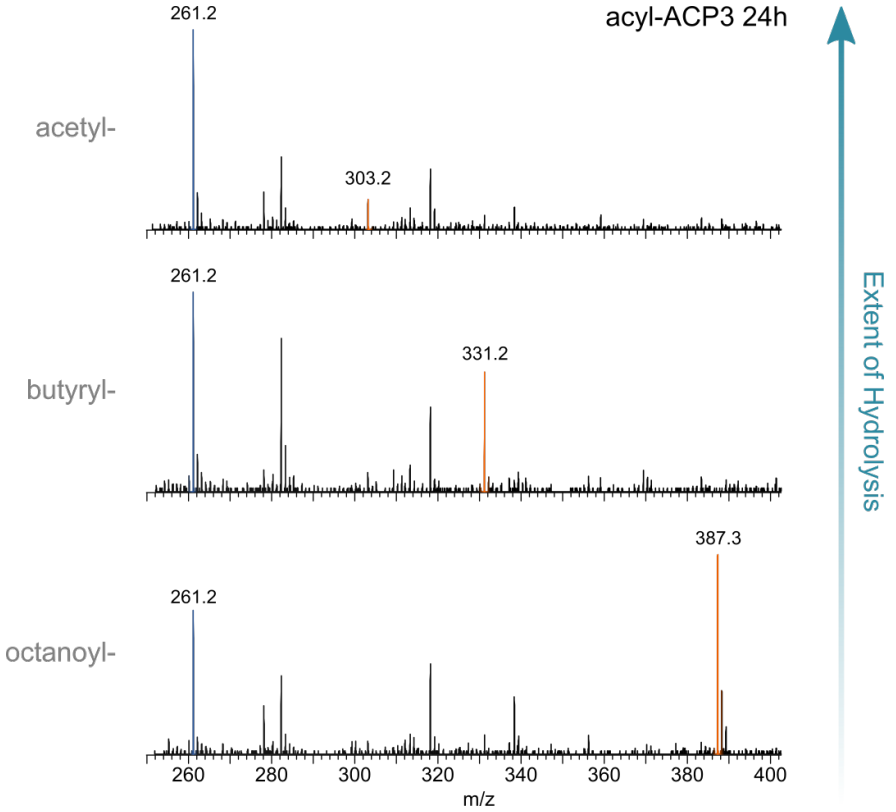


Figure 3.5 | Ppant ejection of acetyl- (*top*), butyryl- (*middle*) and octanoyl- (*bottom*) PsyACP3 incubated for 24 h in the absence of PedC; showing greater hydrolysis of small acyl groups. $Pant^{+}$ ions are coloured blue for *holo*- and orange for acyl-$Pant^{+}$.

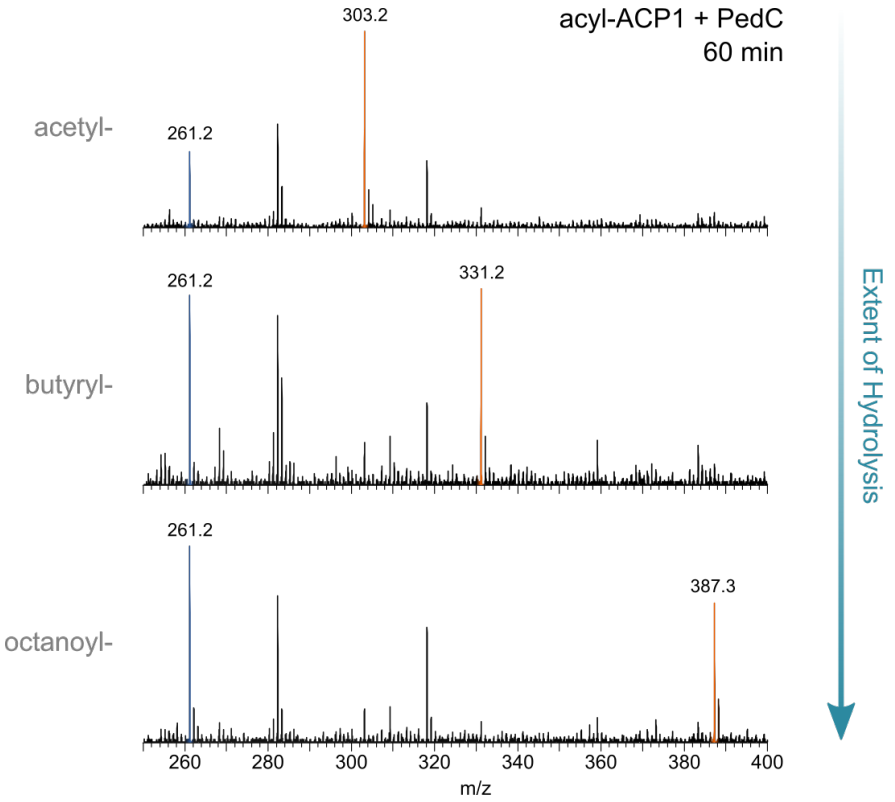


Figure 3.6 | Ppant ejection of acetyl- (*top*), butyryl- (*middle*) and octanoyl- (*bottom*) PsyACP1 incubated for 60 min in the presence of PedC (5 µM); showing protection of small acyl groups from hydrolysis. $Pant^+$ ions are coloured blue for *holo*- and orange for acyl-$Pant^+$.

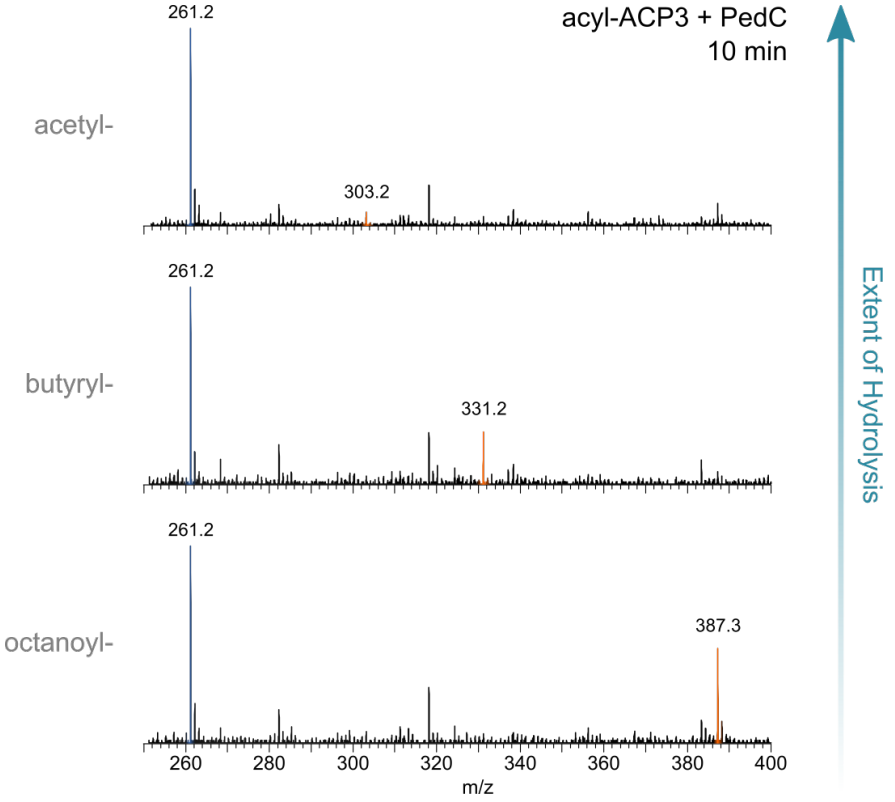


Figure 3.7 | Ppant ejection of acetyl- (*top*), butyryl- (*middle*) and octanoyl- (*bottom*) PsyACP3 incubated for 10 min in the presence of PedC (5 µM); showing enhanced hydrolysis of small acyl groups. $Pant^+$ ions are coloured blue for *holo*- and orange for acyl-$Pant^+$.

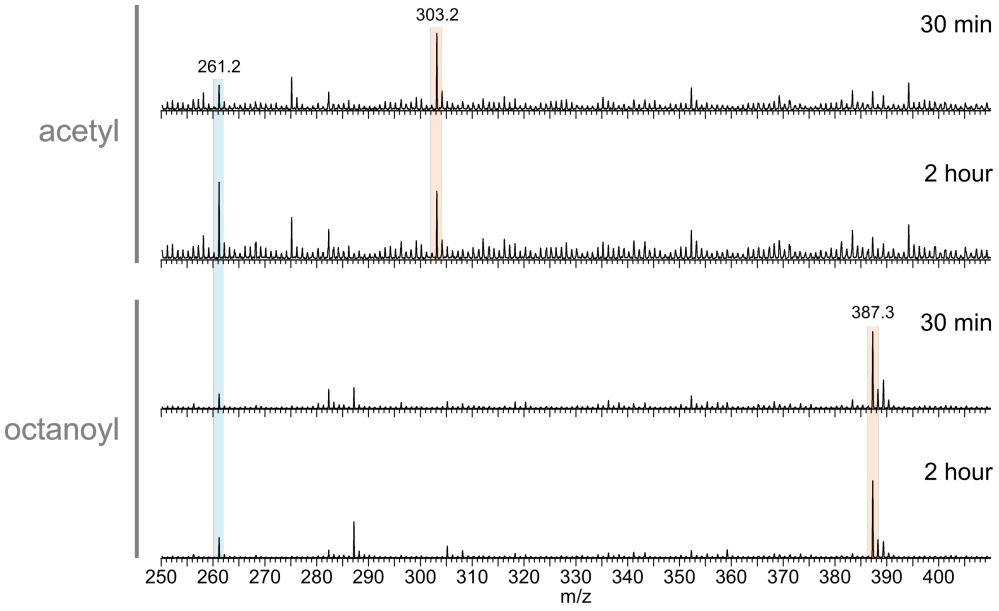


Figure 3.8 | Ppant ejection of acetyl- (*top*) and octanoyl- (*bottom*) PsyACP1 incubated with 5 mM glutathione; showing greater thiolysis of the acetyl group, the same relationship found in Figure 3.4. $Pant^+$ ions are coloured blue for *holo*- and orange for acyl-$Pant^+$.

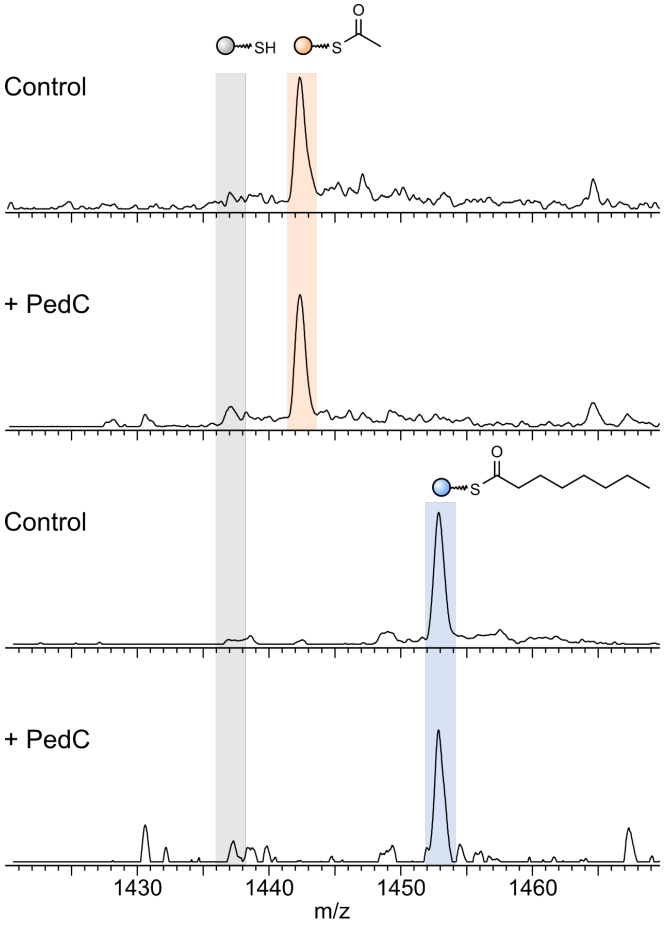


Figure 3.9 | Stacked ESI-MS of the 8+ charge state of acetyl- (*top*) and octanoyl-(*bottom*) PsyACP4 (20 µM) after 30 min incubation with PedC (5 µM). The *holo*, acetyl and octanoyl species are highlighted in grey, orange and blue respectively.

3.4 Generating Malonyl-ACPs

To probe whether the ACPs of loading modules interact with ATs, *holo*-PsyACP1 and *holo*-PsyACP3 were incubated with malonyl-CoA in the presence of PedD, the AT1-type AT from the closely related pederin PKS. PedD and PedD^{R97Q} were cloned and expressed as His-tag fusion proteins by Dr José Afonso and Dr Matthew Jenner as previously described.[124] Much like PsyACP4, *holo*-PsyACP1 does not malonate in the presence of PedD, or by self-malonation for general interactions with PedD, see Figure 3.10. PsyACP3 on the other hand, shows 70 % conversion, see Figure 3.11.

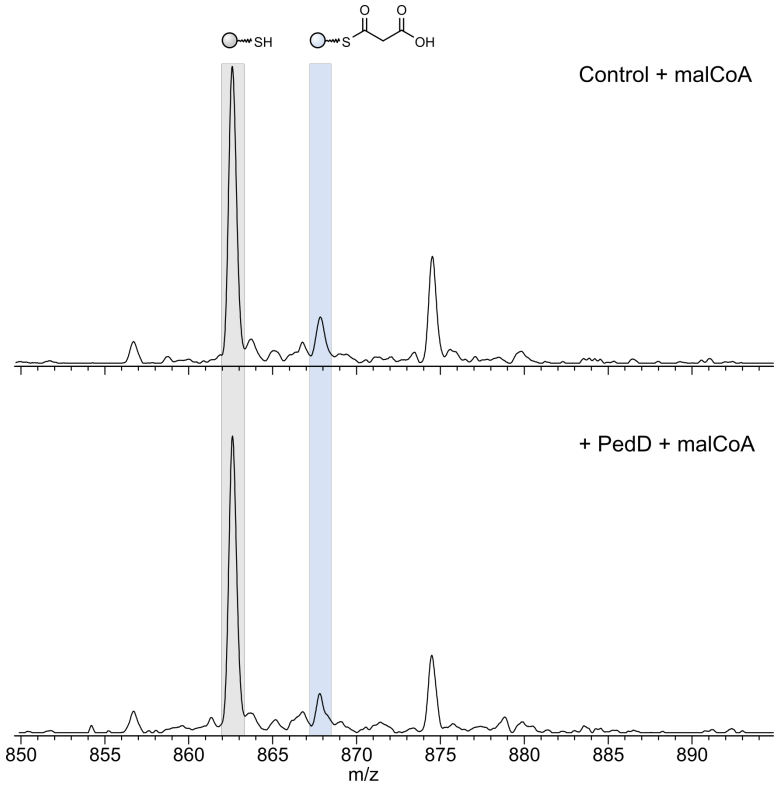


Figure 3.10 | Stacked ESI-MS of the 15+ charge state of PsyACP1 (20 µM) after 25 min incubation with PedD (5 µM) and malonyl-CoA (1 mM). The *holo* and malonyl species are highlighted in grey and blue respectively.

What is not apparent from the above results, is whether this anti-preference for malonation is due to an interaction between the ACP and the AT, or an interaction between the ACP and the malonyl group itself. By performing an acyl hydrolysis experiment with PedD^{R97Q} on acetyl-PsyACP1 (see Figure 3.12), it is clear that PsyACP1 does indeed interact with PedD — as only the active site residue is mutated, all other interactions should be identical to wild-type PedD. This points to some interaction between the malonyl-PedD and *holo*-PsyACP1, or the malonyl-Ppant and PsyACP1.

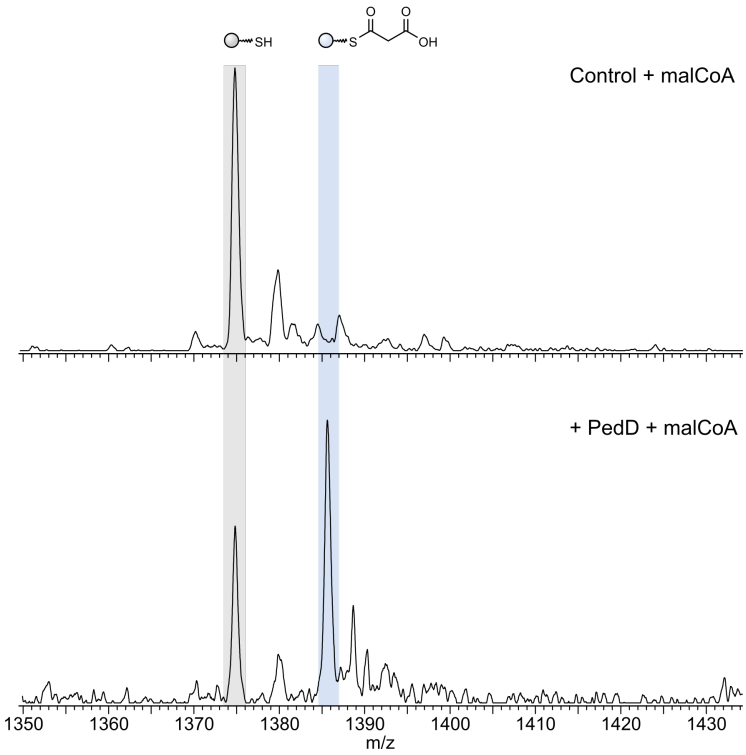


Figure 3.11 | Stacked ESI-MS of the 8+ charge state of PsyACP3 (20 µM) after 15 min incubation with PedD (5 µM) and malonyl-CoA (1 mM). The *holo* and malonyl species are highlighted in grey and blue respectively.

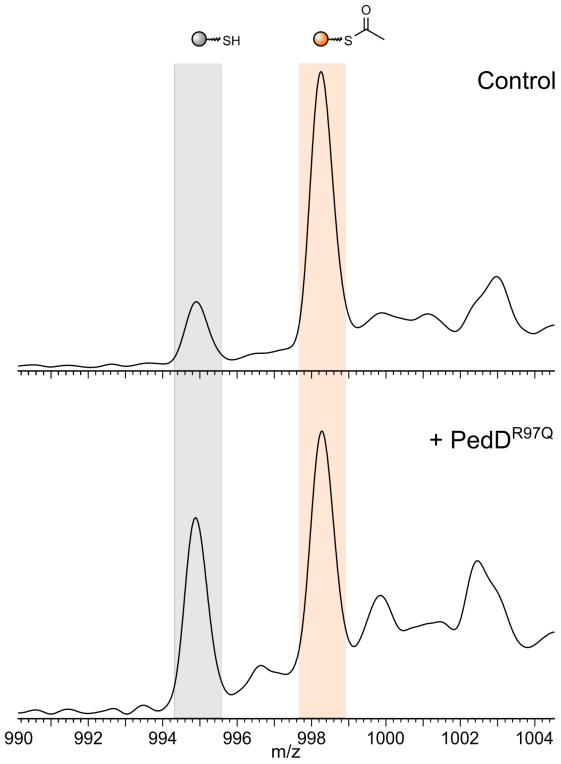


Figure 3.12 | Stacked ESI-MS of the 13+ charge state of acetyl-PsyACP1 (30 µM) after 1 h incubation with PedD^{R97Q} (30 µM). The *holo* and acetyl species are highlighted in grey and orange respectively.

One way to examine this is to directly load a malonyl-Ppant onto the ACP using a PPTase, in the same way as other acyl-ACPs have been generated above. Incubating PsyACPs with malonyl-CoA and Svp showed that malonyl-CoA is less preferred for loading onto ACPs than CoA, with at most 40 % conversion being observed for PsyACP3 and loading of free CoA to generate some *holo*-ACP, see Figure 3.14. PsyACP1 shows no loading of malonyl-Ppant under conditions that would otherwise generate complete conversion to *holo*/acyl-Ppant, see Figure 3.13. Seemingly, PsyACP1 is resistant to malonation, possibly due to some interaction between the malonyl unit and the ACP itself.

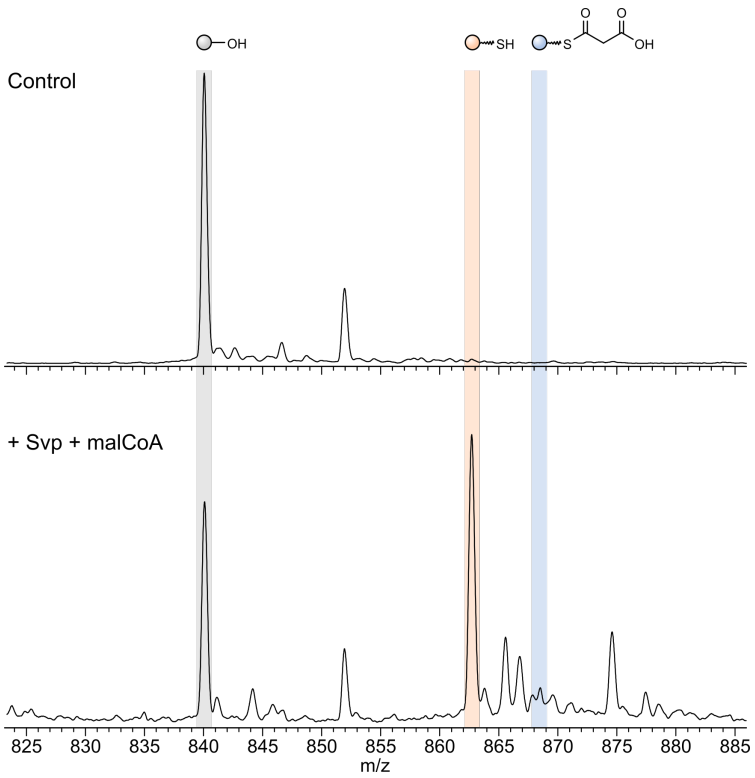


Figure 3.13 | Stacked ESI-MS of the 15+ charge state of PsyACP1 (20 µM) after 25 min incubation with Svp (1 µM) and malonyl-CoA (200 µM). The *apo*, *holo* and malonyl species are highlighted in grey and blue respectively

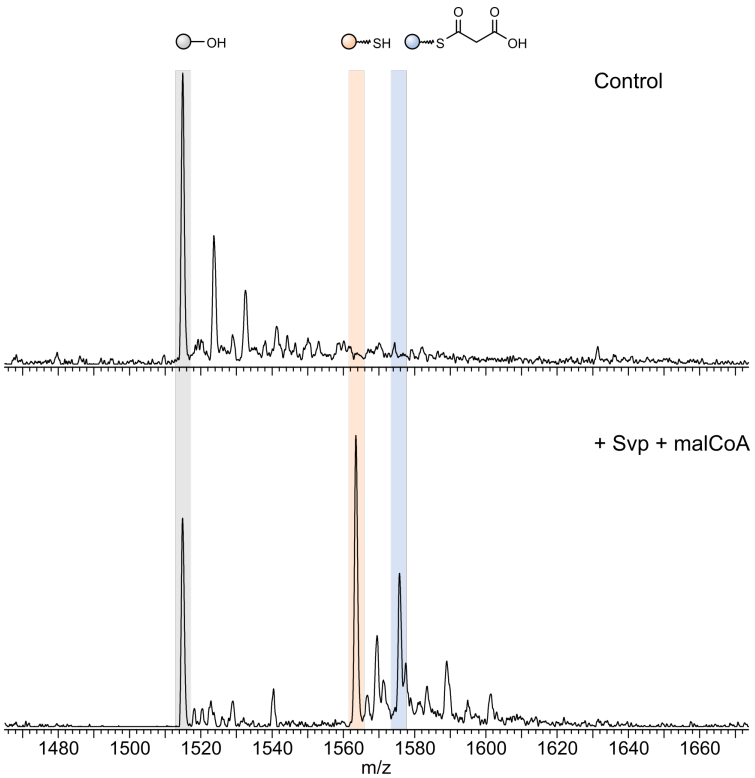


Figure 3.14 | Stacked ESI-MS of the 7+ charge state of PsyACP3 (20 µM) after 25 min incubation with Svp (1 µM) and malonyl-CoA (200 µM). The *apo*, *holo* and malonyl species are highlighted in grey and blue respectively.

3.4.1 GNAT–PsyACP1 Interactions

In the psymberin PKS, PsyACP1 is loaded directly with an acetyl group from the neighbouring GNAT domain. In principle, the ACP should not be malonated as it is possible that a malonyl unit on this ACP might 'stall' biosynthesis, much like PsyACP3 would be stalled by an acetyl group. If the GNAT were only capable of decarboxylating free malonyl-CoA and was hindered from decarboxylating malonyl-PsyACP1, biosynthesis would not proceed. It should be noted that several other ACPs from GNAT loading modules (CurA,[123] AprA[125] and StxA[126]) have been shown to produce malonyl-ACP when loaded with a PPTase, two of the GNATs from these modules have been shown to decarboxylate malonyl-ACP. This implies they can accept a substrate into their active site from the rear-side.

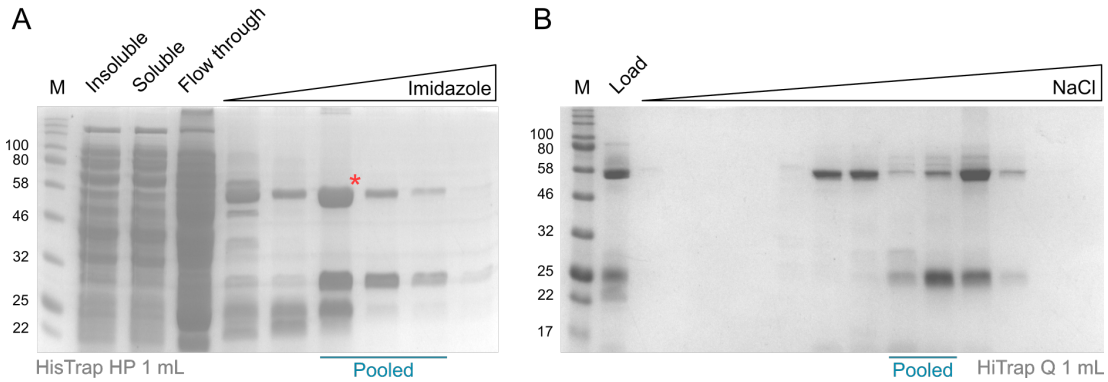


Figure 3.15 | (**A**) SDS-PAGE of PsyGNAT HisTrap affinity purification, fractions pooled for further purification are indicated. GroEL is highlighted with a red asterisk. (**B**) SDS-PAGE of PsyGNAT HiTrap Q ion exchange purification. **M**: Color prestained protein standard broad range (11–245 kDa, New England Biolabs)

PsyGNAT was expressed and purified in *E. coli* as a His-tag fusion protein as described in Section 2.2. After IMAC purification, PsyGNAT was shown to have co-expressed with a 57 kDa protein, see Figure 3.15A, confirmed by LC–MS/MS to be the chaperone protein GroEL. PsyGNAT was further purified with strong anion exchange chromatography to reduce the amount of GroEL present in the sample, see Figure 3.15B. The other GNAT containing constructs PsyGNAT-ACP1 and PsyAR-GNAT-ACP1 were subjected to expression tests as shown in Figure 3.16. Unfortunately, neither construct appears to be soluble under the standard lysis conditions. No further action was taken with these constructs.

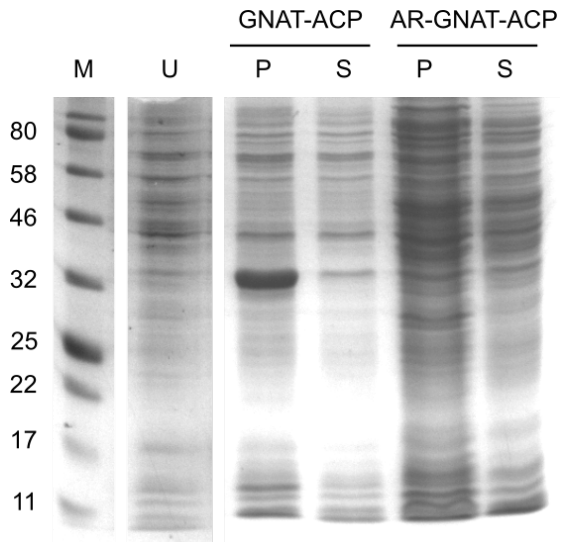


Figure 3.16 | SDS-PAGE of expression tests for PsyGNAT-ACP1 and PsyAR-GNAT-ACP1. **M**: Color prestained protein standard broad range (11–245 kDa, New England Biolabs), **U**: Uninduced BL21 (DE3), **P**: Pellet fraction after sonication, **S**: Supernatant fraction after sonication.

Enzymatic loading of malonyl- or acetyl- moieties onto *holo*-PsyACP1 by PsyGNAT was negligibly different from the background, as shown in Figure 3.17. To investigate the decarboxylative activity of GNAT, an modified HPLC assay[123] was used. This showed no decarboxylation above the background suggesting that this activity is not present in PsyGNAT unlike the CurA-GNAT,[123] see Figure 3.18. An alternative explanation, given the presence of GroEL in the purification, is that the PsyGNAT is not correctly folded after purification and hence inactive.

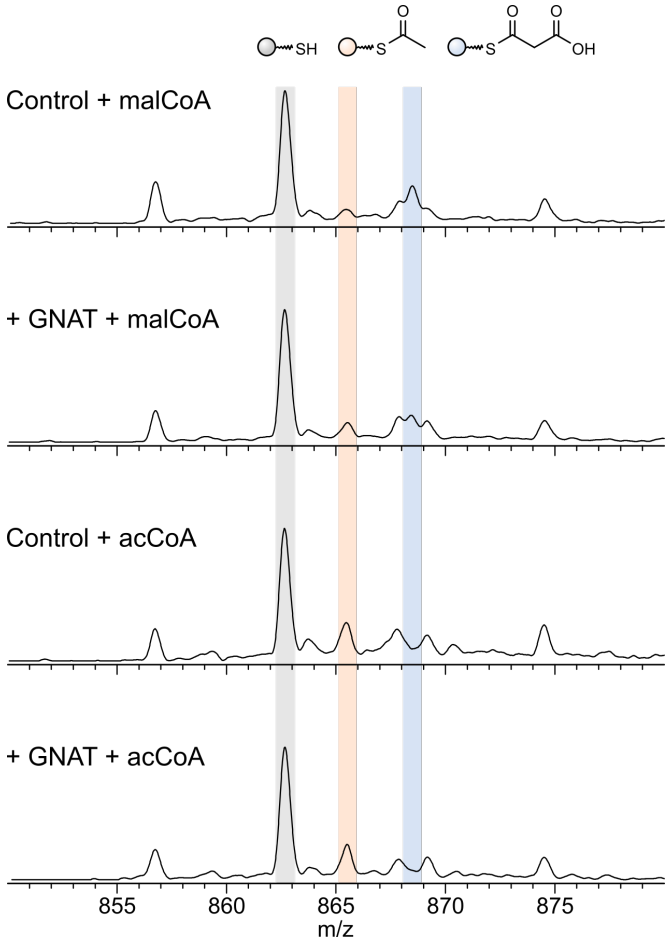


Figure 3.17 | Stacked ESI-MS of the 15+ charge state of *holo*-PsyACP1 (20 µM) incubated with PsyGNAT (10 µM) and malonyl-CoA (400 µM) for 20 h (*top*). Stacked ESI-MS of the 15+ charge state of *holo*-PsyACP1 (20 µM) incubated with PsyGNAT (10 µM) and acetyl-CoA (400 µM) for 20 h (*bottom*).

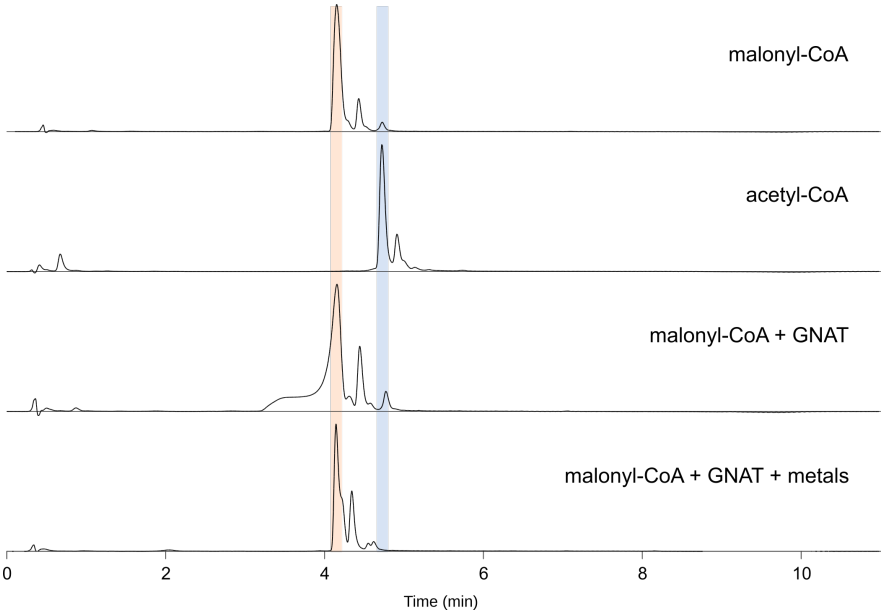


Figure 3.18 | HPLC UV chromatograms recorded at 260 nm for malonyl-CoA (200 µM) incubated with PsyGNAT (2 µM) after 20 h. Peaks corresponding to malonyl-CoA and acetyl-CoA are highlighted in orange and blue respectively. Metal mixture contained 2 mM Mg^{2+}, 2 mM Ca^{2+}, 0.5 mM Mn^{2+} and 0.1 mM Fe^{3+} to account for possible requirement in active site. [125]

3.5 A Bioinformatic Approach to ACP Specificity

It has been shown that KS domains can be highly substrate specific, and that the KSs of *trans*-AT PKS can be grouped into phylogenetic clades that closely match this specificity and general module architecture. Whether due to the supposed insignificance of ACPs, or the availability of sufficient sequences, such an approach has not been explored for ACPs, until recently. If ACPs were specific to their native acyl groups, or at least to their native modular architectures, it would be expected that they would clade, much like KSs. To explore this possibility, two separate routes were taken: (1) ACPs were directly grouped based on the assigned clade of neighbouring KSs, and (2) ACPs were indirectly grouped after phylogenetic clustering. For both routes, the same collection of 315 ACPs from 22 *trans*-AT PKSs were used.

A final count of 315 ACP sequences and 260 KS sequences were collected from the 22 *trans*-AT PKSs as described in Section 2.5.1 from the MIBiG repository. Occasional ACPs were missing from the MIBiG assignment and were instead collected from NCBI. All KS were present in the MIBiG assignments, the only exception to this was the PedI protein (the first PKS in the pederin cluster) where the entire protein was retrieved from NCBI. ACP and KS numbering refers to the modular position within the putative PKS assembly starting from the upstream end, e.g. BaeACP1 is the first ACP in the bacillaene PKS. In the case of tandem ACPs — where multiple ACPs are within one module — the sequences are denoted alphabetically in order of appearance, e.g. PedACP3b is the second tandem ACP in the third module in the PKS.

Due to the varied assignment of sequences with sequence lengths of 43 residues (TaACP16) up to 84 residues (PedACP3a) and the position of the conserved serine in these assignments varying greatly, the sequences were pruned to being 35 residues either side of the conserved serine, this distance was chosen to encompass the majority of the ACPs but not include any interdomain regions which may have higher variability. Additionally, any sequences with an identity of 90 % or greater were reduced to one entry to account for replicate ACPs — this was only the case for DipACP3, BonACP10 and LkcACP1.

3.5.1 Direct Grouping by KS Clade

After first assigning the clades of all neighbouring KSs, the ACPs were grouped by the clade of their native upstream or downstream KSs (see Table 3.1). The conservation of amino acids at specific positions can be displayed in several ways, the most common is in the form of a sequence logo — where the frequency of a particular amino acid for a given residue position is converted to a height. Figure 3.19 shows an overview of such sequence logos for these ACPs when grouped by downstream KS clade. Residues that may be involved in interactions with other proteins are highlighted with a triangle, typically these are highly conserved within the group and are not observed within the other groupings.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	C	—
Upstream	41	12	10	19	17	9	11	25	45	14	6	2	3	6	11	12	20	52
Downstream	48	13	10	18	17	9	13	18	44	15	5	2	2	6	10	11	18	56

Table 3.1 | Frequency of ACPs with an upstream or downstream KS by clade, including C/Cy domains. No assigned clade is represented as —, which includes occasions where an KS closely matches multiple clades and where there are no KS/C domains in that location.

Each group was then aligned by sequence and a hidden Markov model (HMM) profile produced. Every ACP in the dataset was scored against each of these profiles, as well as a profile for all ACPs in the dataset. These scores were then used to assess the similarity of these groups. By plotting these bit scores against each other, the ability to discriminate between these groups and all ACPs can be determined.[127] Plots for all of the downstream KS clade groupings can be found in Figure A.1.

As an example, Figure 3.20 shows the corresponding plot for a HMM profile for ACPs with a downstream KS^{0}_{III}. As can be seen, this grouping provides good discrimination from ACPs in general, which suggests that there are conserved sequence features specific to this group. There is a single ungrouped ACP with a better bit score for this group than general ACPs; this is TaACP14, an ACP at the C-terminus of the TaI protein, that docks to a C domain at the N-terminus of the Ta-1 protein. Both TaACP14 and the ACPs in this group dock to another PKS/NRPS protein, therefore the conserved residues in this HMM profile may be involved in this interaction. Several of the members of this group (BaeACP14, TaACP2, RhiACP1 and TaiACP4) are not PKS-terminating domains, but are instead non-elongating modules within a PKS protein.

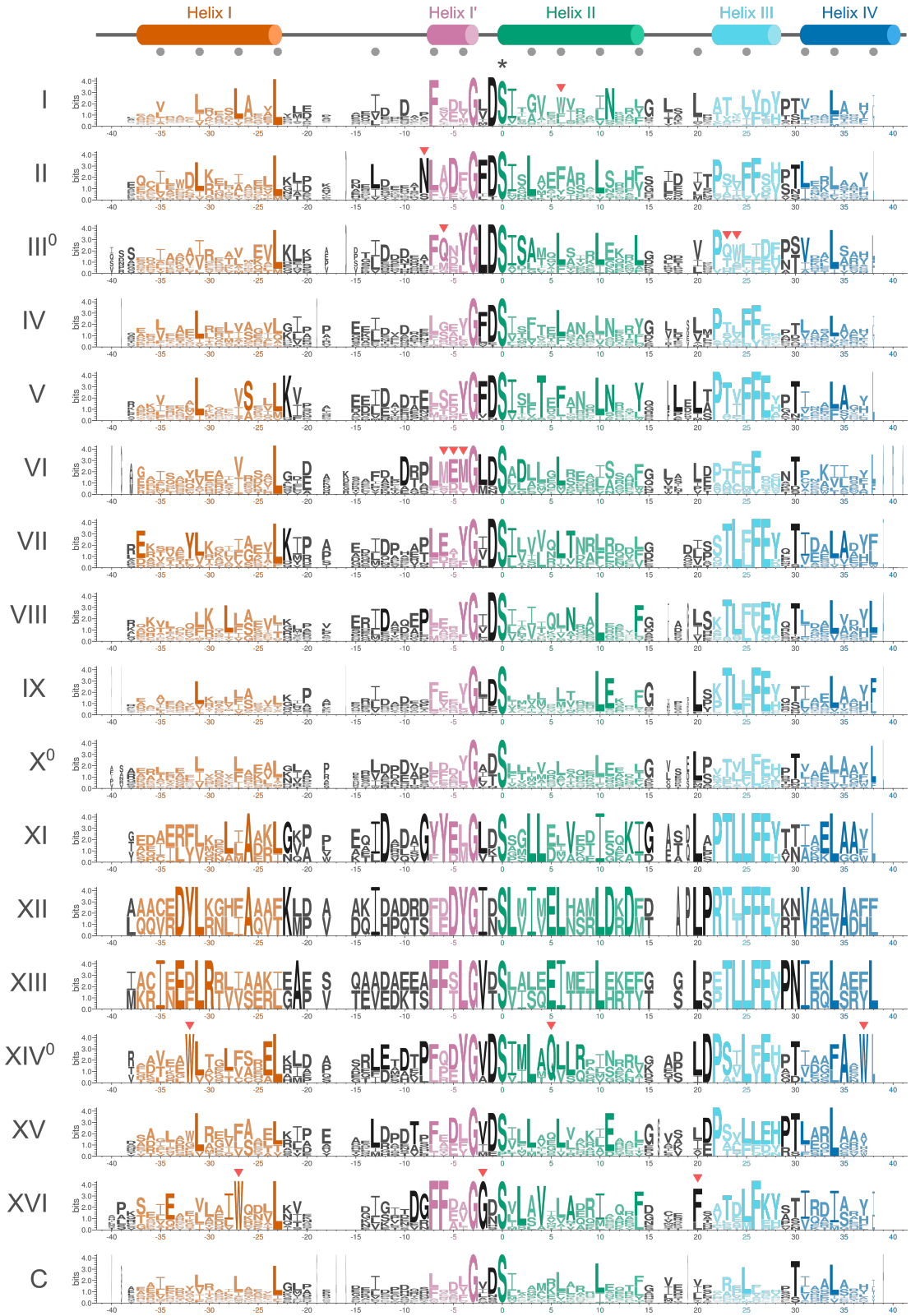


Figure 3.19 | Sequence logos for ACPs grouped by downstream KS clade. Sequences were taken from an alignment of 313 ACPs, without RhiACP8 and LkcACP2 as these introduced two large sections of gaps between helices II and III. Logos were generated using webLogo[183] v3.7.4, and manually coloured to reflect helix positions and conservation. The conserved serine for Ppant attachment is highlighted with an *, and all positions are relative to this. Potentially important residues are highlighted with a triangle. Helices are derived from a solution NMR structure of MmpACP7a[127] (PDB: 2L22, 1–76) and the positions of residues in the hydrophobic core indicated with circles.

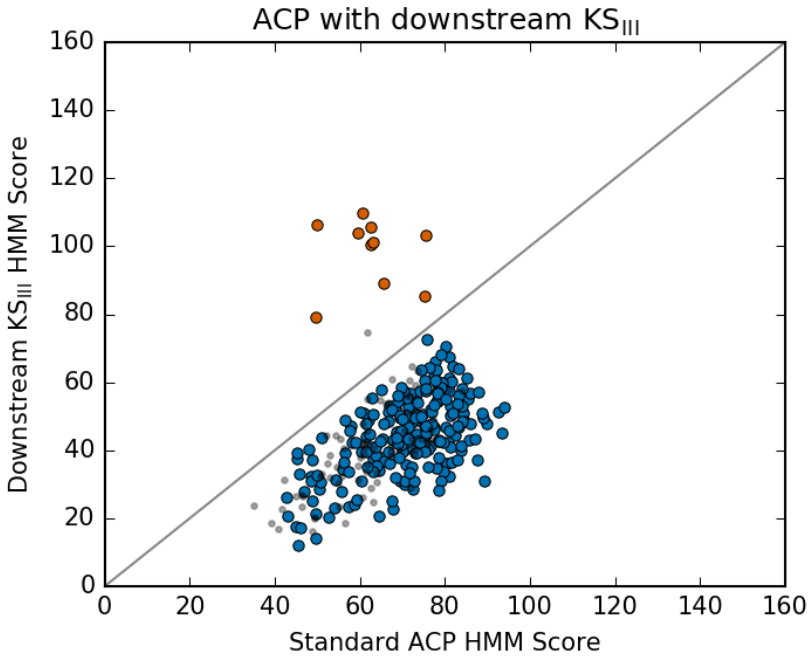


Figure 3.20 | Plot of bit scores for HMM profiles for ACPs with a downstream KS_{III} versus a standard HMM profile for all ACPs. ACPs from the KS_{III} group, assigned to other groups or with no assignment are shown as dots coloured orange, blue and grey respectively. A line $(y = x)$ divides the plot, ACPs above this line are more similar to the downstream KS_{III} group than to ACPs in general.

When these members are discounted from the group, the high conservation of several motifs becomes apparent. These are FQ(N/D)Y starting at -7 to the conserved Ser in helix I' and PQ**W**LID starting at +22 in helix III. Curiously, these ACPs are not followed by a docking domain, unlike many other module–module docking ACPs[197] (see Figure 3.21), which might imply that these motifs are necessary for successful docking to the downstream PKS protein. It has been proposed that tryptophan (Trp) residues facing into the hydrophobic core of ACPs 'lock' the structure to allow for correct recognition by *trans*-acting enzymes, such is the case for β-branching ACPs (see Figure 3.22) where a Trp residue +6 to the conserved Ser is critical for interaction with the HMGS, possibly by ensuring helix III and the Ppant are correctly presented.[127]

There are several other cases of highly conserved Trp in downstream-clade groups, highlighted in Figure 3.19, the XIV non-elongating group feature conserved Trp at -32 and +37 positions, as well as a Gln in position +5, the latter of which may be required for reduction by the upstream KR. Both of the Trp positions appear to be surface residues rather than core residues when examining known ACP structures and with energy minimised homology models, therefore it is possible that these Trp residues may aid

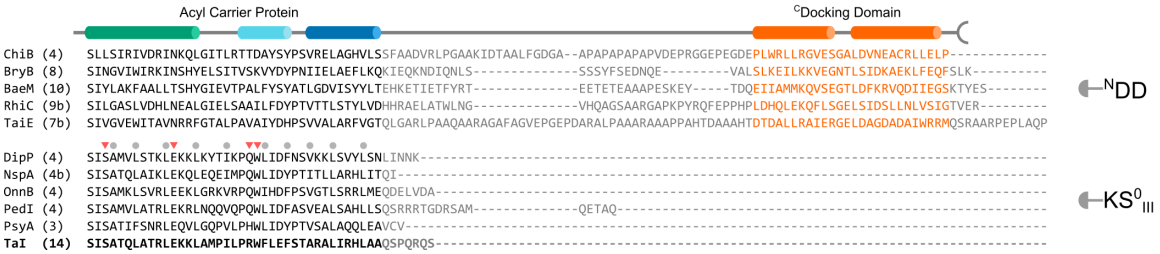


Figure 3.21 | Comparison of PKS proteins terminating in an ACP with downstream KS^0_{III} (*bottom*) to those with known docking domains (^{C}DD)[197] (*top*). The C-terminal region of PKS proteins are shown from the conserved Ser of the terminal ACP to the end of the protein. Names for the proteins are shown with ACP numbers shown in brackets. TaI is shown in bold with the KS^0_{III} downstream ACPs. Uncommonly conserved residues in these KS^0_{III} downstream ACPs are highlighted with triangles. Residues that form part of the hydrophobic core are indicated with grey circles. Sequences aligned using MUSCLE[198] v3.8.31 with default settings.

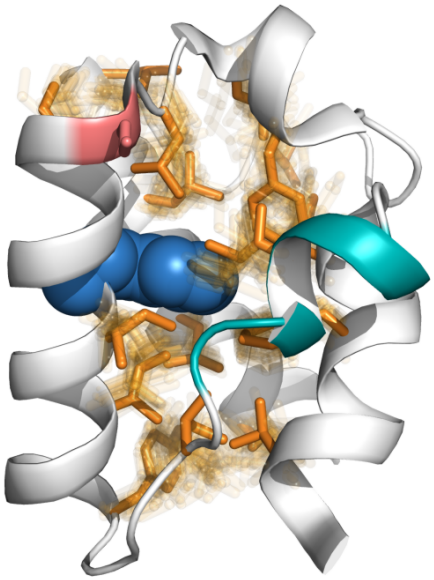


Figure 3.22 | Cartoon representation of solution NMR structure of MmpACP7a[127] (PDB: 2L22, 1–76) showing interaction of Trp+6 and hydrophobic core residues. The conserved Ser is shown as a stick (red), the conserved Trp at position +6 is shown as spheres (blue), the hydrophobic core residues as indicated in Figure 3.19 are shown as sticks (orange, all states shown from ensemble) and helix III is highlighted (cyan). Rendered in PyMol.

successful interaction with the adjacent ACP in the next PKS protein. PCPs upstream of a KS_{XVI} have a conserved Trp at -27, as well as a Gly at -2 and a Phe at +20; these residues point into the core, and may cause helix I′ to be differently positioned for interaction with the C and A domains. Other surface residues may be involved in specifying ACP–enzyme interactions: a highly conserved Asn at -8 in the II grouping might be involved in correctly presenting the β-keto-Ppant to the KR, as the β-hydroxyl product of these modules has L-stereochemistry[83] but the corresponding KRs all lack the conserved Trp for the expected A-type[140] KR, see Figure 3.23A. Likewise the conserved Gln in XIV

Figure 3.23 | Comparison of ACPs with downstream KS_{II} and KS_{XIV} and their upstream KRs. (**A**) Aligned sequences for several KRs from modules with a downstream KS_{II} and KS_{XIV}. Key motifs for determining A (blue) or B (orange) type are highlighted, and catalytic Lys, Ser and Tyr are highlighted in red. Sequences aligned using `MUSCLE`[198] v3.8.31 with default settings. (**B**) Cartoon representation of MmpACP7a (PDB: 2L22, 1–76) modified to have Asn at -7. Conserved Ser is represented as a sphere with an *. Conserved Gln and Asn residues, for XIV and II groups respectively, are shown as sticks. Helices are coloured to match Figure 3.19. Rendered in PyMol.

grouped ACPs may play the complementary role to generate the D-stereochemistry typically accepted by DH domains,[156] the corresponding KRs for these ACPs also lack the conserved LDD motif for B-type[140] KRs. As these two residues appear on opposite faces of the conserved Ser (Figure 3.23B), it is possible that these residues cause the ACP to dock to the KR with different rotations and potentially influence the resulting stereochemistry of the β-hydroxyl. Overall the apparent discrimination using this grouping is good, allowing for the identification of potentially relevant residues for explaining ACP behaviour.

The ability to discriminate using HMM profiles can be seen by using the ratio of bit scores for the grouped against all ACPs, where scores of greater than one imply better matching to the clade-grouped HMM profile than generic ACPs. Figure 3.24 and Figure 3.25 show clustered heat-maps for such scores for the downstream- and upstream-KS clade profiles respectively. The HMM profiles generated for downstream-KS clade grouped ACPs seem to be more discriminatory than those of upstream-KS clade grouped ACPs, showing more defined groups and less scattering. Interestingly, clustering the HMM profiles themselves based on these ratio scores reveals some correlation between downstream-KS grouped ACPs. In particular, ACPs with a downstream KS_{VII}, KS_{VIII}, KS_{IX} or KS_{X} have similar bit score ratios for all of these profiles. Some caution needs to be taken when making such comparisons, clade groups that are over-represented (I, IX and VIII, see Table 3.1) have a disproportionate influence on the all-ACP HMM profiles. This results in score ratios being reduced compared to under-represented groups (XI–XIV).

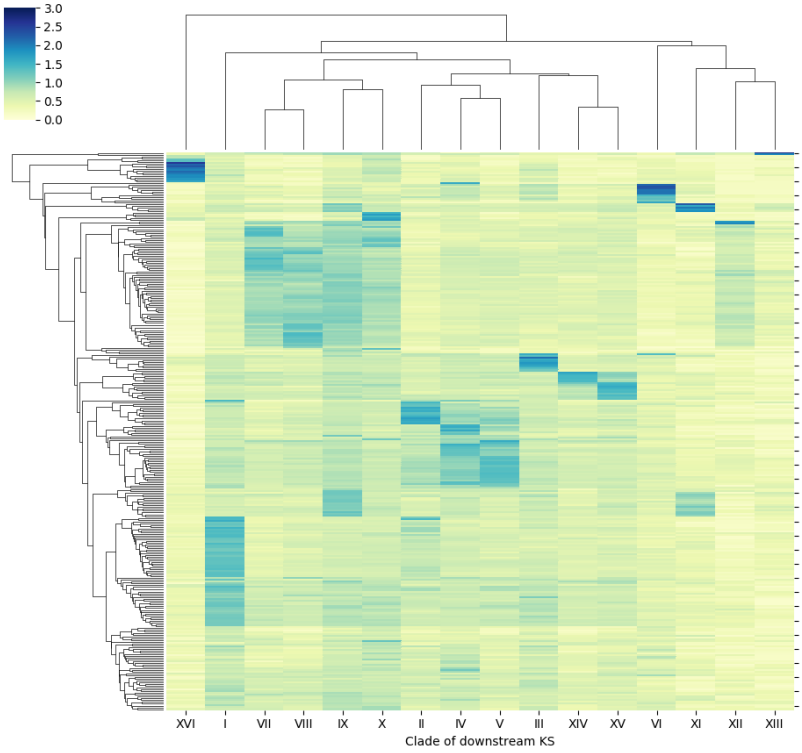


Figure 3.24 | Clustered heat-map of HMM score ratios between downstream-KS clade and all-ACP HMM profiles. Generated using the Python package `seaborn` with the `clustermap` function. Scores are clustered in both rows and columns. A value of >1 (green $\rightarrow$ blue) implies a greater match to clade assignments.

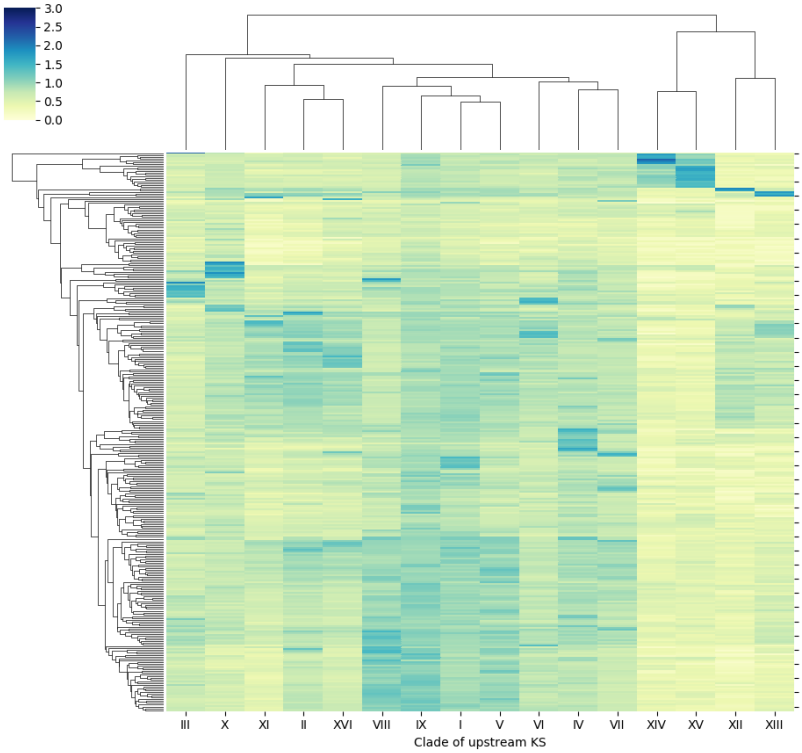


Figure 3.25 | Clustered heat-map of HMM score ratios between upstream-KS clade and all-ACP HMM profiles. Generated using the Python package `seaborn` with the `clustermap` function. Scores are clustered in both rows and columns. A value of >1 (green $\rightarrow$ blue) implies a greater match to clade assignments.

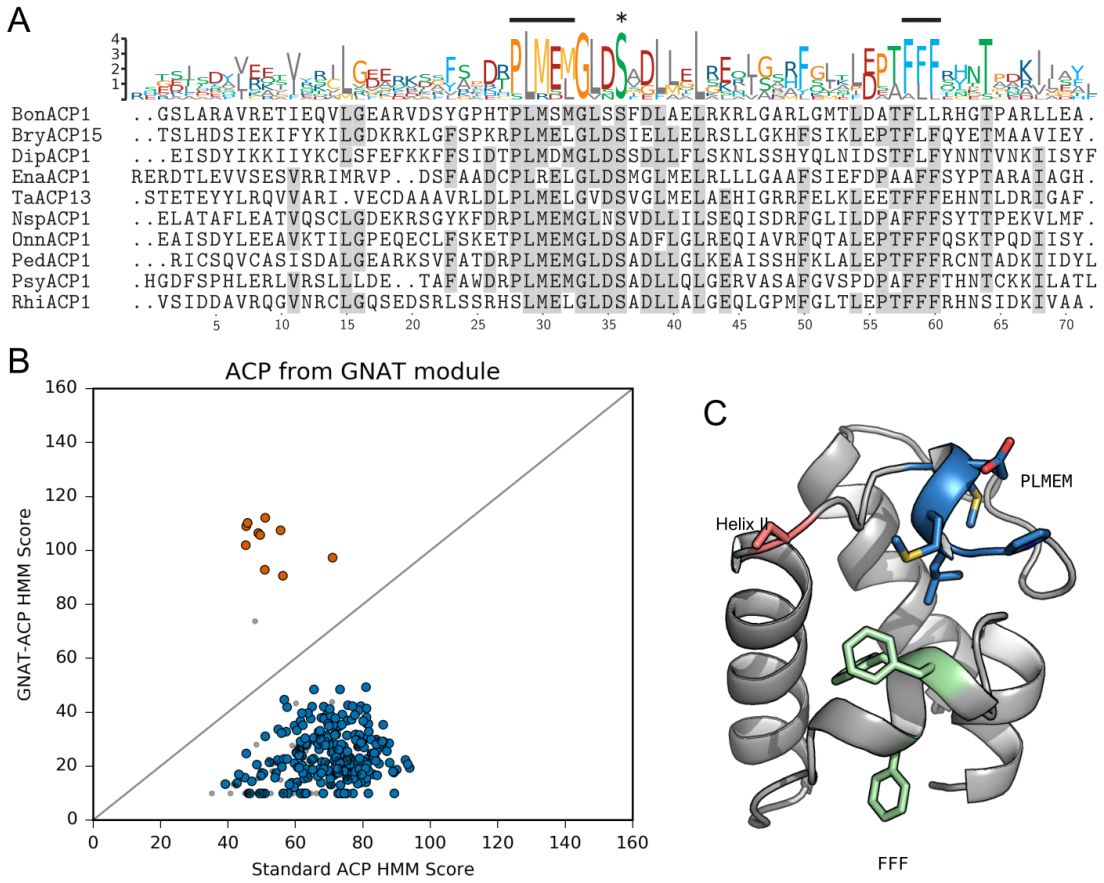


Figure 3.26 | (**A**) Multiple sequence alignment for GNAT-loaded ACPs. Sequence logo of alignment with potential interaction sites (—) and conserved serine (*) highlighted (*upper*), and the multiple sequence alignment with conservation threshold 60 % (*lower*). (**B**) Plot of bit scores for GNAT-loaded HMM against standard HMM. (**C**) Rendering of PsyACP1 homology model (PDB: 2AFD as template) with conserved serine (red), PLMEM motif (blue), and FFF motif (green) highlighted. Rendered in PyMol.

It is evident however, that KS clade is not always categorical for ACP sequence. ACPs upstream of a KS_I can in-fact be split into two groups: those that are involved in β-branching and those that are not. The key difference in sequence between these is the conserved Trp at +6, as previously identified by Haines *et al.* [127] ACPs upstream of a KS_{VI} are responsible for initiating the polyketide, typically with an acetyl or propionyl starter unit provided by a GNAT domain. When examining only those ACPs in GNAT-loading modules, the PLMEM motif N-terminal to the conserved Ser and the FFF motif in helix III predominate, see Figure 3.26A. The HMM score plot (Figure 3.26B) shows good discrimination using this profile, with a single unassigned ACP showing better matching to the GNAT-grouped HMM. This is BatACP1, an ACP that initiates with acetyl-Ppant but no GNAT or KS^Q present in the module. Within the grouping one ACP lacks the Met in the PLMEM motif, this is also the only one in this group which

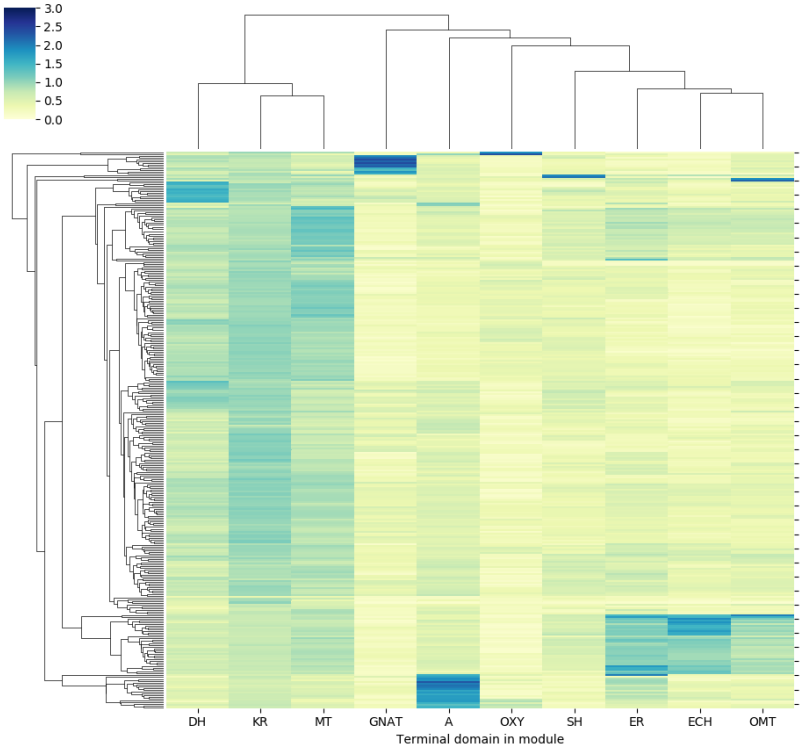


Figure 3.27 | Clustered heat-map of HMM score ratios between module terminating domain and all-ACP HMM profiles. Generated using the Python package `seaborn` with the `clustermap` function. Scores are clustered in both rows and columns. A value of >1 (green $\rightarrow$ blue) implies a greater match to clade assignments.

incorporates propionyl, not acetyl, as a starter unit. When mapped to a model of PsyACP1 (Figure 3.26C), several of these residues are surface accessible and may be involved in the interaction with the GNAT domain, or perhaps with the acyl-group itself.

To see whether ACPs could be likewise grouped using the contents of the modules rather than by KS clade, the ACPs were split into groups based on the processing domain immediately N-terminal to the ACP. This was done — rather than simply by the presence of a domain in the module — to avoid multiple contributions to, and scrambling of, the HMM profiles. The clustered heat-map of the HMM scores is shown in Figure 3.27. Most groupings show a high level of discrimination against the standard ACP background. DH, KR and MT terminal groupings, however, are less diagnostic, most likely due to the aberrant nature of *trans*-AT modules where functionally similar modules can have different configurations or *trans*-acting domains. Table 3.2 shows the occurrence of the different processing domains within the data-set, illustrating a clear problem with using the terminal domain of a module for grouping. Most KRs occur

	A	AL	B	DH	DUF	ECH	ER	GNAT	KR	MT	OMT	OXY	PS	SH
In Module	19	1	1	111	2	10	7	8	180	56	4	2	4	2
Terminal	18	—	—	19	—	9	6	8	135	49	3	2	—	2
In-*trans*	—	—	—	4	—	—	12	—	3	3	—	3	—	—

Table 3.2 | Frequency of ACPs in modules by processing domains in the module, terminating the module and *trans*-acting with the module.

as the last domain within a module regardless of the remaining modular composition, while DHs are only found as terminal domains in split modules for double bond epimerisation.

The high number of different modular configurations known (>50) would make grouping based on these types difficult. In many cases, different modular configurations lead to the same product and using downstream-KS clade to group by these is sufficient. In a few cases, the ACP–protein interactions in a module dominate, but even these are in some way related to downstream-KS clade. Overall, this approach to classifying and identifying key residues in ACPs has been successful. It is already apparent that ACPs have differing interactions with the stand-alone AT and AH of *trans*-AT PKSs (Section 3.3 and Section 3.4). It remains to be seen whether the residues identified herein are involved in these interactions, or are involved in other interactions. As these interactions are transient, it is difficult to perform experiments such as protein footprinting or NMR to determine interaction sites on either the ACPs or the enzymes. Site-directed mutagenesis of these residues may allow their impact to be indirectly measured, particularly for the hydrolysis of acyl-chains.

Computational studies of ACPs may also yield insight: the Trp+6 in β-branching ACPs has been shown to rigidify ACP structure under molecular dynamic simulations;[127] and a recent study[199] comparing type I and II PKS and type II FAS ACPs by molecular dynamics showed that the identity of G**X**DS and the residue composition of helix III dictate the sequestration of acyl-chains, where bulky residues in these areas prevent tunnel formation that protects acyl-chains from hydrolysis when being shuttled between stand-alone domains. Similar studies could be performed on a large ensemble of ACPs, such as used here, to allow the overall structural changes imbued by these conserved residues. In so far as using the HMM approach to predict the role of an ACP, or interacting *trans*-acting domains that serve modules, the results herein are promising. However, the diversity of ACPs used must be increased to provide more representative

HMM profiles. There is also the potential to pursue machine learning for identification of key features within these ACP sequences; this would likewise require a sufficiently representative training set to provide robust analyses. Furthermore, using enzymatic results as a means for grouping these ACPs would be desirable, as a direct measure and to provide a scalar, rather than binary, assignment of residue–function correlation.

3.5.2 Indirect Phylogenetic Clustering

The same set of aligned ACP sequences was clustered using a maximum likelihood phylogenetic approach, as described in Section 2.5.1. The resulting phylogram, or tree, was rooted to the *cis*-AT out-group and clades manually assigned to KS clades or interacting partner domains, shown in Figure 3.28. The strong correlation between ACP and downstream KS, rather than upstream KS, suggests possible evolutionary co-migration of ACP-KS pairs and a potential interplay between the ACP and KS for substrate selectivity. Furthermore, ACPs from GNAT-loading, β-branching and NRPS modules claded together rather than just by KS clade alone, suggesting that these interactions are mediated by specific residues on the surface. Recently, a similar analysis has been independently performed[200] on an extended set of ACP sequences (526 ACPs from 33 *trans*-AT assembly lines) and found the same trend, going further to suggest a redefinition of module boundaries in *trans*-AT PKS such that modules terminate with a KS/C. ACP module types were also defined on the basis of the moiety introduced by the module and the processing domains involved, these are signified alphabetically (*a*–*z*, where *z* represents undefined module type), shown in Figure 3.29. These results are supportive of the direct approach used in Section 3.5.1.

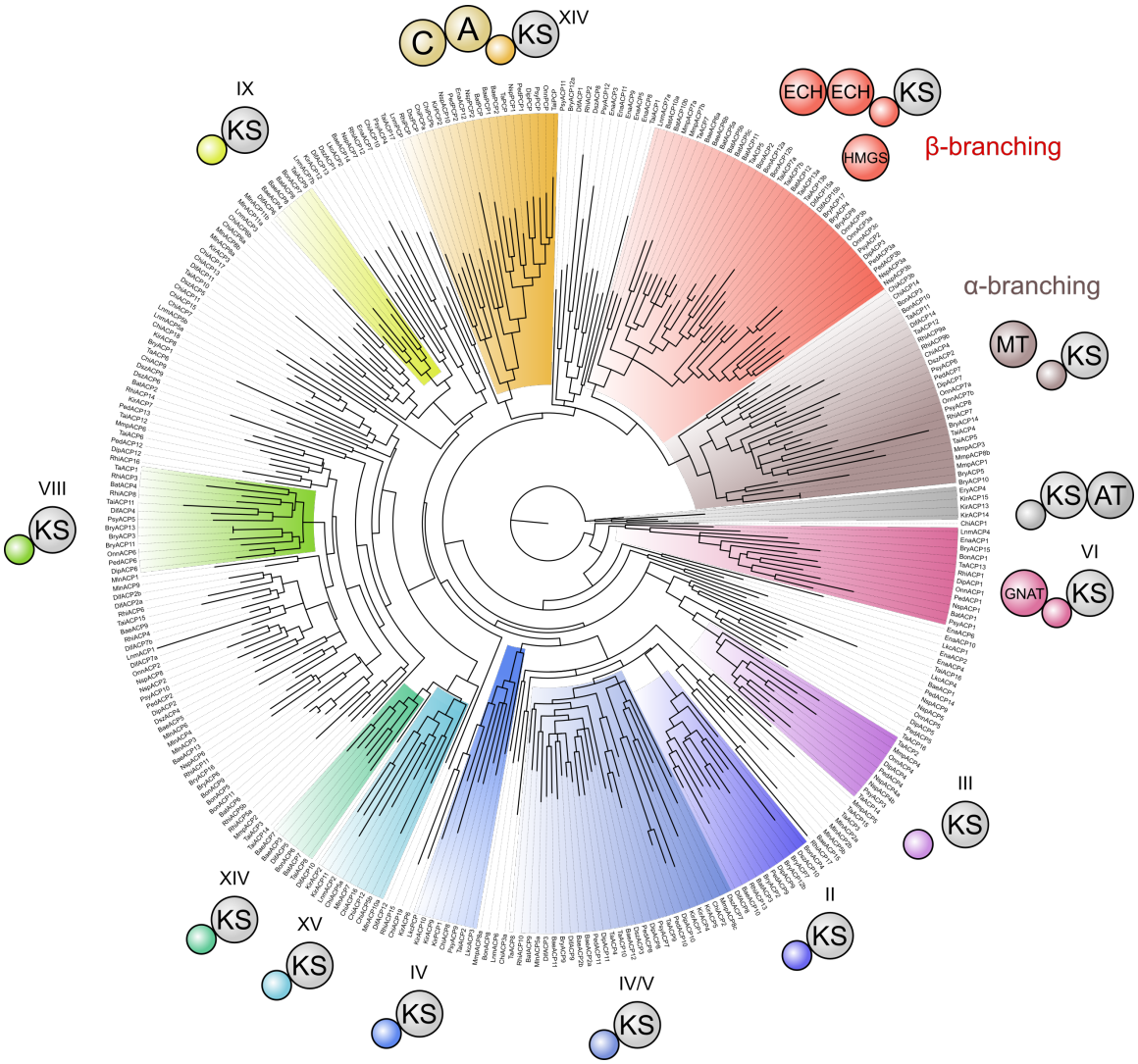


Figure 3.28 | Circular maximum likelihood phylogram of 315 ACP sequences from 22 *trans*-AT PKS. The tree was rooted to *cis*-AT ACPs and manually interpreted. Prominent clades are highlighted with downstream KS clade identified. See Table 1.1 for KS clade substrate specificity. Sequences aligned using Clustal Omega,[201] tree calculated using PhyML[178] and visualised using FigTree (github.com/rambaut/figtree) and Inkscape.

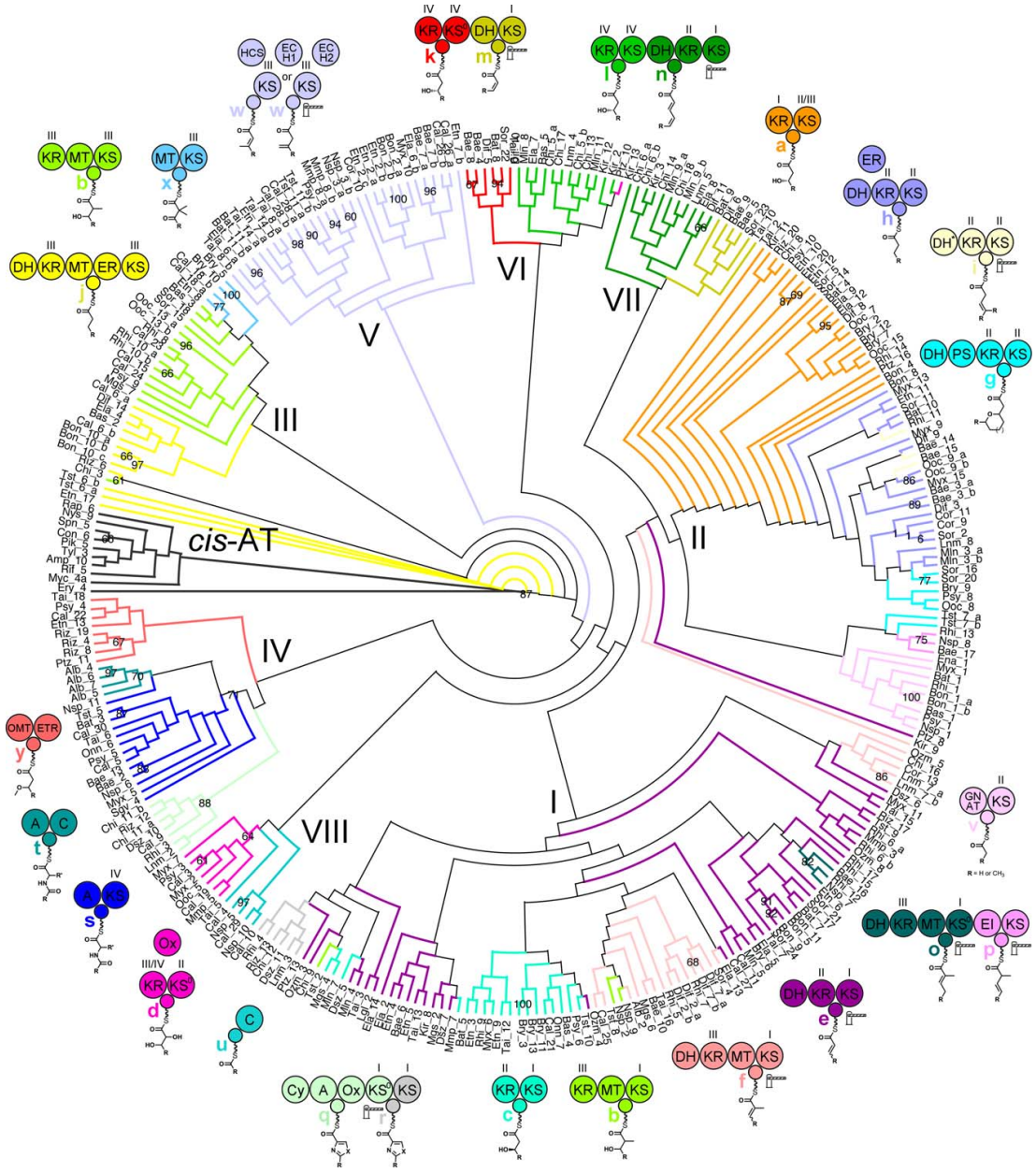


Figure 3.29 | Circular cladogram of ACPs coloured by module type. Featuring ACPs from 33 *trans*-AT PKSs and 10 ACPs from *cis*-AT PKSs. ACP modules types (*a*–*y*) are defined by moiety added as well as processing domains present. Reproduced from reference 200.

3.6 Conclusions and Outlook

Acyl carrier proteins are more diverse and influential than initially imagined. The ACPs investigated here show various interactions with different enzymes to yield the naturally preferred acyl-ACP. PsyACP1 protects smaller acyl-chains from hydrolysis by an AH (PedC) while PsyACP3 is preferentially de-acetylated, both allowing biosynthesis to proceed effectively. ACP–PedC interactions mediate these results, with non-enzymatic hydrolysis showing no disparity between the two ACPs. In the presence of an AT (PedD), PsyACP3 is readily malonated but PsyACP1 is resistant. This resistance to malonation seems to be an ACP–Ppant interaction, rather than ACP–PedD, as direct loading of malonyl-Ppant with a PPTase gave only *holo*-ACP and acetyl-PsyACP1 is successfully hydrolysed by the mutant PedD^{R97Q}. This is in contrast to other known ACPs from GNAT loading modules. PsyGNAT expressed in our hands was non-functional, a possible explanation for the resistance to malonation. There may be another unidentified mechanism for which PsyACP1 is loaded with the appropriate acetyl unit, and loading of a malonyl unit might act to stall biosynthesis, in the same way acetyl units would stall a normally elongating ACP. Regardless, these results show a clear sequence–function correlation within ACPs.

The role that ACP sequence plays in their interactions can be, in part, rationalised through the generation of HMM profiles and phylogenetic trees. ACPs can be categorised by their interactions and native acyl-intermediates, with clear trends in sequence observed. It is probable that bulky residues that point into the core of the helical bundle control the rigidity and shape of the ACP to allow specific recognition by processing domains. In some cases the residues appear on the surface of the ACP, and may be more directly involved in interactions. Future studies in this vein should aim to establish the link between these conserved residues and the final 3D structure of the ACPs. Mutagenesis experiments, both computational and biological, provide a convenient way of assessing the impact of ACP sequence on their interactions. A nuanced understanding of ACP sequence will allow synthetic biology of PKSs to be fine-tuned for the production of feed stocks and mediating difficult synthetic steps in drug production.

4

Native Mass Spectrometry of PKS Domains

4.1 Introduction

The reactions of PKS domains and modules can be readily monitored offline, typically by: releasing polyketide intermediates,[154,155,202] Ppant ejection of ACP-bound polyketide intermediates[97,112,115,156] or trapping of 'crypto'-ACP interacting domains.[99,157,159] Yet, convenient methods are still sought to allow live monitoring of reactions with the specifics of protein–protein interactions within PKSs. Native mass spectrometry (nMS) can be used to probe the interactions of complex systems[47] and to give insight into solution-phase interactions through gas-phase interactions. The coupling of ion mobility spectrometry (IMS) to nMS allows the structure of gas-phase proteins to be determined, especially in the case of conformational shifts upon change of protein state. IMS has been exploited to mimic thermal-melt type experiments through collision-induced unfolding (CIU),[41,57–59] where a gas-phase ion is subjected to increasing collisional energies to cause unfolding of the ion. Comparing between variants of a protein — either mutants, covalent chemical changes or non-covalent ligands — can allow the effective stabilisation of highly homologous proteins to be measured.[63,64]

4.2 Native Mass Spectrometry of PKS Proteins

PsyACP1 and PsyKS1 were recombinantly expressed in *E. coli* as His-tag fusion proteins, followed by purification with IMAC as described in Section 2.2. PsyACP-KS1 and PksM3 were recombinantly expressed and purified by Dr Matthew Jenner (University of Warwick). Proteins were prepared for nMS by exchanging into a compatible volatile solution of AmAc, as described in Section 2.3.4. Typically, final concentrations of 25 mM AmAc for PsyACP1 and 200 mM AmAc for PsyKS1, PsyACP-KS1 and PksM3 were used; with all proteins at ~10 µM final concentration. However, PsyACP-KS1 was found to precipitate in 200 mM AmAc when on ice and was therefore initially exchanged into 25 mM AmAc before dilution into 200 mM AmAc immediately before loading into the nESI tip.

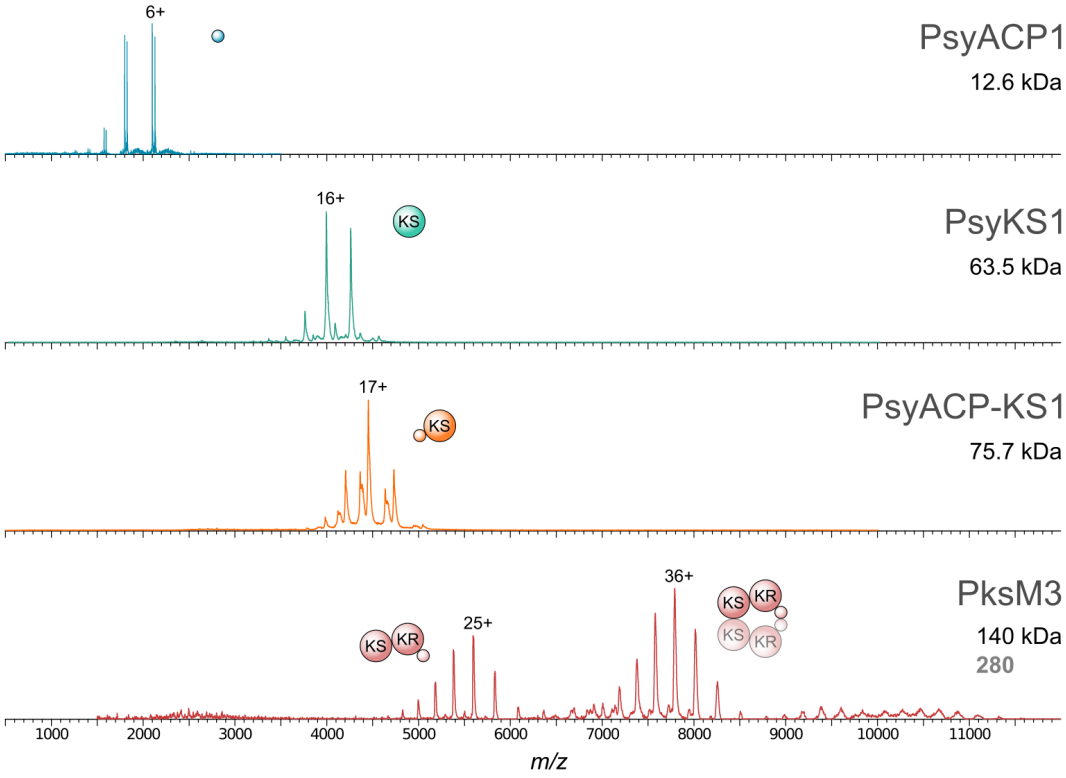


Figure 4.1 | Stacked native mass spectra from nESI-MS of PsyACP1, PsyKS1, PsyACP-KS1 and PksM3, all with His-tags attached sprayed from 25 mM or 200 mM AmAc. The experimental mass calculated from these spectra shown to 3 s.f. and the most intense charge states are indicated. The configuration of each protein is shown.

Spectra were typically acquired using nESI tips with a back-fitted platinum wire electrode, as described in Section 2.3.5. Example nMS spectra are shown for these four proteins in Figure 4.1. Under native conditions these proteins showed experimental

masses of 12 587, 63 541, 75 690 and 139 826 Da as monomers. PksM3 also showed significant dimer (280 395 Da).

4.3 Collision Induced Unfolding of Acyl-PsyACP1

A set of acyl derivatives were generated for PsyACP1, using the method described in Section 2.4.1. For the purposes of this comparison: *apo*, *holo*, acetyl and octanoyl states were chosen. The His-tag was cleaved using thrombin to provide more native-like IMS, as the flexible nature of this tag can provide unwanted conformations that might mask subtle changes in mobility. CIU data for the *holo*-, acetyl- and octanoyl-ACP Ppant variants were recorded with 1:1 *apo*-ACP, to provide an internal control between acquisitions and account for any variations in temperature or pressure within the instrument. All collision voltages (V_a) are converted to E_{lab} ($z \times V_a$, where z is the charge state of an ion) units for better comparison of the collision energies used.

4.3.1 Stabilisation by Ppant

CIU experiments on the 50:50 *apo*:*holo*-PsyACP1 mixture shows stabilisation by the Ppant chain in both the 6+ and 7+ charge states examined. For the apo^{6+} there are two conformers, a compact folded population centring at at an arrival time of 5.5 ms and an extended unfolded form centring at an arrival time of 7.5 ms as shown in Figure 4.2A. The $holo^{6+}$ exhibits the same two populations with a shift in arrival time caused by the additional mass of the Ppant. These two profiles can be compared by subtracting the *holo* profile from the *apo* profile, creating a difference plot as shown in Figure 4.2B, where areas of blue indicate a positive difference between *apo* and *holo* (*apo* > *holo*) and *vice versa*. The overall difference between the two data sets at each energy can be compared using a root-mean-squared deviation (RMSD), which can be used to represent the extent of change as a function of collision energy (Figure 4.2B), this value can be averaged to give an RMSD for the entire difference plot. For the 6+ charge state, this shows a steady and consistent difference between the *apo* and *holo*. Extracting the ATD for both profiles at 93 eV, the collision energy of this transition point, shows that the *holo*-PsyACP1 is more folded at this collision energy than the *apo*, indicating stabilisation caused by the Ppant chain (Figure 4.2C).

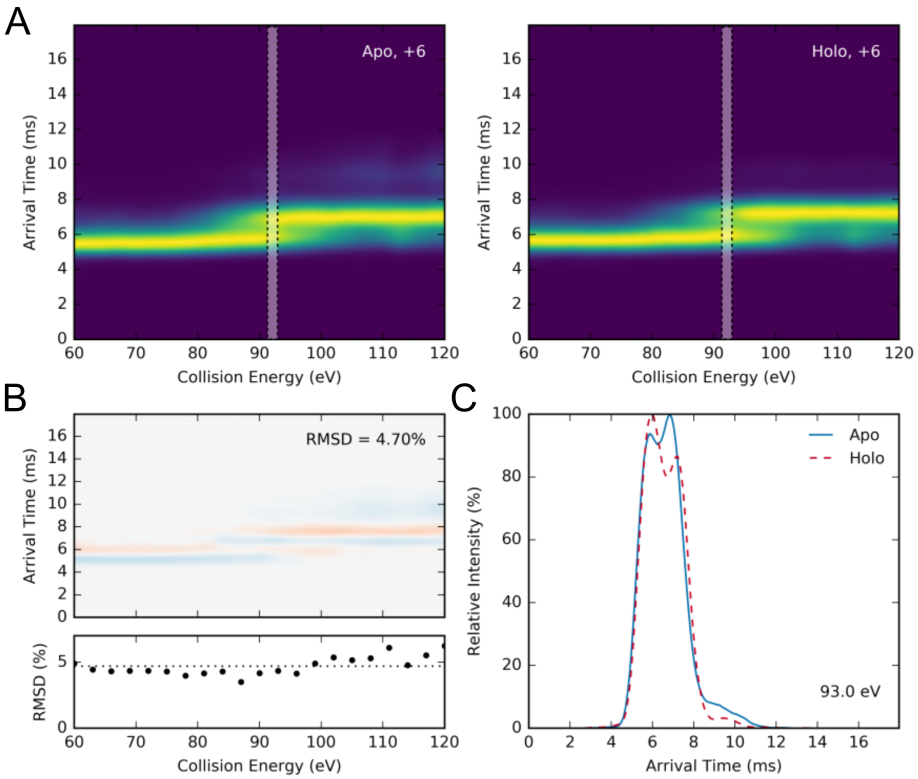


Figure 4.2 | (**A**) CIU profiles of *apo*- and *holo*-PsyACP1^{6+} with highlighted region indicating the 93 eV shown in C. (**B**) Difference plot of the two profiles (*apo* − *holo*) and RMSD plot by collision energy. (**C**) Overlaid ATD of *apo*- (blue, –) and *holo*- (red, ···) PsyACP1^{6+} at 93 eV.

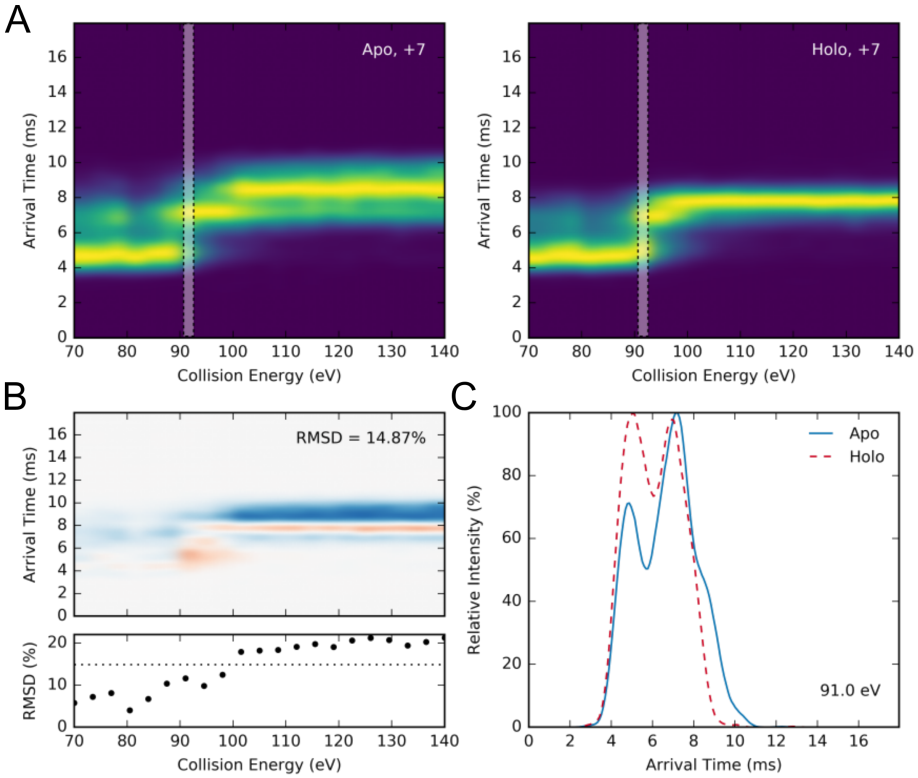


Figure 4.3 | (**A**) CIU profiles of *apo*- and *holo*-PsyACP1^{7+} with highlighted region indicating the 91 eV shown in C. (**B**) Difference plot of the two profiles (*apo* − *holo*) and RMSD plot by collision energy. (**C**) Overlaid ATD of *apo*- (blue, –) and *holo*- (red, ···) PsyACP1^{7+} at 91 eV.

This can be further seen by examining the 7+ charge state of the mixture. Here three populations are observed for the *apo*: a compact folded conformer at 5 ms, intermediate transition state at 7 ms and the extended unfolded conformer at 9 ms (Figure 4.3A). Notably, the unfolded population for the $holo^{7+}$ appears to be more compact and lower in arrival time than for the apo^{7+}, this is especially apparent in the difference plot where the rolling RMSD is much higher after 100 eV as indicated in Figure 4.3B by the wide blue band. One explanation for this phenomenon is the unfolded conformer of the *apo* is a set of unresolved conformers at least some of which do not appear in the *holo* due to interactions with the Ppant chain. Extracting the ATDs at 91 eV, the collision energy of the first transition point, shows that the first transition is appreciably stabilised by the addition of the Ppant chain.

4.3.2 Stabilisation by Acyl-Groups

To determine whether *holo*, acetyl and octanoyl datasets could be directly compared, the above analysis was performed on the *apo* component of the *holo* and acetyl datasets. For the 6+ charge state it is clear that the two profiles are within standard error (RMSD=1.48 %) as shown in Figure 4.4. It is therefore reasonable to directly compare the different acyl chains. There is a compaction at 114 eV which is present in both samples, at this energy the unfolded population shifts towards the more compact arrival times. This is also seen in the 7+ charge state data (Figure 4.5), although it occurs at 80.5 eV instead. The origin of this phenomenon is yet to be determined, although work by Hall *et al.* [203] showed that gas-phase proteins exhibit compaction due to gas-phase collapse, previously prevented by conformation.

While comparing the CIU fingerprint visually can be informative for major differences, it is insufficient for subtly; as can be seen in Figure 4.6, there are little to no apparent differences with the exception of $octanoyl^{6+}$, which has a delayed onset for the unfolded conformer. When looking at ATD extracted for specific collision energies, however, the differences between the *holo* and acyl groups is clear, see Figure 4.7 and Figure 4.8. Acetyl-Ppant is very similar to *holo*-Ppant with only slight stabilisation, most noticeable in the 7+ charge state at 91 eV. Octanoyl-Ppant provides a greater level of stabilisation, visible at 87, 93 and 99 eV in the 6+ charge state and at 98 and 105 eV in

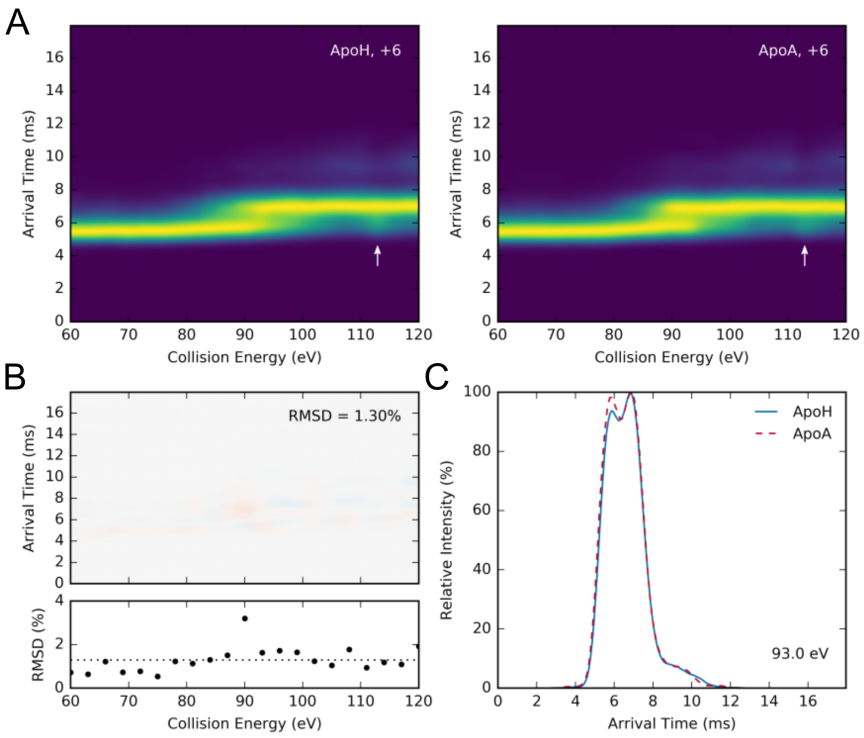


Figure 4.4 | (**A**) CIU profiles of *apo*-PsyACP1^{6+} from the *holo* (ApoH) and acetyl (ApoA) datasets, conserved compaction indicated by arrow. (**B**) Difference plot of the two profiles (ApoH − ApoA) and RMSD plot by collision energy. (**C**) Overlaid ATD of ApoH (blue, –) and ApoA (red, ···) at 93 eV.

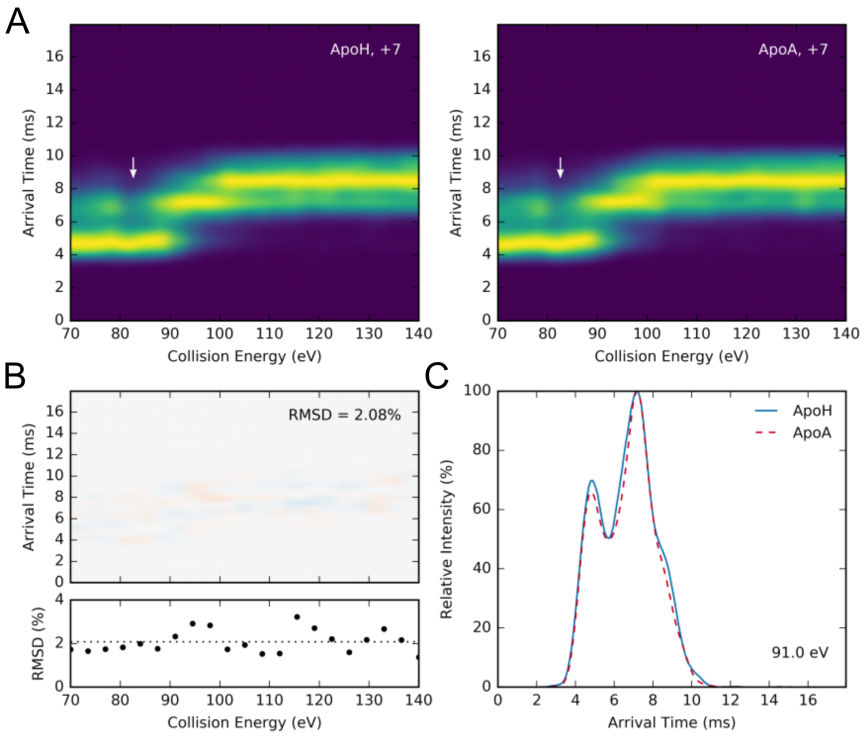


Figure 4.5 | (**A**) CIU profiles of *apo*-PsyACP1^{7+} from the *holo* (ApoH) and acetyl (ApoA) datasets, conserved compaction indicated by arrow. (**B**) Difference plot of the two profiles (ApoH − ApoA) and RMSD plot by collision energy. (**C**) Overlaid ATD of ApoH (blue, –) and ApoA (red, ···) at 91 eV.

the 7+ charge state. A compact conformer is preserved in the octanoyl^{7+}, even at higher energies, which is visible in the CIU profile shown in Figure 4.6.

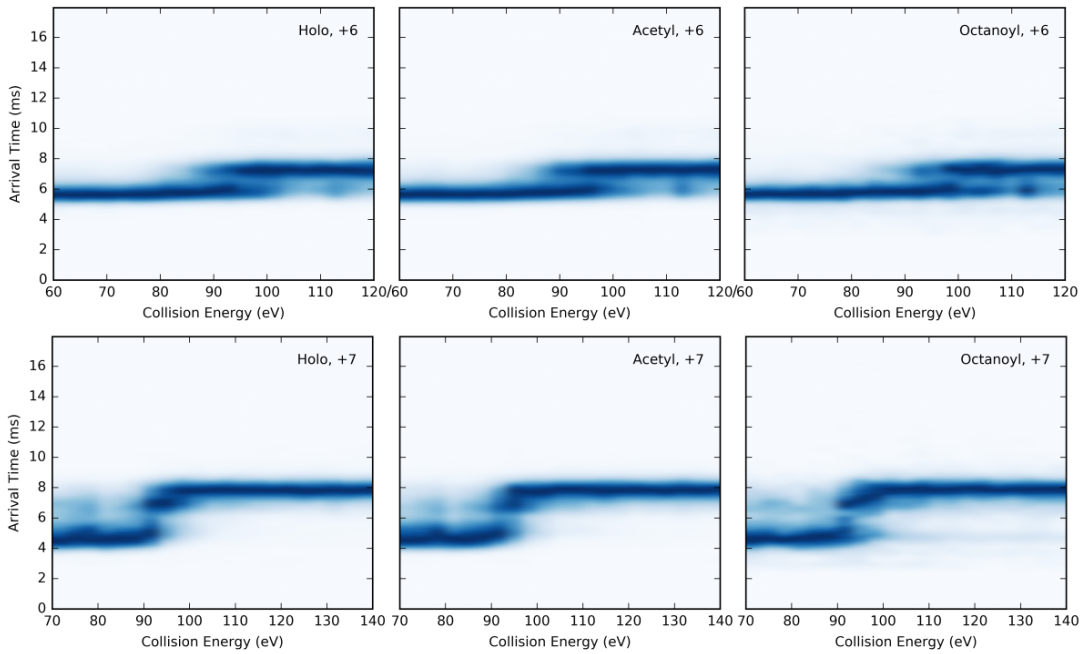


Figure 4.6 | Grid of CIU fingerprints for both 6+ (*top*) and 7+ (*bottom*) charge states of *holo* (*left*), acetyl (*middle*) and octanoyl (*right*) -PsyACP1. Visual inspection shows little change between *holo* and acetyl.

Quantifying stabilisation was attempted by calculating the weighted-average arrival time or centroid arrival time at a given collision energy, in this way the ATD is abstracted into a single number.[57] Once these centroid times were calculated, a four-parameter (x_0, k, a and c) logistic sigmoid curve was fitted in Python using the `scipy.curvefit` module, as described in Section 2.3.6.1, the fitted plots are shown in Figure 4.9. The midpoint parameter (x_0) of this curve was taken to be the value for E_{C50} — the collision energy at which 50 % of the analyte is unfolded, relative changes in this energy (ΔE_{C50}) would indicate (de)stabilisation. The c and $a + c$ values were taken to be the centroid arrival time for the folded state (C_{fold}) and centroid arrival time for the unfolded state (C_{unfold}) respectively. These values calculated from the weighted arrival times are shown in Table 4.1. For the 6+ charge state, the addition of the Ppant chain causes the E_{C50} to decrease (−2.2 eV), with subsequent chain length increases also showing an increase in E_{C50}. In contrast, the Ppant increases the E_{C50} for the 7+ charge state (0.6 eV) and increasing length further stabilises. For all *holo*/acyl^{7+} variants, the C_{fold} drops to 5.4 ms. Surprisingly, the C_{unfold} was found to decrease with increasing

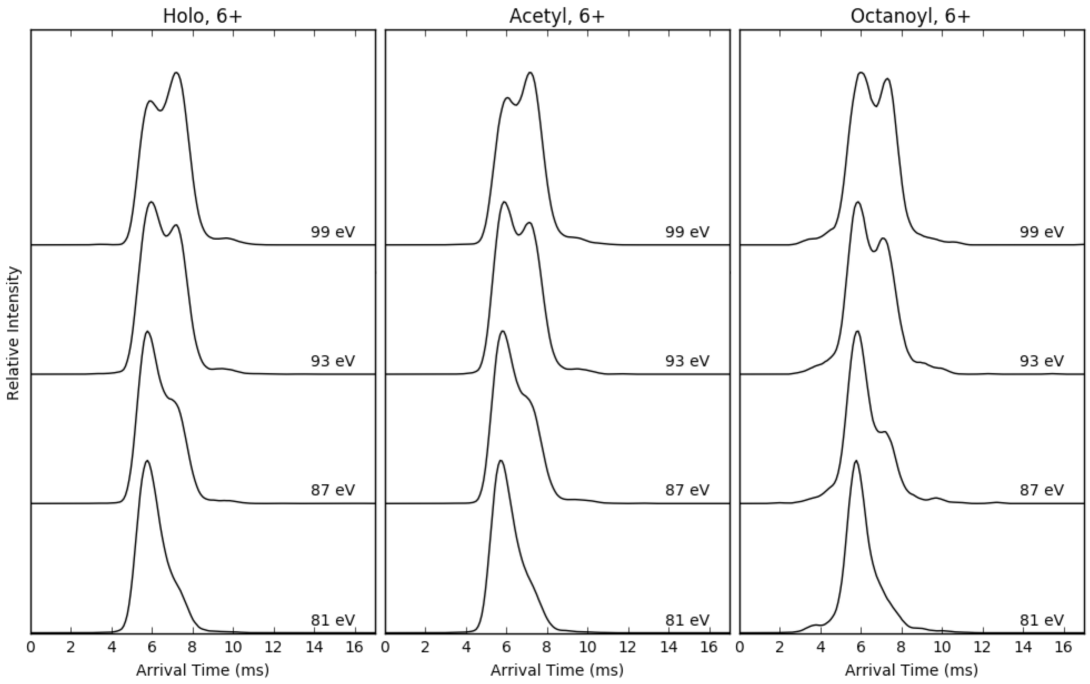


Figure 4.7 | Stacked ATD of *holo* (*left*), acetyl (*middle*) and octanoyl (*right*) -PsyACP1 for the 6+ charge state at 81, 87, 93 and 99 eV.

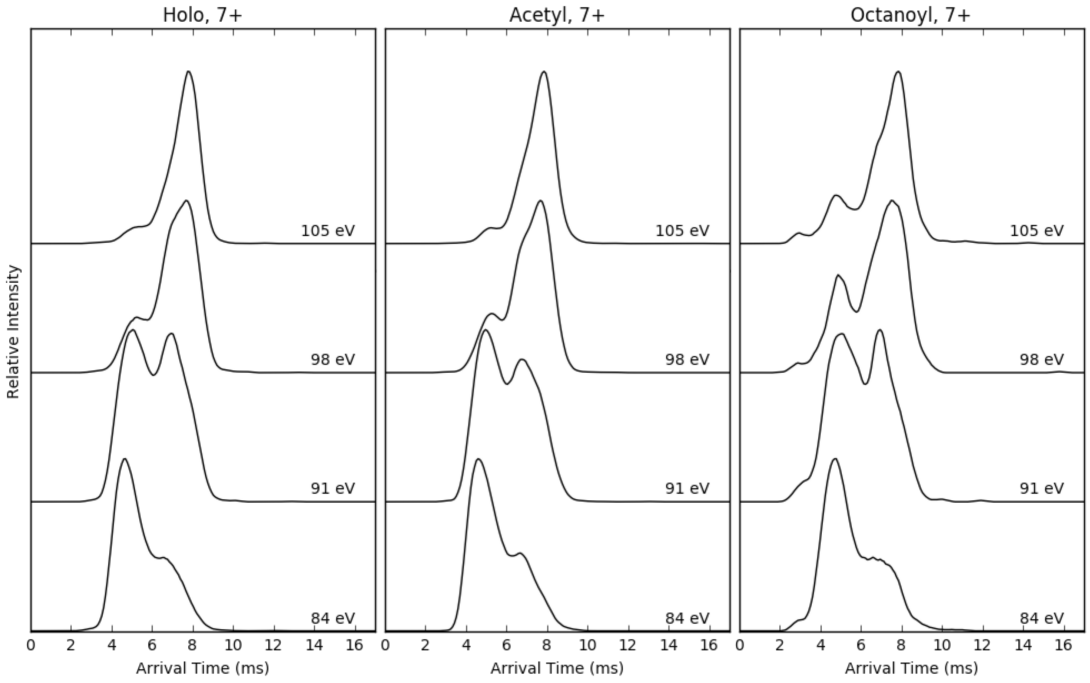


Figure 4.8 | Stacked ATD of *holo* (*left*), acetyl (*middle*) and octanoyl (*right*) -PsyACP1 for the 7+ charge state at 84, 91, 98 and 105 eV.

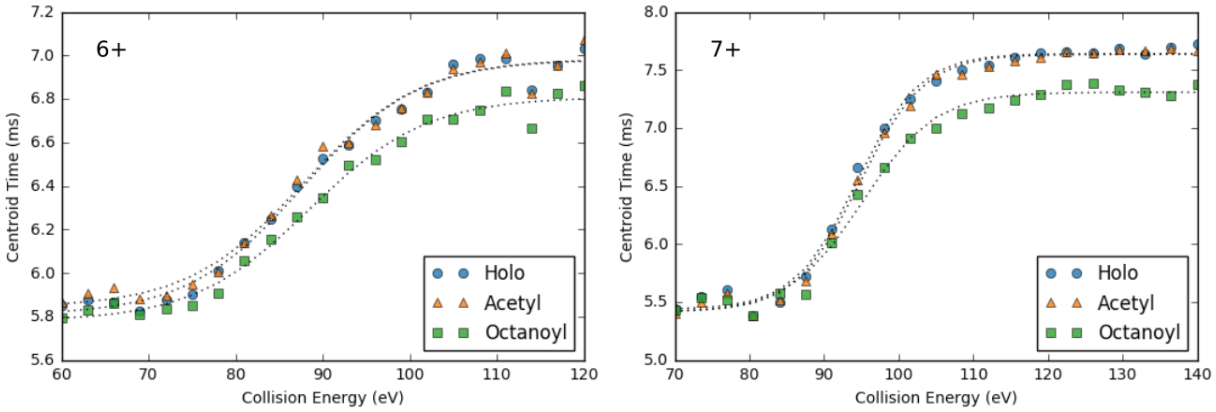


Figure 4.9 | Weighted average arrival times for *holo* (blue ●), acetyl (orange ▲) and octanoyl (green ■) -PsyACP1 for the 6+ charge state (*left*) and 7+ charge state (*right*).

chain length for both 6+ and 7+ charge states, where a larger unfolded state would be expected because of increased mass and potential size from the 18 Å Ppant arm. This would imply that the unfolded conformer is compacted, perhaps by stronger van der Waals interactions between the acyl chain and the hydrophobic 'core' residues, which would now be surface exposed in the gas-phase.

The use of E_{lab} as a measure for energy transfer is limited, as described in Section 1.1.3.2, the masses of the two colliding bodies attenuate the energy converted. When accounting for the mass of these acyl groups to calculate E_{com} (Equation 1.23) — the maximum energy transferred to internal energy during elastic collisions — these E^*_{C50} values show a decrease in stability with similar values for all *holo*/acyl-Ppant. It should be noted, however, that these corrections do not account for number of collisions, which is dependent upon the collisional cross-section.[204,205] It remains to be seen whether these subtle stabilising effects are a result of specific interactions between the Ppant and the ACP with some biological relevance, or more likely, general gas-phase phenomena. The covalent nature of this system might allow for the gas-phase effects to be studied at higher energies, which would otherwise cause dissociation of a non-covalent ligand.

	6+					7+				
	E_{C50}	C_{fold}	C_{unfold}	ΔE_{C50}	ΔE^*_{C50}	E_{C50}	C_{fold}	C_{unfold}	ΔE_{C50}	ΔE^*_{C50}
Apo	89.9	5.7	7.3	—	—	93.8	5.7	8.2	—	—
Holo	87.7	5.8	7.0	-2.2	-0.018	94.4	5.4	7.6	0.6	-0.008
Acetyl	87.8	5.8	7.0	-2.1	-0.019	94.8	5.4	7.6	1.0	-0.008
Octanoyl	88.5	5.8	6.8	-1.4	-0.019	95.1	5.4	7.3	1.3	-0.010

Table 4.1 | Parameters from sigmoid fitting shown in Figure 4.9. C_{fold} is the centroid arrival time for the folded state, C_{unfold} is the centroid arrival time for the unfolded state and ΔE_{C50} is the difference in E_{C50} compared to *apo*.

4.3.3 Collision Induced Unfolding of Larger PKS Proteins

The CIU profiles of larger PKS proteins such as PsyKS1 (Figure 4.10) and PsyACP-KS1 (Figure 4.11) can reveal insight into their complex structure. PsyKS1 shows four possible features: the folded and unfolded states, as well as two intermediate unfolded states; while PsyACP-KS1 shows six such features. Between the two proteins' profiles there is a high degree of similarity, both sharing a small intermediate conformer at low energies, as well as some coalescing conformers at high energies. The number of discrete features in CIU profiles correlates with the number of domains within a protein.[206] The high number of features, albeit short-lived, would suggest a more complex set of sub-domains that sequentially unfold. The addition of the PsyACP domain in PsyACP-KS1, is likely responsible for the increase in conformers, with the two-three conformers observed in Section 4.3.1 adding to those of PsyKS1.

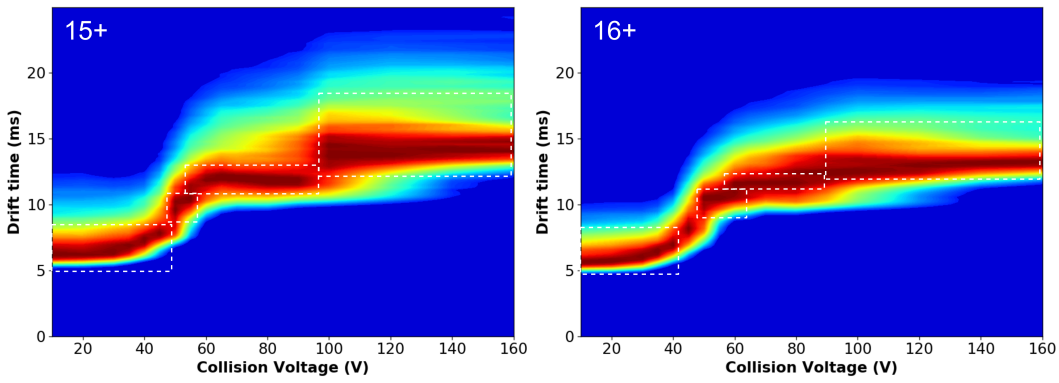


Figure 4.10 | Collision induced unfolding profiles of PsyKS1 15+ (*left*) and 16+ (*right*) charge states. Features are highlighted with dashed white boxes.

Identification of the features provided by unfolding of the KS and ACP respectively could allow the interplay between these domains to be probed under various acyl conditions. Unfortunately, CCS is a non-additive property, as atoms can be occluded

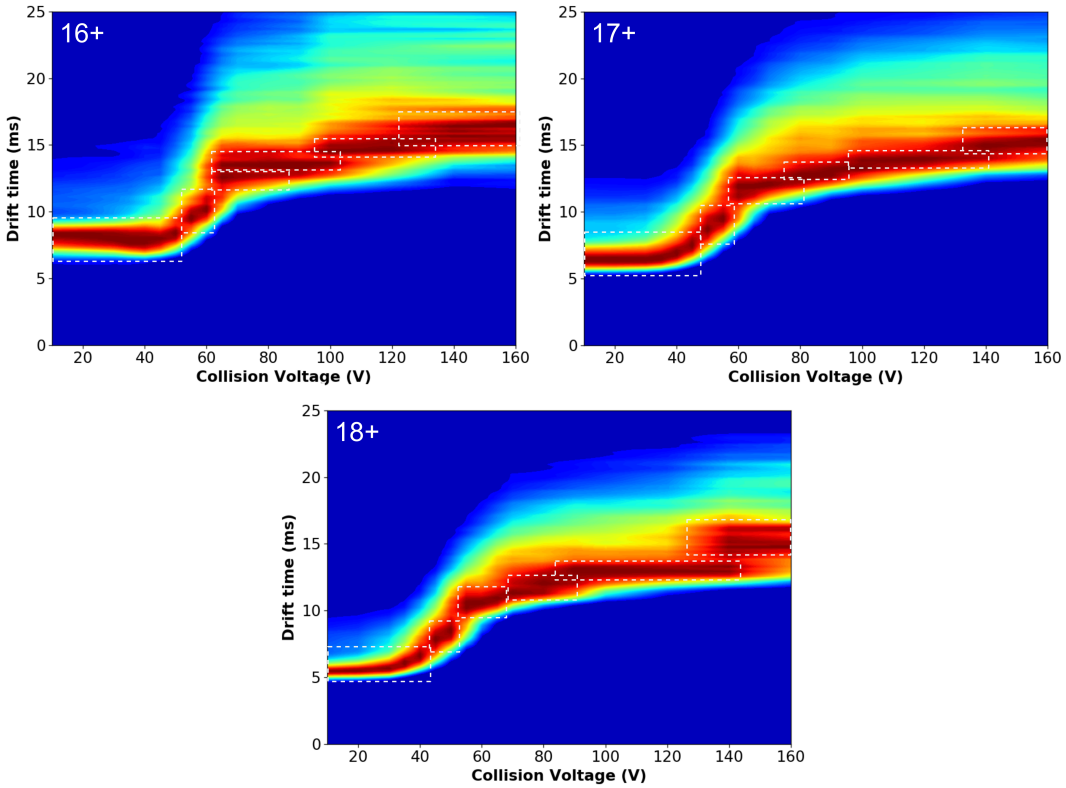


Figure 4.11 | Collision induced unfolding profiles of PsyACP-KS1 16+ (*left*), 17+ (*right*) and 18+ (*bottom*) charge states. Features are highlighted with dashed white boxes.

from certain directions.[168] Even super-coarse grained approaches, where two proteins are treated as spheres with a radius related to their CCS, are limited by interpenetration and overlap.[167] Applying such an approximation (as described in Section 2.3.6.2) to the above data, along with data acquired under identical conditions for PsyACP1, a reasonable approximation to the weighted average CCS (CCS_{avg}) of PsyACP-KS1 can be made (Figure 4.12). While there have been several developments in modelling quaternary structure to fit CCS with coarse-grained integrative modelling,[207,208] to date these have not been applied to CIU data, where the CCS are dynamic.

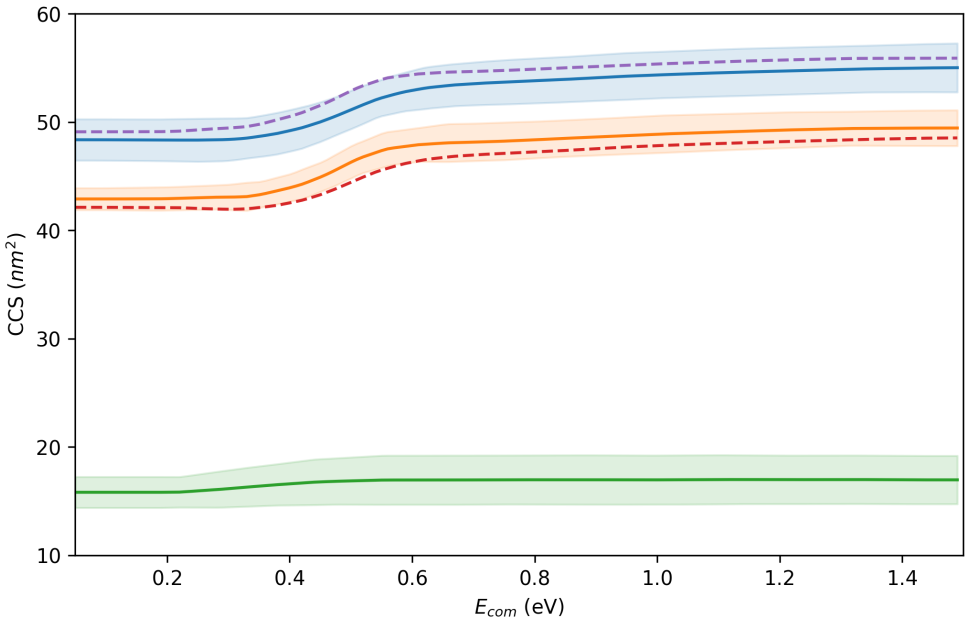


Figure 4.12 | Weighted average CCS (CCS_{avg}) for PsyACP1 (green), PsyKS1 (orange) and PsyACP-KS1 (blue) as a function of E_{com}. The solid lines represent the mean CCS_{avg} for experimental data across available charge states and shaded areas represent the standard deviation. Predicted (see Section 2.3.6.2) CCS_{avg} for PsyACP-KS1 from PsyACP1 and PsyKS1 data (purple ···); and PsyKS1 from PsyACP-KS1 and PsyACP1 data (red ···) are within the standard deviation of the experimental data.

4.4 Native Ppant Ejection

The charge state and size of native PKS modules makes identification of the polyketide intermediate impossible — the mass difference between malonyl (45.02 Da), acetoacetyl (43.05 Da) and β-hydroxyl (45.06 Da) states is 0.1 *m/z* for a 20+ charge state ion. Ppant ejection provides smaller singly-charged ions for accurate identification of acyl state.[115] By coupling Ppant ejection to nMS of PKS modules, it would be possible to monitor the flux of these modules in real time, with information about discrete steps in elongation and reductive processing.

4.4.1 Ppant Ejection from PsyACP-KS1

To see whether Ppant ejection from modules is viable, a minimal modular configuration, ACP-KS, was chosen. PsyACP-KS1 was loaded with CoA by Svp to generate *holo*-PsyACP-KS1. This was then exchanged into 25 mM AmAc as described in Section 2.3.4, before 4-fold dilution into 200 mM AmAc. Samples (9 µM) were sprayed by nESI. The 17+ charge state (4397 *m/z*) was isolated using the high-mass modified quadrupole and collided with SF_6 gas at increasing collision voltages (20–160 V). SF_6 gas was used in place of Ar gas to allow for greater energies to be transferred during collision, as massive molecules have reduced E_{com} (see Equation 1.23), thus the greatest potential for Ppant ejection. Shown in Figure 4.13A, no ejected $Pant^+$ ions are observed at higher voltages with the precursor remaining largely intact until 140 V, where significant degradation of the spectrum can be seen. There is also no corresponding appearance of apo^{17+}+HPO_3^- (Figure 4.13B) but instead just charge stripping to *holo*-PsyACP-KS1^{16+}.

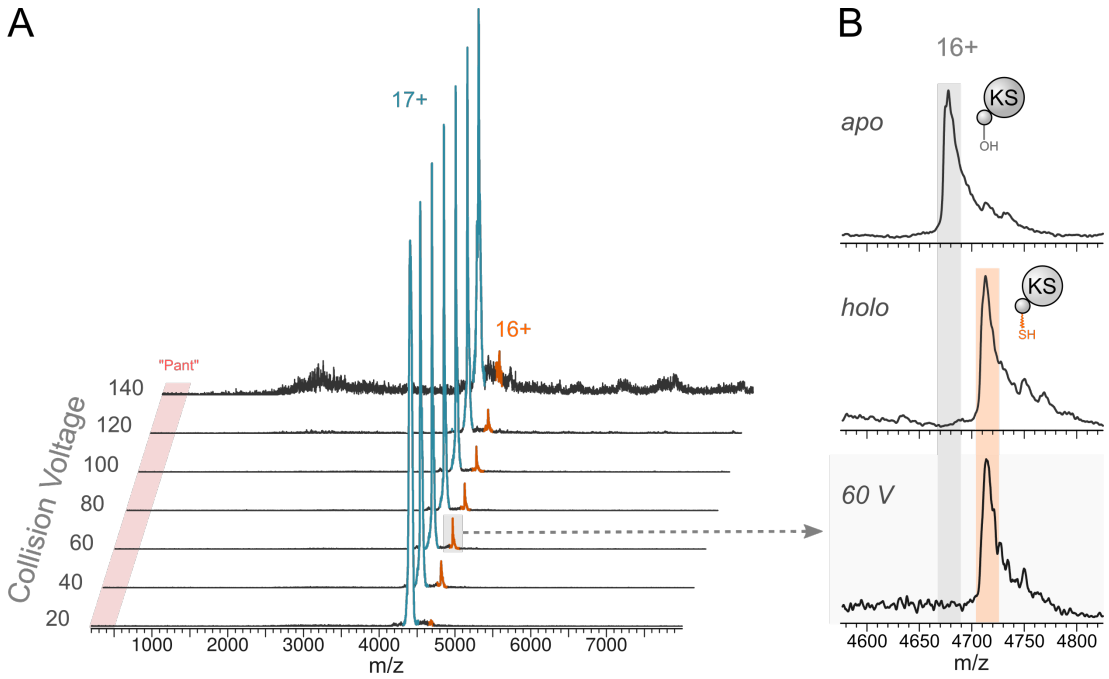


Figure 4.13 | (**A**) Stacked nESI-MS/MS spectra of selected PsyACP-KS1 17+ charge state with collision voltages 20–140 V. Key *m/z* regions are highlighted for precursor (blue), $Pant^+$ (red) and corresponding *apo*+HPO_3^-. (**B**) Stacked nESI-MS spectra for 16+ charge state region with *apo* and *holo* examples. Zoomed region from (A) at 60 V shows charge stripping and not ejected ions (*bottom*).

4.4.2 Native vs Denatured Ppant Ejection of PsyACP1

To understand whether the native gas-phase structure of an ACP affects Ppant ejection, the 7+ charge state of *holo*-PsyACP1 was isolated using a quadrupole and ejected under native (25 mM AmAc), acidified (25 mM AmAc, 0.1 % FA) and organic (60 % MeCN, 0.1 % FA) conditions, as described in Section 2.3.7.2. With the charge as a common factor, only the gas-phase structure would impact the ejection. Ppant ejection spectra are shown for all three conditions at 75 V collision voltage in Figure 4.14A. For the 7+ charge state under all conditions, the most abundant product ion is not the $Pant^+$ (261.2 *m/z*), but instead an ion at 432.3 *m/z*. The relative abundances of these two product ions and the *holo* precursor (Figure 4.14B) are consistent between the conditions, suggesting that the Ppant is not stabilised by structural features removed upon denaturation.

Upon top-down fragmentation of the apo^{7+} charge state (Figure 4.15), it is clear that the 432.3 *m/z* ion does not correspond to Ppant ejection but instead the y_4 product ion, corresponding to the cleavage of a labile Ala–Pro bond. Likewise the 1219.2 *m/z* peak corresponds to the $y_{22}^{+\cdot+}$ product ion, a labile Asp–Pro bond. The abundance of

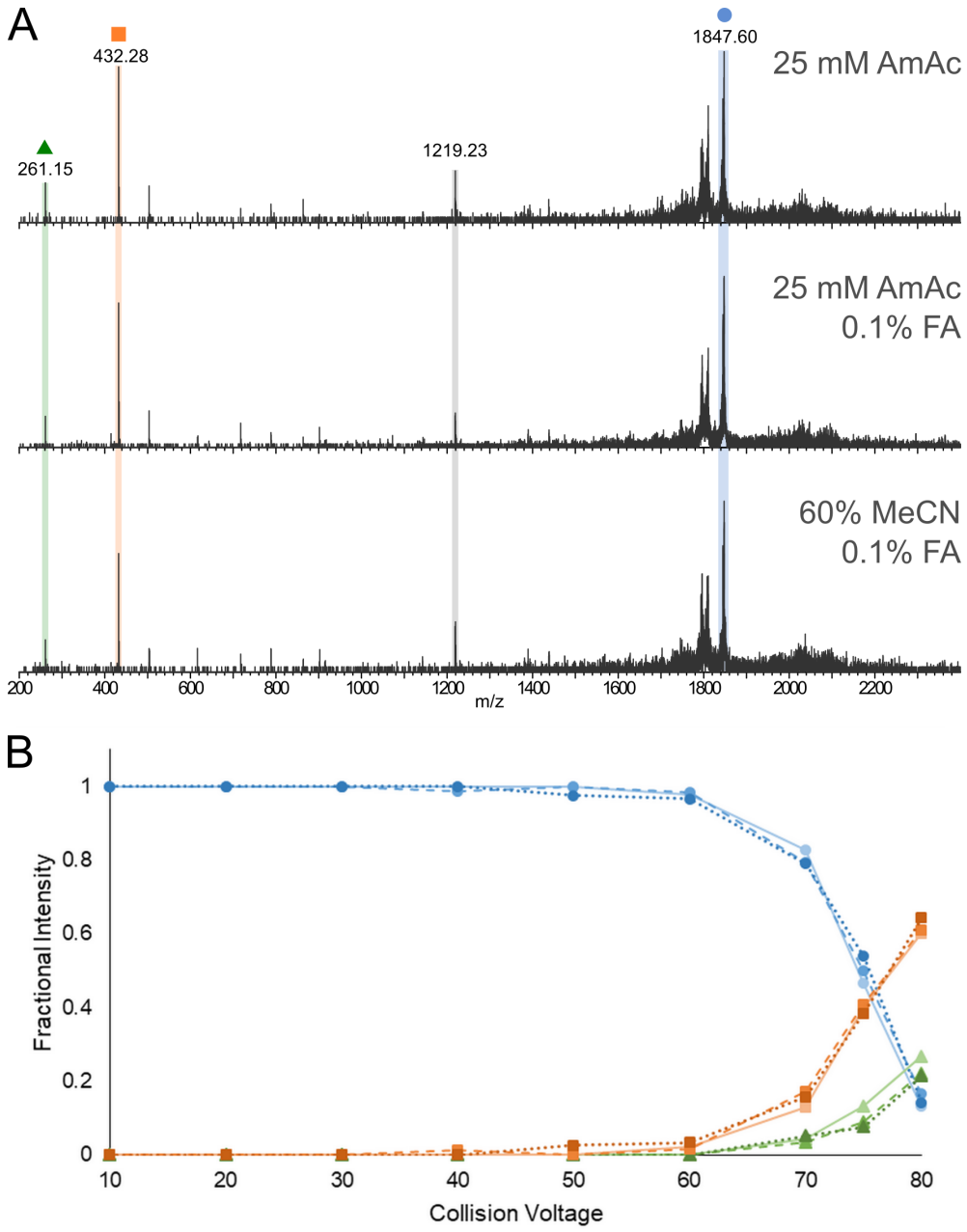


Figure 4.14 | (**A**) Stacked nESI-MS/MS spectra of selected PsyACP1 7+ charge state under native-like, acidified and organic conditions at a collision voltage of 75 V. Spectra are minimally smoothed (Savitsky-Golay, 2×5). Key ions are highlighted. (**B**) Fractional intensity of 1847.6 *m/z* (blue, ●), 261.2 *m/z* (green, ▲) and 432.3 *m/z* (orange, ■) ions from Ppant ejection spectra with 10–80 V accelerating voltage for native-like (−), acidified (┄) and organic (⋯) conditions.

these product ions over the $Pant^+$ ion would suggest that there is some stabilisation of the phosphoester bond by the ACP, that is not affected by the structural changes caused under acidified and organic solutions.

When comparing the Ppant ejection spectra of the 7+ charge state to those of the 13+ charge state (Figure 4.16) these peptide fragment ions are greatly diminished compared to the $Pant^+$ ion. This suggests that the charge state may be playing a role

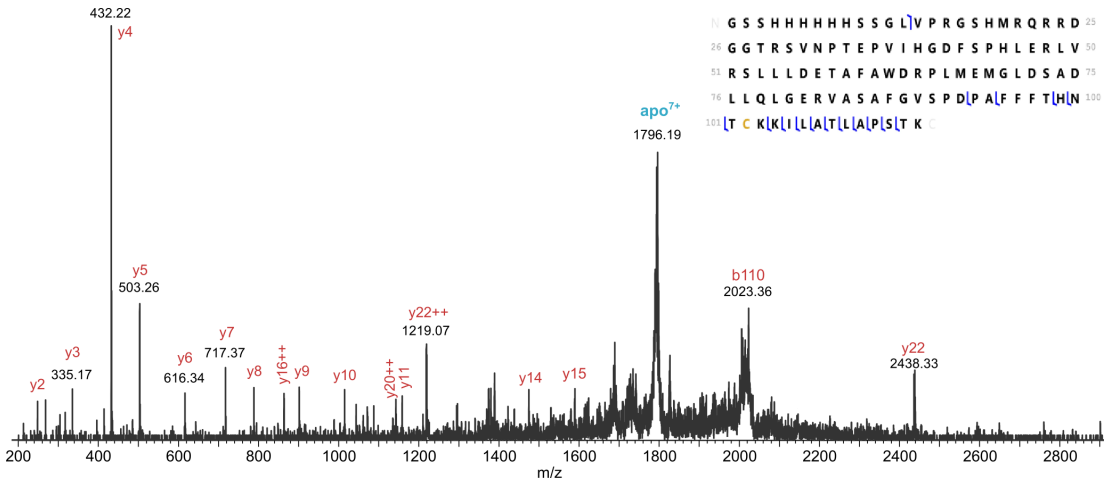


Figure 4.15 | Annotated MS/MS spectrum from top-down CID of *apo*-PsyACP1^{7+}. Identified ions are annotated. An inset shows the identified peptide fragment ions on the sequence of PsyACP1, output from ProSight Lite[171] v1.4.

in the ejection, either through direct charging on the Ppant or coulombic repulsion within the *holo*-protein. PsyACP1 has an equal number of basic (Arg+Lys=12) and acidic (Asp+Glu=12) residues, therefore a charge of 7+ would neutralise seven acidic residues while a charge of 13+ would result in complete neutralisation of the acidic residues and protonation elsewhere.

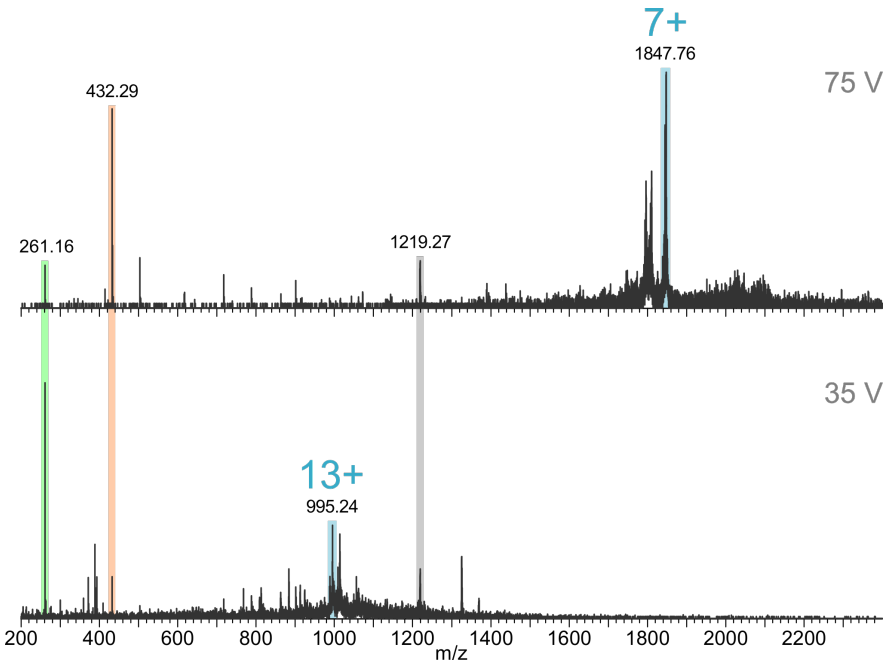


Figure 4.16 | Stacked nESI-MS/MS spectra of selected PsyACP1 7+ (1847.8 *m/z*, *top*) and 13+ (995.2 *m/z*, *bottom*) charge states, generated with collision voltages of 75 V and 35 V corresponding to E_{lab} values of 525 eV and 455 eV respectively. Spectra are minimally smoothed (Savitsky-Golay, 2×5). Key ions are highlighted.

4.4.3 Excessive Fragmentation of Pant$^+$ Ions

Under CID, multiple collisions occur as the precursor ion travels through the collision cell until sufficient internal energy is transferred. This can mean that product ions of collision may themselves undergo collision and fragmentation before leaving the cell. The collision voltages required for successful Ppant ejection should increase with protein size and decreasing charge state, approximately following Equation 1.23. It is probable that these high collision voltages could cause excessive fragmentation of ejection ions, removing their diagnostic utility. Using CoA as a *holo*-ACP analogue, the stability of Ppant ejection ions was tested under the same native conditions as above, see Figure 4.17. The intensity of the Pant$^+$ ion (261.2 m/z) increases with collision voltage until 35 V, with appreciable degradation after 40 V. Exposure of the Pant$^+$ ion to higher voltages for sufficient time within the collision cell will result in similar degradation. In order to reduce this excessive fragmentation, higher charge states or more directed energy transfer methods should be sought.

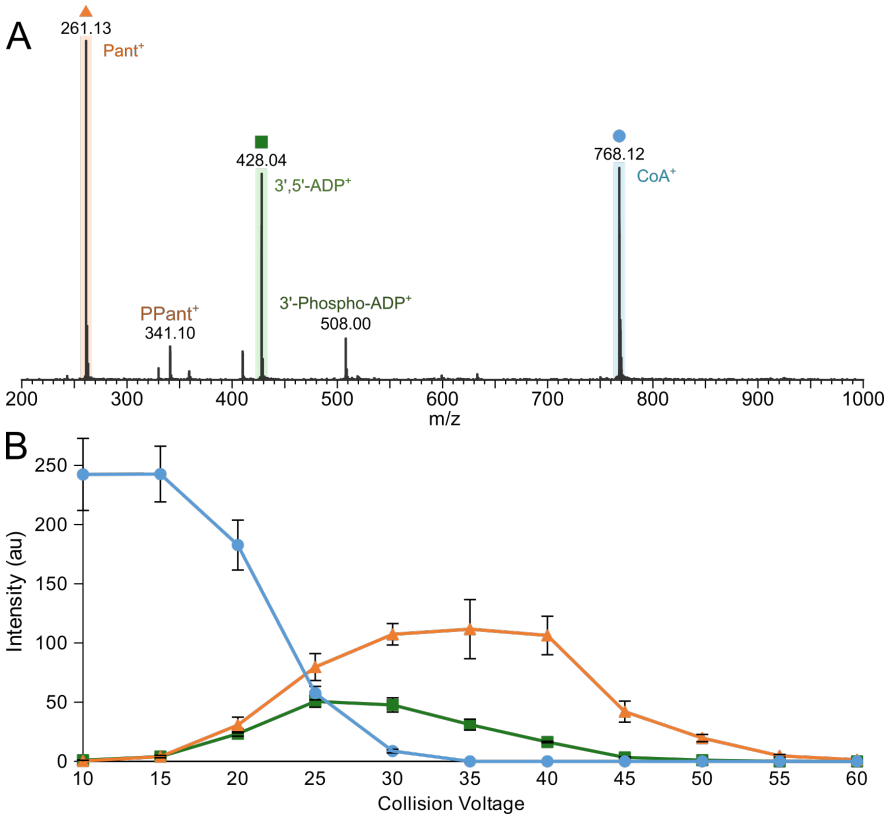


Figure 4.17 | Ppant ejection from CoA. (**A**) Ppant ejection spectra at 25 V, key ions are highlighter and annotated. (**B**) Ion intensity for precursor (768.1 m/z, blue ●); Pant$^+$ (261.1 m/z, orange ▲) and 3',5'-ADP$^+$ (428.1 m/z, green ■) product ions from Ppant ejection at increasing collision voltage, $n = 3$.

4.4.4 Supercharging of PsyACP-KS1

Supercharging additives can drastically increase the average charge state of ions while seemingly maintaining native gas-phase structure.[43,44] Increasing the charge state of native PKS proteins could allow Ppant ejection to be performed at lower collision energies, preserving $Pant^+$ ions. Sulfolane was used to supercharge PsyACP-KS1 from its typical z_{av} of 17.5 up to 23.5. The higher m/z ions generated were found to have broader peaks, possibly due to adduction of sulfolane itself.[44] The intensities of these peaks were also consistently lower than those in samples without sulfolane present. Because of this, isolation of the charge states and subsequent collisional activation gave weak spectra and it was not possible to determine if Ppant ejection was occurring.

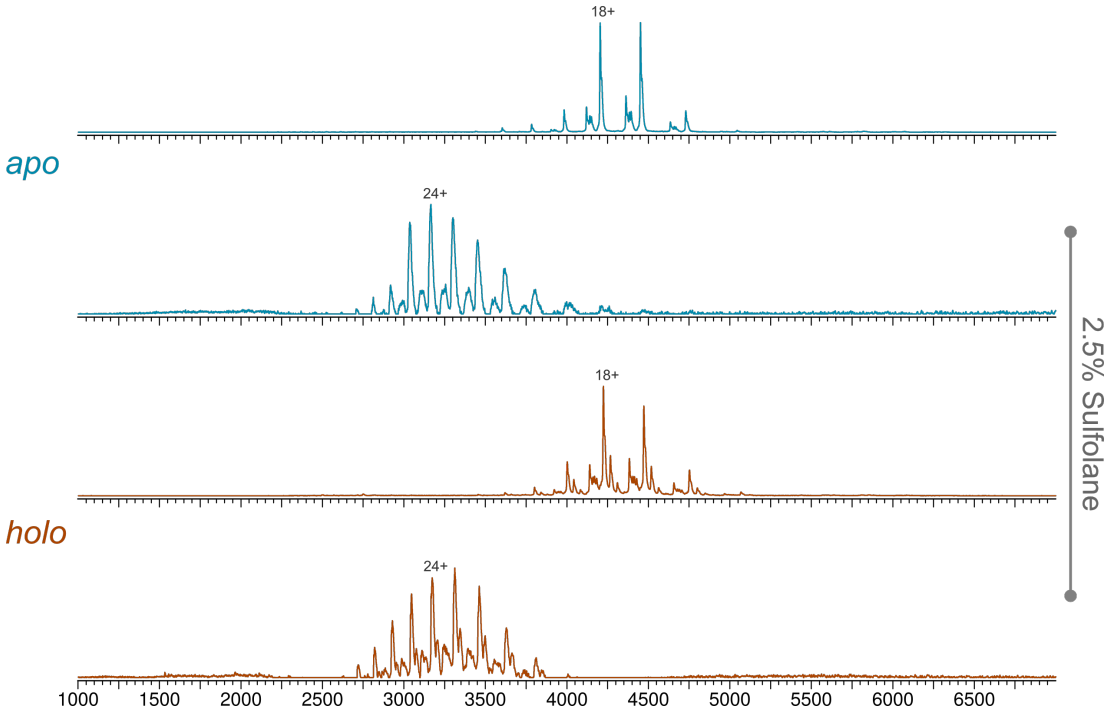


Figure 4.18 | Stacked nESI-MS spectra of native *apo* (*top*) and *holo* (*bottom*) PsyACP-KS1 supercharged with 2.5 % sulfolane.

4.5 Conclusions and Outlook

Several PKS proteins have been investigated by nMS, from an isolated ACP domain up to an intact module, showing the promise that nMS techniques offer to understanding PKS systems. Using the preparation methods described, even intact modules can be readily analysed by nMS in small quantities. CIU offers a window into the structure of proteins in the gas-phase, in particular it has been demonstrated that the Ppant and its acyl variants are responsible for stabilisation in the gas-phase unfolding of PsyACP1, as well as compaction of the unfolded conformer. This work can be expanded to understand key interactions within ACPs for gas-phase stability, this would be most easily achieved through mutagenesis. The study of larger and more complex PKS proteins by CIU might allow structural changes during polyketide turnover to be investigated, herein only preliminary studies on larger PKS proteins (PsyKS1 and PsyACP-KS1) have been performed.

The small quantities, sensitivity and mass resolution provided by nMS could allow for PKS reactions to be monitored in real-time. While Ppant ejection of intact PKS modules has proved difficult with CID methods, there are alternative options that may provide some potential. Surface-induced dissociation (SID)[29,209] allows for all of the energy to be applied in a single impact, preventing excessive fragmentation of the $Pant^+$ ions. Native complexes of ACPs and the remnants of modules may yet provide a better alternative, allowing the acyl-ACP to be dissociated from the module before Ppant ejection. Such complexes would require the addition of recombinant docking domains, either engineered (SYNZIP)[210] or PKS derived.[197] Using such non-covalent modules might allow the module flux of 'legoized' modules to be determined in a high-throughput manner.

5 Development of a Semi-Automated Software for Protein Footprinting

5.1 Introduction

Protein footprinting is an emerging technique for probing the structure and interactions of proteins. Several protein footprinting techniques exist, foremost are HDX, HRF/FPOP and carbene footprinting, as described in Section 1.1.4.3.[211] Unlike HDX, covalent methods alter the chemistry of target proteins, resulting in different retention times for labelled peptide isomers under typical LC–MS conditions. This significantly adds to the complexity of data analysis with typical analysis times running into weeks for simple differential experiments, ruling out high-throughput usage of these footprinting techniques. While HDX data can be analysed readily with a variety of software packages,[212–218] which use isotope distribution fitting to determine the extent of deuterium uptake; packages for covalent methods leave much to be desired. Two key software packages that exist are ProtMapMS[219] and ByoLogic, proprietary and commercial packages respectively. Both applications rely on MS/MS data for target peptides to fully automate assignment and quantify modifications. In an effort to

The work presented in this chapter has been published as: Bellamy-Carter, J.; Oldham, N. J. PepFoot: A Software Package for Semiautomated Processing of Protein Footprinting Data. *J. Proteome Res.* **2019**, *18*, 2925–2930. Source code and binaries for PepFoot are available from github.com/jbellamycarter/pepfoot.

increase accessibility and visualisation of covalent footprinting data, PepFoot,[220] an open-source software package that allows semi-automated processing of LC–MS data from covalent footprinting experiments, was developed using the Python programming language and a graphical user interface created using the Qt framework.

5.1.1 Manual Data Processing

At present the majority of covalent footprinting data is processed manually, typically using instrument vendor software (e.g. Xcalibur or MassLynx) to navigate ion chromatograms and mass spectra, and to integrate chromatographic peaks. The data are then often processed in a spreadsheet (e.g. Excel) before mapping to a protein structure (e.g. PyMOL or Chimera). A typical manual workflow for processing footprinting data would be as follows:

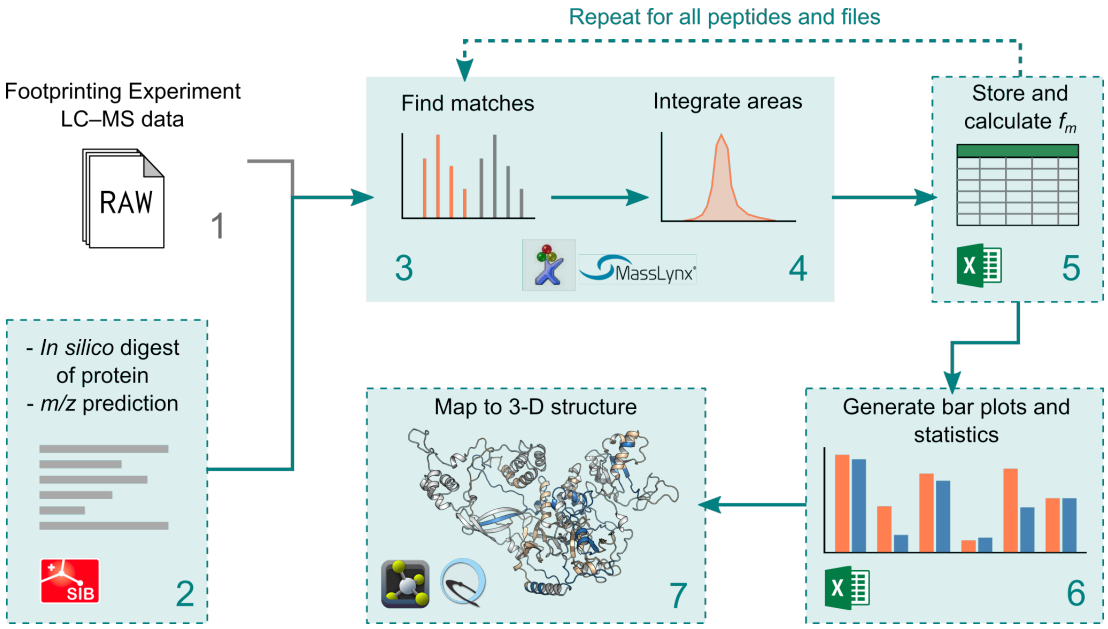


Figure 5.1 | Typical manual footprinting analysis workflow, including common software for each step. (**1**) Perform footprinting reaction and follow up LC–MS. (**2**) Generate *in silico* list of peptides with theoretical m/z for unlabelled and labelled ions. (**3**) Find m/z matches in raw data. (**4**) Integrate chromatograms for positive matches. (**5**) Store areas and calculate f_m (as defined in Equation 1.25), then repeat steps 3 and 4 for all peptides and data files. (**6**) Generate bar plots and statistics from full data set. (**7**) Map results onto 3-D model of protein structure. All steps in blue are performed directly in PepFoot, with automation available for steps with dotted outline.

Assuming that chromatography is reasonably reproducible (i.e. the same peptide isomer will have the same retention time after multiple runs), multiple data files can be batch processed with the same m/z and retention time values from the first file. By reducing user-analysis time to that of a single data file, many runs can be

quickly processed and high-throughput analyses become more feasible. Furthermore, by combining all of the various steps into a single integrated software the analysis becomes less burdensome and less prone to human error.

An obvious question regarding automation of such data processing is, "Can the process be *fully* automated?". Certainly, within the scope of HDX experiments, this is already possible. A caveat for the covalent modifications, as mentioned above, is the variable retention of labelled isomers in LC–MS. As such, this leads to the high possibility of false negatives, losing information about low abundance species. Therefore the primary focus of PepFoot development has been a manual-first approach, to provide the scrutiny and, in some cases, lenience from a user and the speed of a computer for the intensive processing.

5.1.2 Data Formats

Protein footprinting experiments can be performed on any LC–MS setup designed for peptide analysis, as the labelling reactions themselves are performed offline. As such, an accessible software would allow for data from the majority of instrument vendors. Access to many MS vendor formats is provided by binaries available from most of the major vendors, however, writing compatibility for all such binaries is time-intensive. Fortunately, an open-source software, ProteoWizard has been developed to allow access to, and conversion from, all major vendors.[194] By utilising a standardised open-source framework, the initial stage of data analysis, namely data access, is significantly simplified. Of the open-source formats available through ProteoWizard the most prominent is the .mzML format, an XML-based hierarchical data format, which is accepted as input for a variety of software packages and repositories.[221,222]

mz5 Format

Due to the XML framework it is based upon, .mzML suffers from slow read-write speeds and large file sizes. A variant of this format based on HDF5 (Hierarchical Data Format version 5), dubbed .mz5, overcomes these shortcomings due to the inherent speed of the HDF5 format and storage of *m/z* data as a delta mass representation.[223] The .mz5 file format consists of a set of HDF5 datasets, the most important are as follows:

`CVParam` Contains information about the data acquisition and per-scan parameters. Using the controlled vocabulary described in `CVReference`

`CVReference` Reference for `CVParam`, using the controlled vocabulary outlined by the Proteomics Standards Initiative for Mass Spectrometry data.[224,225]

`SpectrumIndex` 1-D array of scan indices, corresponding to the last index in a scan + 1 for `SpectrumMZ` and `SpectrumIntensity`.

`SpectrumMZ` 1-D array of delta mass ($\Delta m/z$) values for contiguous scans. The first index of a scan contains the lowest m/z for that scan.

`SpectrumIntensity` 1-D array of intensity values for contiguous scans, corresponds exactly to the m/z in `SpectrumMZ`.

5.2 Data Processing in PepFoot

5.2.1 Extracting Data from .mz5

A custom Python class object was created to interact with the .mz5 file format, `mz5` (Listing A.2). This class is initialised with a .mz5 data file and provides custom functions to parse scan information (`fill_lookup`), generate mass spectra (`spectrum`), extracted ion chromatograms (`chromatogram`) and integrate chromatographic peak areas (`get_area`).

Mass Spectra

For a given scan, a spectrum can be generated by first searching the `SpectrumIndex` array to find the scan index `i`. The $\Delta m/z$ and intensity arrays for this can be extracted with range `[i-1:i]`. The original m/z array can then be reconstructed using a cumulative sum function (e.g. `numpy.cumsum`), see Figure 5.2A. By default, this is performed *ad hoc*, but the *in memory* option will perform this once upon opening the data file to generate a ragged array of reconstructed m/z arrays. To generate a combined spectrum from multiple scans, each individual spectrum must be generated. The m/z arrays of first two scans are compared with each other for similarity, in PepFoot this is achieved with the function `in1d` from the numpy[226] (`np`) Python library. Briefly, the scans are first compared in both directions, generating a boolean mask for m/z from the

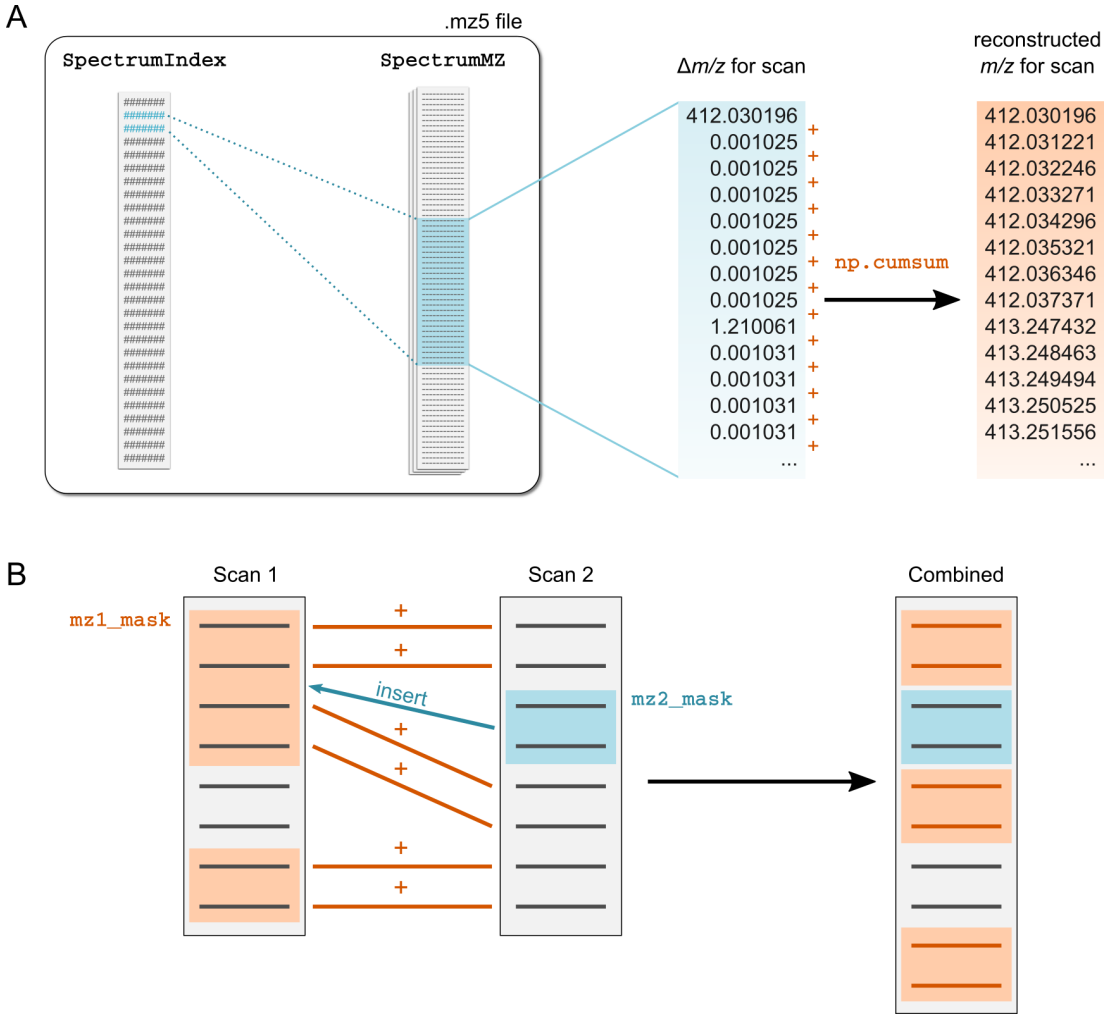


Figure 5.2 | Schematic representation of spectrum generation from the .mz5 file format. **(A)** The scan indices are extracted from the `SpectrumIndex` array and used to extract the $\Delta m/z$ data from `SpectrumMZ` before cumulative summation to yield a reconstructed m/z array. **(B)** Two scans are combined by first generating masks of m/z from Scan 1 present in Scan 2 (`mz1_mask`), and m/z from Scan 2 absent in Scan 1 (`mz2_mask`). These masks are used to sum intensities between scans and to insert new [m/z, intensity] entries.

first scan present in the second scan (`mz1_mask`) and a mask for the m/z from the second scan absent in the first scan (`mz2_mask`). The masks are then used to sum intensities for m/z found in both arrays in a new combined intensity array, then to insert new entries in the combined m/z and intensity arrays for m/z unique to the second scan, see Figure 5.2B. This is then repeated for all subsequent scans against the combined arrays.

Chromatograms

Generating an extracted ion chromatogram (EIC) is somewhat simpler. For each scan, the intensity and reconstructed m/z arrays are generated, the m/z array is then searched for m/z with values between lower and upper limits. The indices of those values are

used to slice the corresponding intensity values, which are then summed and appended to a chromatogram intensity array. Chromatographic peak areas are generated by integrating peaks, the integration is an approximation with several methods available. The most rudimentary way of doing so is to sum chromatographic intensities with no account of spacing in the time-domain; while quick, the curve is approximated with steps and variations in time-spacing lead to gross misestimations of area. The more common numerical method is the trapezoidal rule, where the area between two adjacent points is approximated to a trapezoid (e.g. `np.trapz`). In this approximation the line between two point is linear, leading to both under- and over-estimation. Other methods can involve fitting the peak to a Gaussian curve and integrating it algebraically. For rapid and reasonably accurate area approximation, the trapezoidal rule has been used in PepFoot.

Batch Processing

The key advantage of PepFoot over the manual approach is the ability to quickly process a batch of experimental files. This is performed by the `batch_process` function. Briefly, a set of *m/z* ranges and retention time ranges for unlabelled and labelled peptide ions are integrated and the areas stored. All user-selected data files are iterated through, with the time for processing proportional to the number of peptides and *m/z* array memory option, shown in Figure 5.3, approximately 10 s per file for 20 peptides. Provided the chromatography is reproducible between all of the selected runs, the resulting f_m are highly accurate. As the batch processing time per file is short, the data generation is time-limiting.

5.2.2 pfoot Format

PepFoot uses the .pfoot project file to store all necessary information for processing an experiment, including the peptide generation parameters, matched peptides, *m/z* and retention ranges for batch processing, integrated areas, and fraction modifications for all peptides and data files. The .pfoot file is a JavaScript Object Notation (JSON) format, which directly mirrors an in-memory Python dictionary called `project`, this

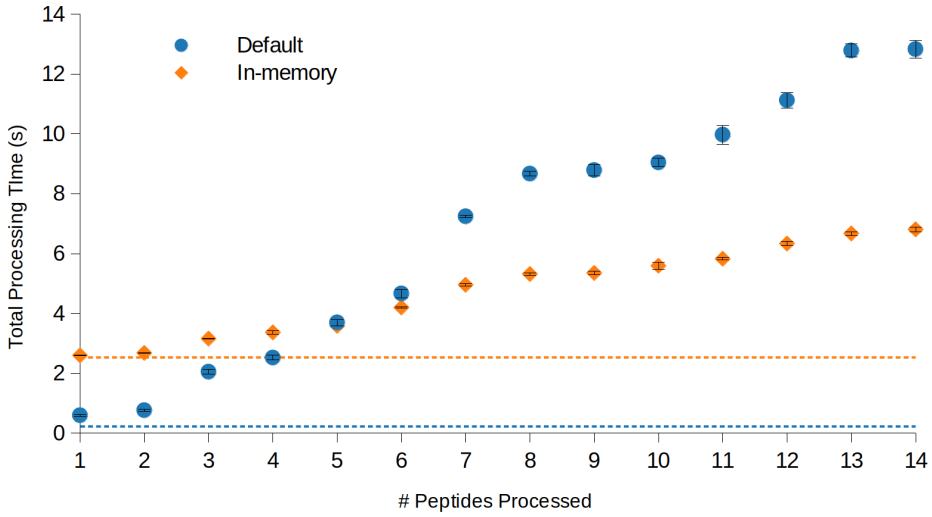


Figure 5.3 | Comparison of total batch processing times in PepFoot for default (blue, ●) and in-memory (orange, ♦) m/z array options. The average times for file loading prior to peptide integration are shown as dashed lines for each option, 0.217 s and 2.52 s respectively. Error bars indicate ± s.d. ($n = 3$)

schema is described in Table 5.1. The .pfoot file is human readable and easily parsed with existing JSON packages. An example of a .pfoot file can be seen in Listing A.3.

5.2.3 Generating Theoretical Peptides

A custom Python class object was created to represent peptides, `Peptide`. In PepFoot, a theoretical list of peptides is generated, where each peptide is represented by a member of this class. The `Peptide` object is initialised with the peptide sequence, position in a protein, modifications and some user set parameters for mass range and peptide charge. Once initialised, a variety of attributes for a given peptide can be accessed using the `.attribute` syntax, see Table 5.2. A set of unique peptide sequences is generated from a user-provided protein sequence using the `parser.cleave` function from the pyteomics[187,188] library. The parser generates peptides using enzyme rules in regular expression form (e.g. trypsin is defined by `[KR](?=[^P])`) and is tolerant to modX notation, where modified peptides are indicated using a lowercase prefix. The most abundant mass for a peptide is determined from the theoretical isotope distribution based on the elemental composition as described by a `Composition` object. This distribution is calculated by the PepFoot function `get_isotopes` (Listing A.4), which uses the stepwise approach for fine-structure derivation. The accuracy of the isotope distribution is dependant upon the mass tolerance and abundance threshold.

Key	Description (Shape)
`data files`	List of data files (1×M)
`sequence`	Unmodified sequence string
`length range`	Peptide length range (1×2)
`charge range`	Charge state range (1×2)
`enzyme`	Enzyme name
`missed cleave`	Number of missed cleavages
`fixed mods`	List of fixed modifications applied to protein
`differential mod`	Differential modification applied to peptides
`peptides`	Array of peptide IDs (N×2)
`charge array`	Array of charge states for `peptides` (1×N)
`m/z array`	Array of *m/z* ranges for both unlabelled and labelled peptides (2×N×2), `[0,0,:]` and `[1,0,:]` for the unlabelled and labelled *m/z* ranges for peptide at index `0`
`rt array`	Array of retention time ranges for both unlabelled and labelled peptides (2×N×2), same as `m/z array`
`areas`	Array of area values from analysis (2×M×N), `[0,1,2]` and `[1,1,2]` for the unlabelled and labelled areas of peptide at index 2 in data file at index `1`
`fractional mod`	Array of f_m values calculated from `areas` (M×N)
`treatment`	Nested list of indices for grouped data files
`pdb file`	Associated .pdb file for mapping f_m

Table 5.1 | Schema for .pfoot file. Keys and descriptions of them are given along with array shapes, if applicable. For a .pfoot file with M data files and N peptides.

Attribute	Description
`sequence`	Unmodified sequence string
`id`	Positions of peptide within protein, i.e. (1, 21) for 1-21.
`composition`	Pyteomics `Composition` object for unlabelled peptide, subclass of dictionary type, e.g. {H: 4, C:2, O:1} for C_2H_4O
`mmass`	Monoisotopic mass of unlabelled peptide
`mass`	Most abundant mass of unlabelled peptide
`chgs`	List of possible charge states for peptide
`isotopes`	List of theoretical isotopes for unlabelled peptide
`mod_isotopes`	List of theoretical isotopes for labelled peptide
`diff_mod`	Differential modification name
`mzs`	List of unlabelled *m/z* for charges states in `chgs`
`mod_mzs`	List of labelled *m/z* for charges states in `chgs`
`mz_ranges`	Nested list of *m/z* ranges for verified unlabelled and labelled peptides. Shape 2×2, `[0,:]` and `[1,:]` for unlabelled and labelled respectively
`rt_ranges`	Nested list of retention ranges for verified unlabelled and labelled peptides. Same shape as `mz_ranges`
`areas`	List of integrated areas for verified unlabelled and labelled peptides
`chg`	Charge state of verified peptides.
`detected`	Detection state of the peptide: -1 (not detected), 0 (unlabelled), 1 (labelled) and 2 (both)

Table 5.2 | Attributes available for instances of the `Peptide` class object. A description of the attribute is given. Attributes that are set during peptide identification are highlighted in grey.

5.3 Visualising Footprinting Data

For a given set of peptides with fractional modifications as calculated by Equation 1.25, the data can be presented in several ways. Most commonly, a bar plot is used to show the fractional modifications against peptides; for differential experiments, the data are grouped (*apo* and *holo*) and clustered bar plots used to compare the differences between the groups. Once grouped, significance testing (e.g. Student's *t*-test) can be used to provide statistical meaning to any differences seen. It is also useful to represent the extent of change in modification. Mapping of any of these results to a 3-D protein structure is standard, with colour used to represent the footprinting data.

5.3.1 Extent of Change in Modification

The extent of change in modification E_m is typically calculated as in Equation 5.1, where f_{apo} and f_{holo} are the fractional modifications of the *apo* and *holo* groups — to represent the absence and presence of a binding partner — for a peptide. This number can then be expressed as a percentage change. Using this formula however, increases in f_m between *apo* and *holo* can be exponentially large and skew the appearance of the data. As such, PepFoot uses a modified version shown in Equation 5.2, where the denominator is the maximum of f_{apo} and f_{holo}. In this way, the value of E_m is bounded by the limit $[-1, 1]$, where $E_m = -1$ represents complete masking and $E_m = 1$ represents complete exposure of a peptide.

$$E_m = \frac{f_{holo} - f_{apo}}{f_{apo}} \tag{5.1}$$

$$E_m = \frac{f_{holo} - f_{apo}}{\max\{f_{apo}, f_{holo}\}} \tag{5.2}$$

5.3.2 Mapping to Protein Structures

Footprinting results, as calculated above, can be readily mapped onto 3-D protein structures in most molecular viewer programs (e.g. PyMOL and Chimera). This is achieved by either manual selection of residues and then colouring, or by using the selection algebra provided by most of these software (e.g. in PyMOL `color red, resi`

`1-21` will colour residues 1–21 red). The algebra between these programs varies and integration of these multifunctional software is somewhat difficult.

To preserve individual user preferences for molecular viewer and simplify 3-D visualisation, PepFoot will modify the b-factor attribute for an atom, such that any viewer can be quickly used to colour or style by footprinting data. This is achieved with the `update_bfactor` and `PDB.bfactor_by_residue` functions in PepFoot. The former simply takes the appropriate footprinting output and converts it to a b-factor for a given residue, while the latter modifies a Python class object representation of a user selected .pdb file `PDB` (Listing A.5). This object is then saved to disk in the .pdb file format. The b-factor value is set so that non-detected residues have a value of -2, detected but insignificantly labelled residues have a value of 0 and significantly labelled residues have a value of $0 \rightarrow 1$. For differential experiments, masked residues have a value of $0 \rightarrow 1$ and exposed residues have a value of $0 \rightarrow -1$. PepFoot provides the option to set the b-factor with discrete or continuous data. For discrete data, significantly labelled peptides are defined as those above a user set threshold for f_m and with p-values below a user set threshold in differential datasets. For continuous data, the f_m is used for b-factor in normal experiments and $-E_m$ used for differential experiments. In addition to modifying .pdb files, PepFoot provides an integrated molecular viewer in the form of the NGL Viewer. NGL Viewer is a lightweight open-source Javascript-based molecular viewer.[190,191]

5.4 User Interface

The PepFoot graphical user interface (GUI) is based in the Qt 5.11 framework. The GUI was designed using the QtDesigner application with the PyQt5 library providing access to the GUI. The GUI (shown in Figure 5.4) consists of a main window featuring a tabbed panel (A) for user interaction with the data, and a sidebar (B) for user defined parameters to generate theoretical peptides. The tabbed panel has three tabs: Peptide Level, Analysis and NGL Viewer. In addition to the main window, a window for setting custom modifications (C) and enzymes (D) is available.

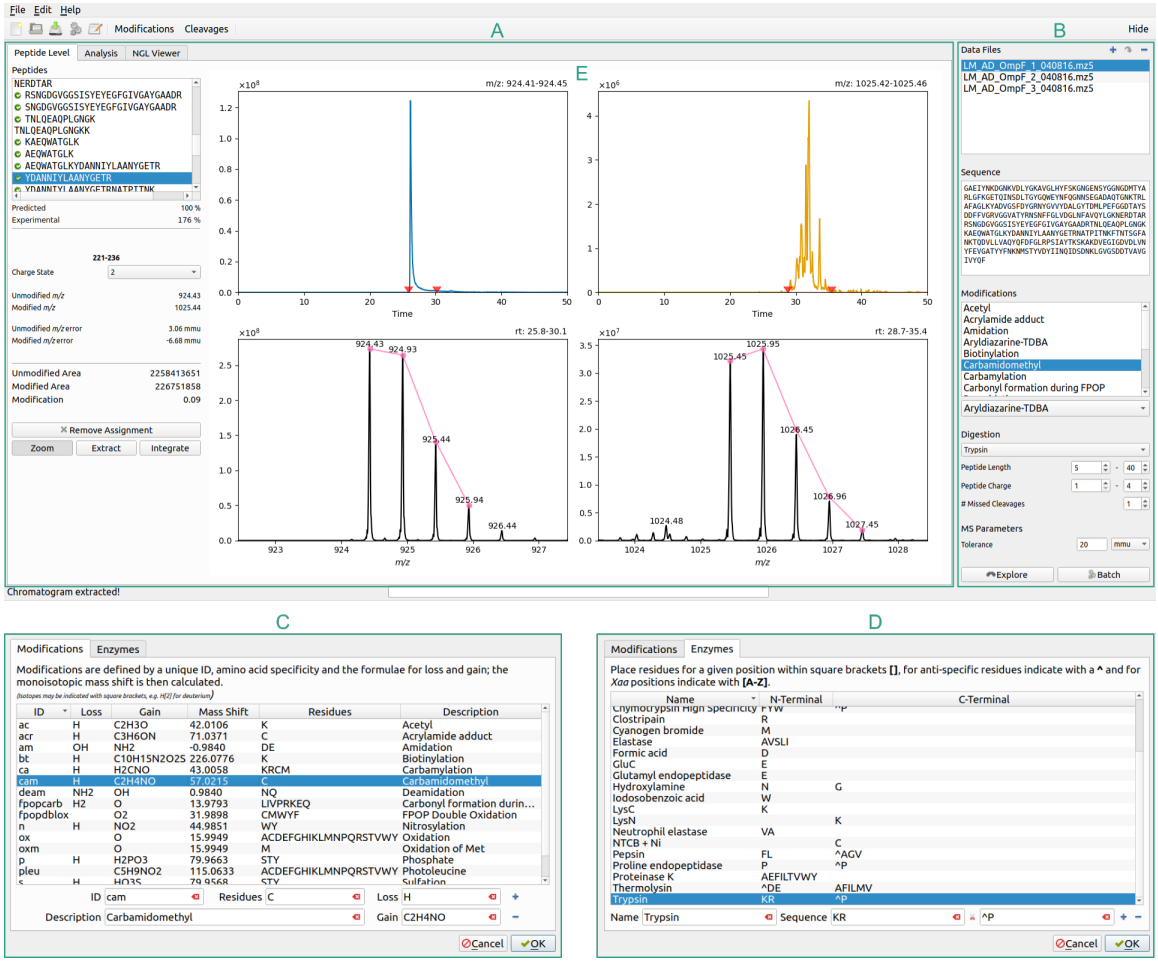


Figure 5.4 | Screenshots of the PepFoot GUI windows. (**A**) Main tabbed panel. (**B**) Sidebar for data and parameter input. (**C**) Dialog for setting modifications. (**D**) Dialog for setting enzymes. (**E**) Interactive plotting panel with EICs (*top*) and mass spectra (*bottom*) for unlabelled (*left*) and labelled (*right*) peptide ions.

5.4.1 Processing Data

After creating a new .pfoot project file, the user can add/remove .mz5 files in the *Data Files* widget. The user may alternatively select .raw files and convert them to .mz5 files within PepFoot, provided an installation of `msconvert` is present. Protein sequence is added in the *Sequence* widget, which can accept modX strings for specific modifications. Fixed modifications and the labelling probe can be selected from a list of user-defined modifications. Digestion parameters can be set including enzyme, peptide length and charge, number of missed cleavages. Lastly, the *m/z* tolerance can be set, this is the window used for generating initial EICs. When ready the user may select the Explore button, which will generate the list of peptides as described above and open the first/selected data file. The *Peptides* widget is filled with any possible peptides that meet the above conditions. Selection of a peptide in this widget will generate an EIC for both unlabelled and singly labelled ions for a given charge state. The EICs are displayed in an interactive panel (Figure 5.4E) with four panes; the left and right panes are for unlabelled and singly labelled ions respectively, while the top and bottom panes are for EICs and combined mass spectra respectively. Using the left mouse button, a data cursor can be used to interact with the data, there are three modes available:

Zoom (default) to zoom on the x-axis with auto-zooming in the y-axis, activated by Z

Extract to generate a combined spectrum from a region of an EIC or an EIC from a region of a mass spectrum, activated by E

Integrate (EIC only) to integrate a region of an EIC, activated by I

When a user extracts a region of an EIC a combined mass spectrum is generated for the region, this region will then be displayed with downward red triangles. PepFoot automatically zooms the spectrum to a region $[-\frac{2}{z}, +\frac{3}{z}]$ around the predicted *m/z* for the peptide ion. The theoretical isotope distribution for the peptide is also overlaid to aid user verification of the peptide. If the user decides an alternative *m/z* window should be applied, a new EIC can be generated by extracting on the combined mass spectrum. Once satisfied, a user can integrate a region of the EIC, storing the *m/z* ranges and retention time ranges and area of the peptide ion in question. This can be applied to both unlabelled and singly labelled ions for a peptide; if ions are not observed for

one of these, a value of 1 or 0 are given for f_m respectively. To process all files in *Data Files*, Batch will iterate over all files as described above.

5.4.2 Visualising Results

PepFoot provides several forms of visualisation for footprinting data in the Analysis and NGL Viewer tabs. Upon reading a .pfoot file, PepFoot will extract any available f_m information for all data files. In Analysis, shown in Figure 5.5, a bar plot and a 2-D map of labelling coverage are generated in an interactive panel (A). Selection of a bar in either plot will show the f_m for the respective data files along with mean and standard deviation statistics in an information panel (B). Data files can be grouped into *apo* and *holo* treatments (C) and the bar plots updated to clustered bar charts where *apo* and *holo* are orange and blue respectively, this is also applied to the 2-D map. When grouped the results of a Student's *t*-test are shown as dots above bars, where the significance threshold used is set by the user (D). The *Unity* checkbox toggles between a stacked view of the 2-D map, or an overlaid view. The *Difference Plot* checkbox will display the bar plot with E_m (as defined in Equation 5.2) between *apo* and *holo* data instead of f_m.

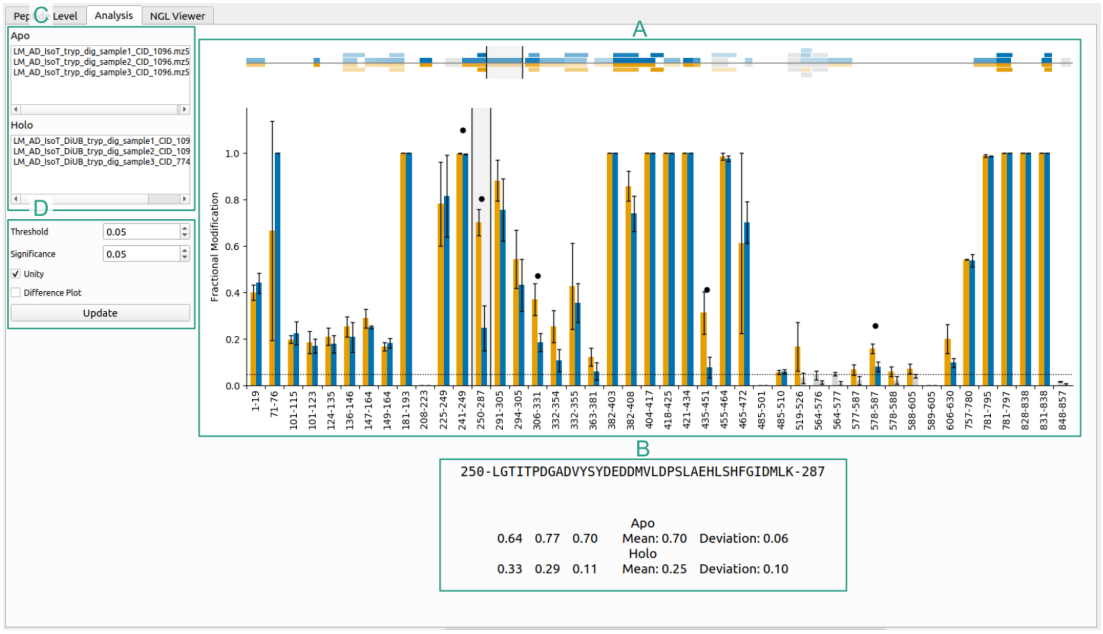


Figure 5.5 | Screenshot of the PepFoot Analysis tab. (**A**) Interactive plotting panel with a 2-D map of the protein coverage (*above*) and a bar plot of f_m (*below*). (**B**) Information panel showing sequence, f_m, and statistics for a selected peptide. (**C**) Widgets for grouping data files. (**D**) Widget to modify (A), will also update any loaded PDB files in NGL Viewer tab. See Figure A.4 for *Difference Plot* view.

The resulting f_m data can be plotted to a 3-D protein structure in the NGL Viewer tab, shown in Figure 5.6, where the interactive NGL Viewer[190,191] is embedded (A). A .pdb file where at least one chain must exactly match the fasta sequence provided can be uploaded. This .pdb file will be overwritten by PepFoot with the relevant b-factor values. The colouring mode is controlled by the *Continuous* colour scaling checkbox (B). Toggling this checkbox will cause the .pdb file to be amended with a b-factor as described in Section 5.3.2.

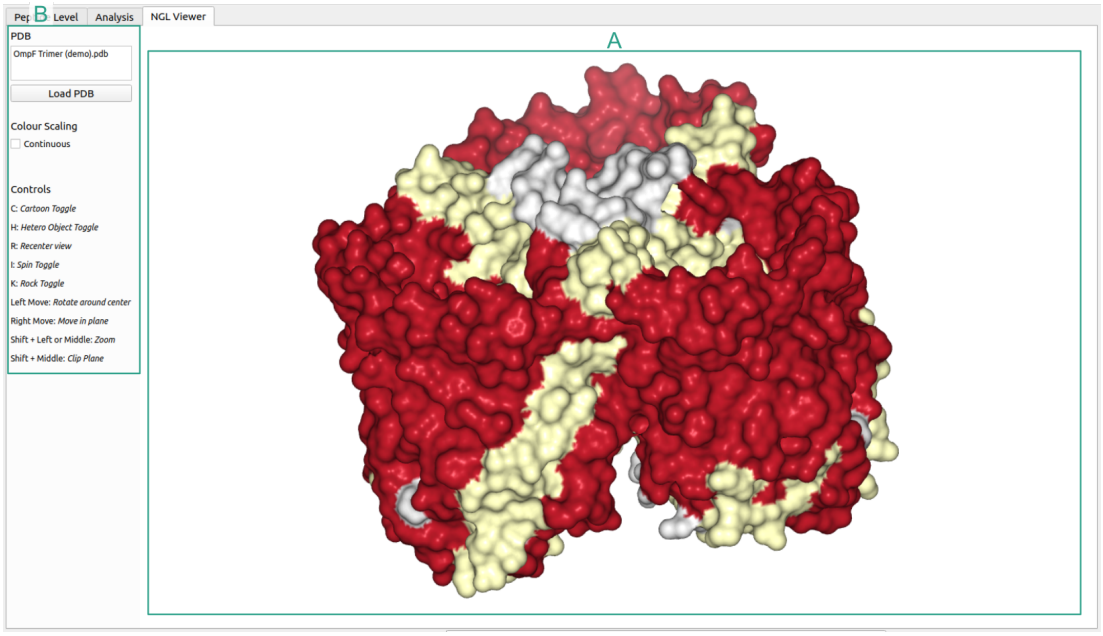


Figure 5.6 | Screenshot of the PepFoot NGL Viewer tab. (**A**) Interactive embedded NGL Viewer panel. (**B**) Control panel for NGL Viewer to load selected .pdb file, colour scaling and a list of keyboard/mouse controls for (A). See Figure A.5 for *Continuous* colour scaling.

5.5 Benchmarking PepFoot

A set of published footprinting data were processed with PepFoot to investigate its efficacy and comparability with published findings. Two publicly available datasets were chosen: the *E. coli* membrane protein OmpF[80] and the human deubiquitinating enzme ubiquitin specific protease 5 (USP5).[79] For both of these datasets, carbene footprinting was used and the MS platform was an LTQ–FT Ultra (Thermo Scientific), the LC–MS data are available in the public Proteomics Identifications Database (PRIDE, https://www.ebi.ac.uk/pride).[192,193]

5.5.1 Revealing Membrane Protein Interfaces

Footprinting of OmpF in the non-ionic detergent octyl-glucoside with the carbene probe 4-(3-trifluoromethyl)-3*H*-diazirin-3-yl)benzoic acid (TDBA, $+C_9H_5O_2F_3$) reveals membrane-binding and protein-trimer interfaces.[80] The data were downloaded from PRIDE dataset PXD007207 as Thermo .raw files. These were converted to .mz5 and processed with PepFoot (aryldiazirine-TDBA, trypsin, peptide length = 5–40, charge range = 1–4, missed cleavages = 1 and mass tolerance = 20 mmu).

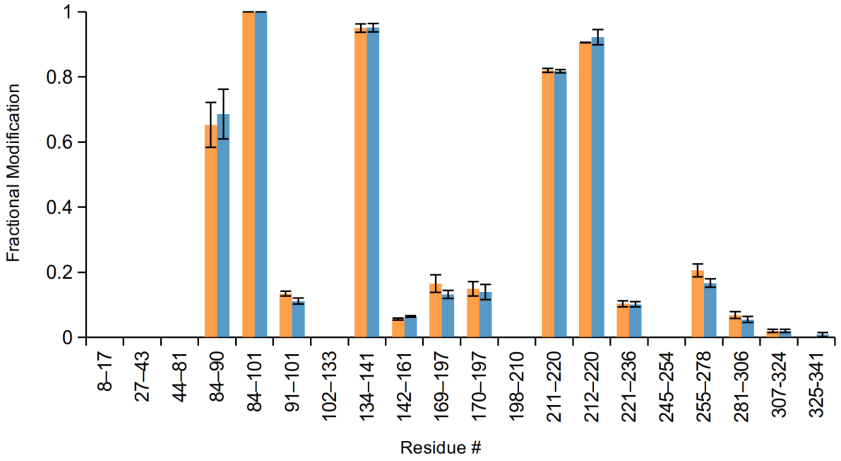


Figure 5.7 | Comparison of peptide fractional modification outputs for PepFoot and from the published data[80] for OmpF. PepFoot data are shown in orange and published manually processed data in blue. Error bars indicate ± s.d. ($n = 3$)

Tryptic peptides yielding 91 % coverage of the protein sequence were identified with 16/27 peptides having significant labelling ($f_m > 0.05$). This is an increase in identification compared to the original study, which found 86 % coverage with 12/20

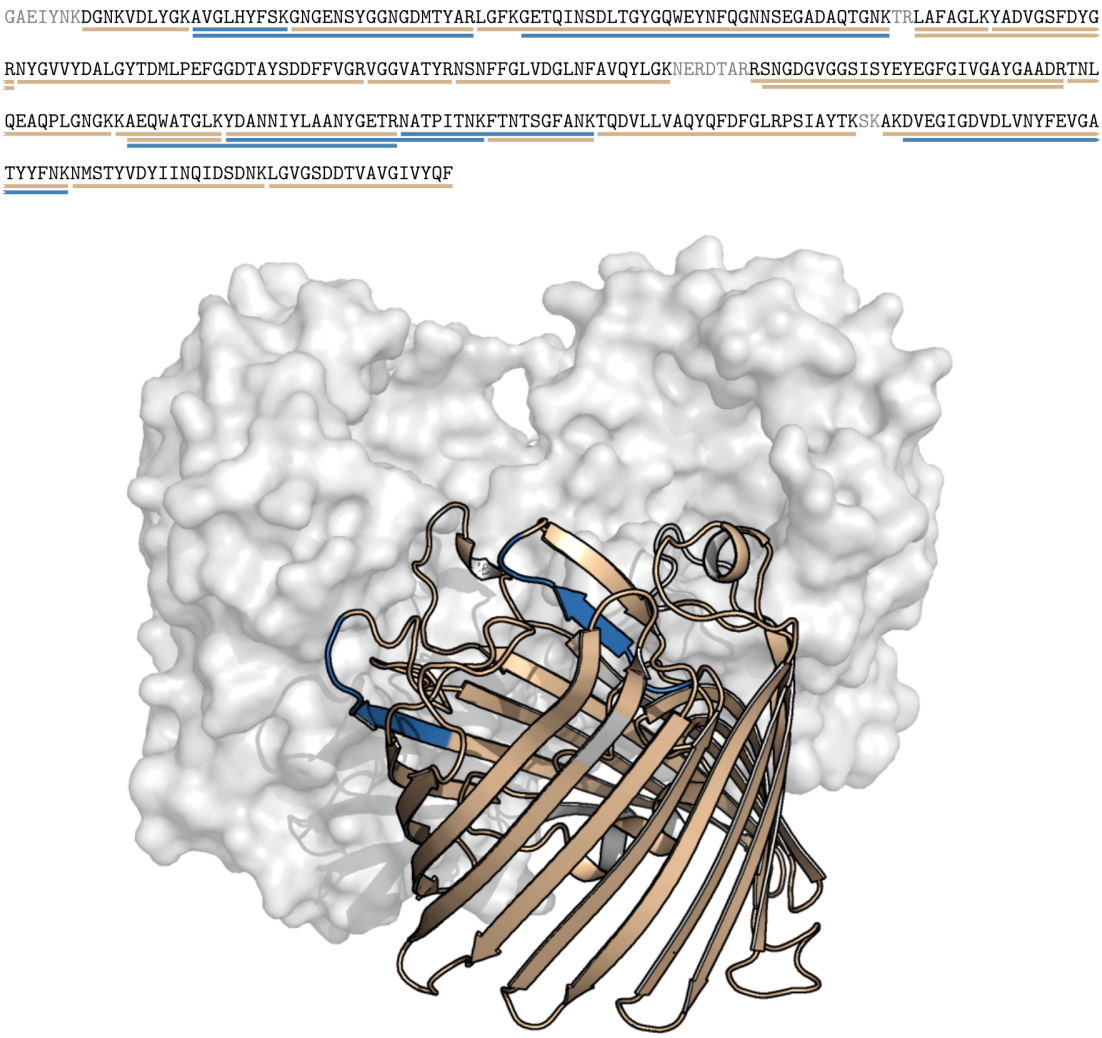


Figure 5.8 | Additional peptides found with PepFoot mapped to the OmpF sequence (*top*) and structure (*bottom*, PDB: 3POX). Undetected residues are coloured light-grey, previously identified regions are coloured wheat and newly identified regions are coloured blue. Two units of the OmpF trimer are displayed with a light-grey surface.

significantly labelled peptides. All previously identified peptides except 134–141 showed remarkably similar f_m and deviations to the original study, shown in Figure 5.7. Peptide 134–141 shows consistent f_m when processed through PepFoot but showed an anomalously low f_m from one data file in the published study, upon further inspection this appears to be the result of manual error in recording the peak area. Manual recalculation of the peak areas (in Xcalibur) for this peptide yielded improved deviation and a close match to PepFoot output. Four of the seven newly-identified peptides (18–26, 18–43, 221–244 and 237–254) accounted for previously unidentified regions of OmpF (18–26 and 237–244), displayed in Figure 5.8.[80]

5.5.2 Identifying Protein–Protein Binding Sites

Differential footprinting of USP5 alone or in complex with diubiquitin revealed interaction of the diubiquitin with the ZnF-UBP and catalytic domains, and showed a biologically significant conformation not accessible through X-ray crystallography.[79] The data were downloaded from PRIDE dataset PXD004971 as Thermo .raw files. These were converted to .mz5 and processed with PepFoot (aryldiazirine-TDBA, carbamidomethyl, trypsin, peptide length = 5–40, charge range = 1–4, missed cleavages = 1 and mass tolerance = 20 mmu). The processed data were grouped by treatment with diubiquitin.

An additional 25 putative tryptic peptides were identified to the 19 from the published study. Of the peptides common to both analyses, there was little difference between the fractional modification and resulting significance testing, as shown in Figure 5.9, with the one exception of peptide 606–630, which is found to be statistically insignificantly different ($p = 0.06$) through PepFoot. Even by manual analysis, the difference is small.

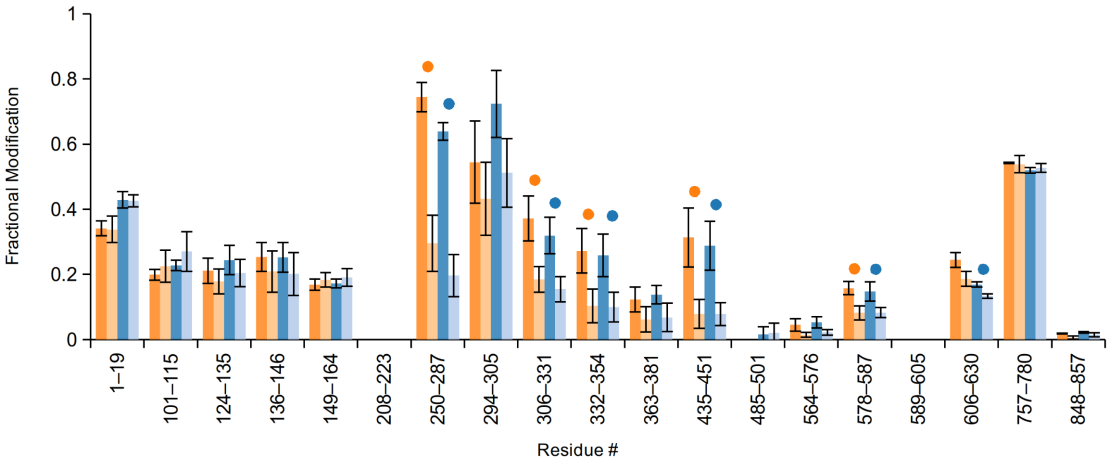


Figure 5.9 | Comparison of peptide fractional modification outputs for PepFoot and from the published data[79] for USP5. PepFoot data are shown in orange and published manually processed data in blue. The data are grouped by the absence (darker shades) or presence (lighter shades) of diubiquitin. Error bars indicate ± s.d. ($n = 3$) and significant masking ($p < 0.05$) is displayed with a dot • coloured according to the dataset.

The newly identified peptides accounted for a further 20 % protein coverage on top of the 41 % previously identified, as shown in Figure 5.10. Of these new peptides, 13 were present with extremely high labelling ($f_m > 0.95$) and may have been previously overlooked for this reason. Possibly due to the high labelling of these peptides, no

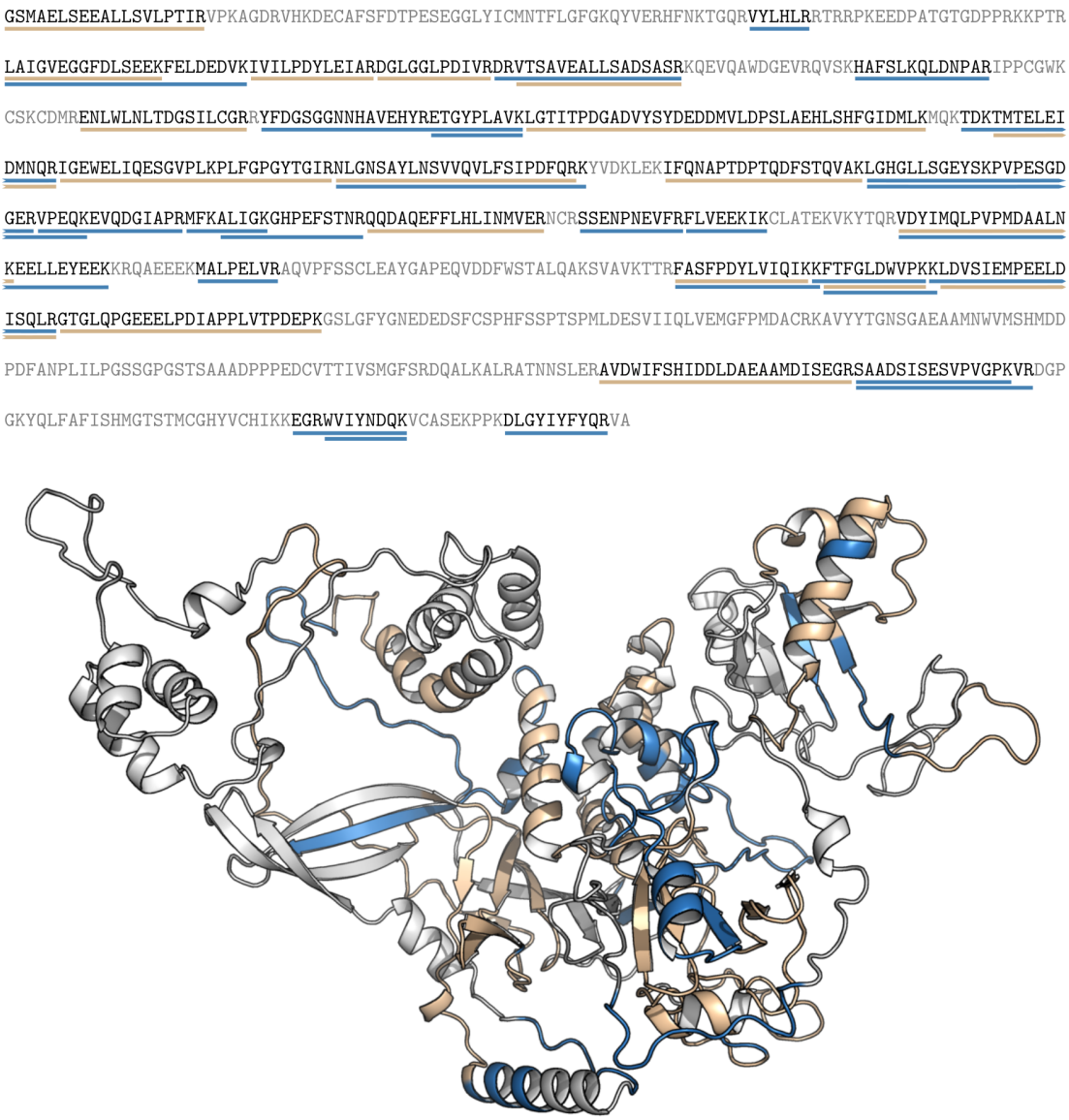


Figure 5.10 | Additional peptides found with PepFoot mapped to the USP5 sequence (*top*) and structure (*bottom*, PDB: 3IHP). Undetected residues are coloured light-grey, previously identified regions are coloured wheat and newly identified regions are coloured blue.

significant changes in labelling were observed for the new peptides. These results show that the streamlined interaction with data provided by PepFoot allows more in-depth analyses to be performed and that the grouping feature in PepFoot allows for reliable and rapid assignment of (un)masking events.

5.5.3 Hydroxyl Radical Footprinting of Myoglobin

To test the performance of PepFoot against other modifications, a data set for myoglobin from an FPOP experiment was kindly provided by the Prof. Alison Ashcroft group (University of Leeds). The data were processed with PepFoot (oxidation, trypsin,

peptide length = 5–40, charge range = 1–4, missed cleavages = 1 and mass tolerance 20 mmu). Unlike carbene footprinting, FPOP generates multiple mass shifts, depending on the amino acid residue. To account for this, the data were processed independently with three variants of oxidation (+16, 32 and 14 Da), as PepFoot can only handle a single variable probe mass shift per analysis. As each of these modifications may be present for any given peptide, the areas for unlabelled and each labelled shift were converted to a relative modification similar to f_m, see Figure 5.11 and Figure 5.12. An appreciable amount of labelling was found for the +16 and +32 oxidations with no significant labelling found for +14, as would be predicted from residue reactivity. Comparable results were obtained with manual analysis using MassLynx ($\text{RMSD}_{\text{total}} = 5.13 \times 10^{-3}$). The predominance of the +16 oxidation indicates that PepFoot can be used to characterise FPOP experiments with a single modification, supplemented by additional investigation of peptides prone to the +32 oxidation where needed.

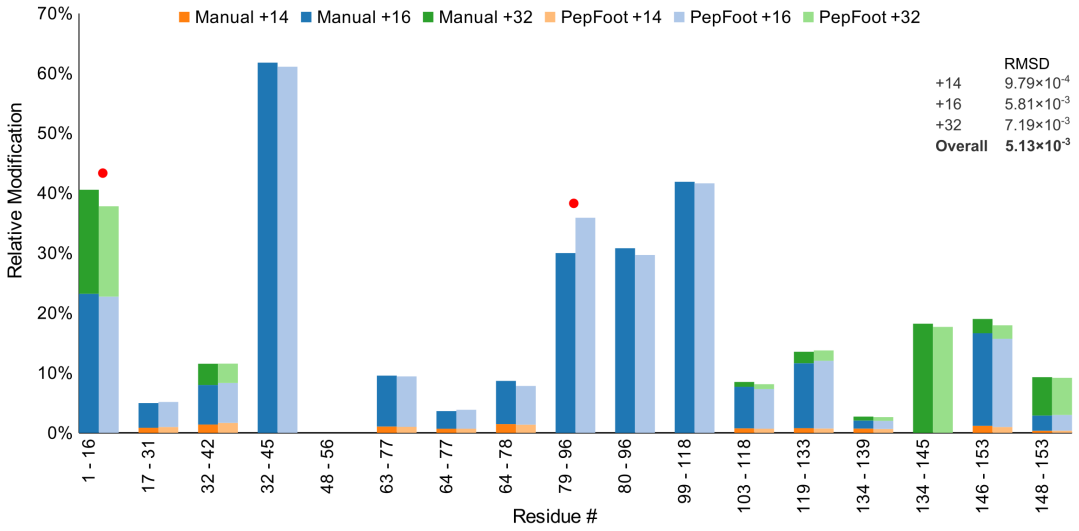


Figure 5.11 | Oxidation of *holo*-myoglobin via FPOP for both manual and PepFoot analysis represented as stacked bars displaying the relative abundance of each modification. Mass shifts of +16, 32 and 14 Da are shown orange, blue and green respectively. Manual analysis shown on left darker bars, PepFoot analysis shown on right lighter bars. Carbonyl (+14) formation is minimal with hydroxyl (+16) predominating, a single peptide is found with double oxidation (+32) alone. To show the agreement between the manual and PepFoot analyses the root-mean squared deviation (RMSD) was calculated per modification as well as for the overall set of peptides. The per peptide RMSD were found to be in good agreement, peptides with RMSD ≥ 0.01 are highlighted with a red dot ●.

5.5.4 Processing Time-of-Flight Data

In addition to being a different probe, the MS platform used for the above FPOP experiment was a Xevo G2-XS QToF (Waters), providing an opportunity to examine the

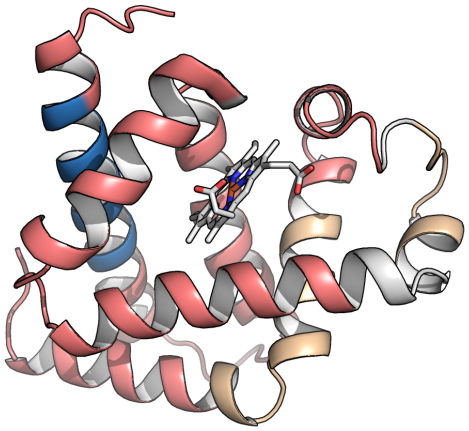


Figure 5.12 | Oxidation of *holo*-myoglobin via FPOP mapped to a structure (PDB: 1AZI). Regions are coloured, light-grey for undetected, wheat for insignificant, red for +16 oxidation and blue for majority +32 oxidation.

differences in processing ToF data. Due to the higher noise-to-signal acquired from ToF instruments, the file sizes for LC–MS experiments are large (2.6 GB for the above) and cumbersome for user interaction with the data, even in vendor software. By applying a threshold to the data, it is possible to drastically reduce the file size and remove much of the noise. The raw data file was converted to .mz5 using MSConvert with an absolute threshold of 0, 200, 500 or 1000 counts per scan; giving 619, 325, 96 and 38 MB files respectively. When applying identical batch processing to these data, it is apparent that there is a reduction in f_m for all peptides with increasing threshold, see Figure 5.13. From this analysis, a threshold of 500 counts per scan gives a reasonable compromise between user interaction and data integrity, see Figure 5.14. This filtering would typically only be applied for the initial manual verification by a user, followed by batch processing on the unfiltered data.

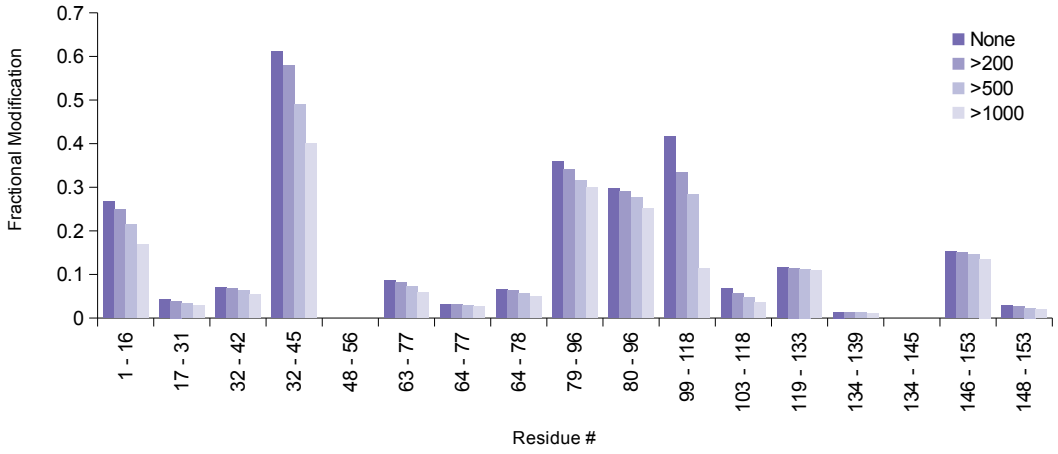


Figure 5.13 | Effect of absolute threshold on f_m for Myoglobin FPOP data, +16 oxidation. Increasing threshold shown with decreasing intensity of colour. Any threshold filter reduces the effective f_m for all peptides.

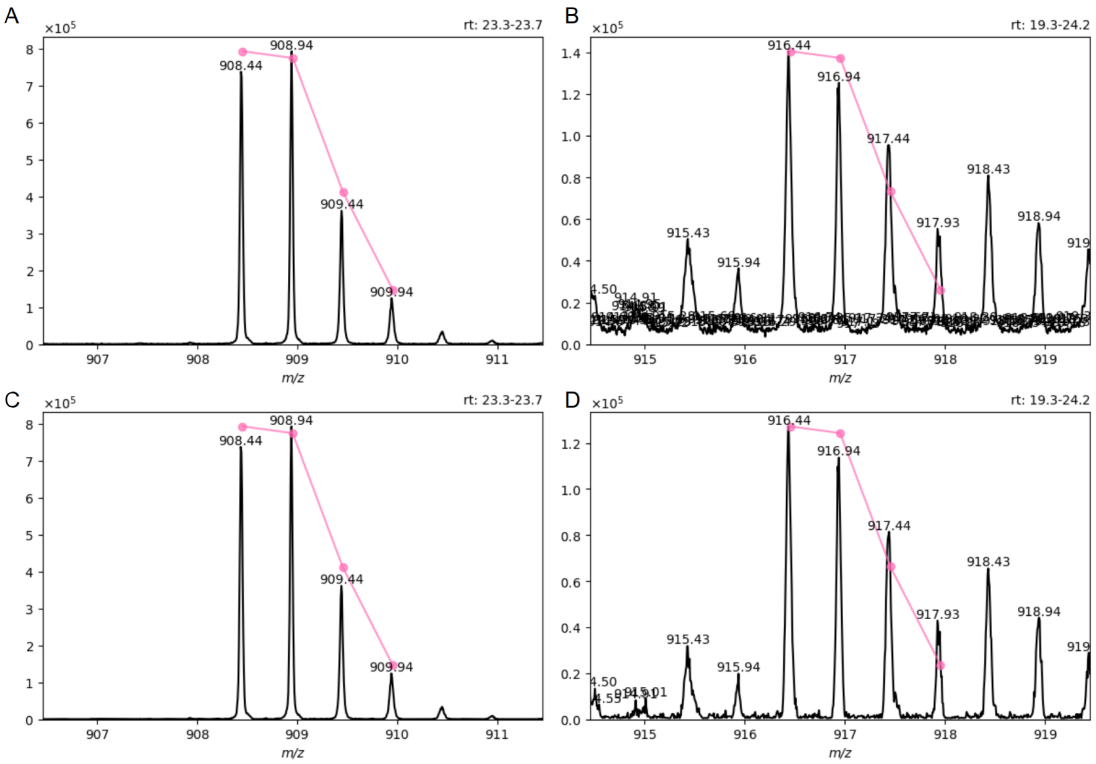


Figure 5.14 | Effect of absolute threshold on spectra quality for peptide 1-16, +16 oxidation. Spectra for unlabelled (**A**) and labelled (**B**) peptide from unfiltered data. Spectra for unlabelled (**C**) and labelled (**D**) peptide from data filtered with 500 absolute count per scan in MSConvert.

5.6 Analysis of ACPs by Carbene Footprinting

To investigate interactions between ACPs and their Ppant arm, carbene footprinting was performed on PsyACP1 *apo* and *holo* variants, as described in Section 2.3.9. The data from these experiments were then processed with PepFoot (aryldiazirine-TDBA, carbamidomethyl, AspN, peptide length = 5–40, charge range = 1–4, missed cleavages = 1 and mass tolerance = 20 mmu). Six peptides were identified, covering 79 % of the total protein, which all showed some decrease in labelling in the *holo* form, shown in Figure 5.15. One of these peptides, 59–75, showed statistically significant masking ($p = 0.01$) in the presence of the Ppant arm. This peptide corresponds to the majority of helix II. Due to the Ppant modification of peptide 47–58, comparison of this peptide between *apo* and *holo* was not possible using the same .pfoot file parameters. By manual inspection this peptide was not detectable in the *holo* data, even when accounting for possible carbamidomethylation of the Ppant thiol. Another peptide with somewhat statistically significant masking, 40–46 ($p = 0.08$), corresponds to part of the loop between helices I and II. These results show that the Ppant interacts with parts of the ACP in solution, however, without MS/MS data it is not possible to say which parts of these helices are involved in the interaction.

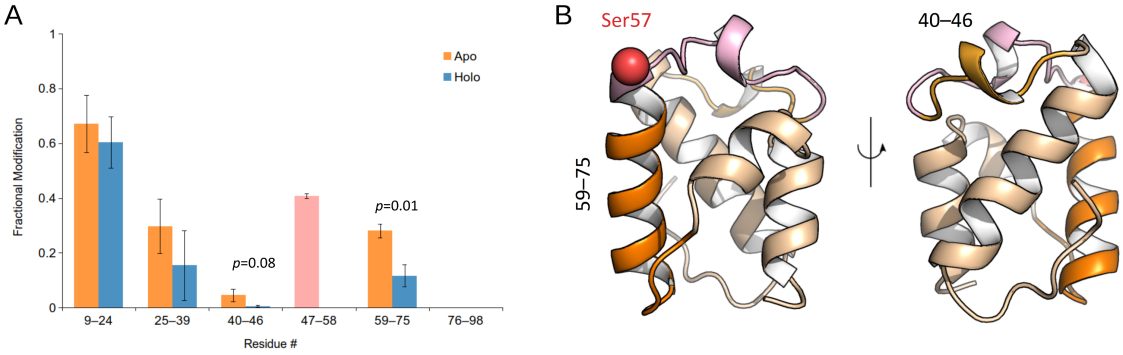


Figure 5.15 | Carbene footprinting of PsyACP1 *apo* versus *holo* variants shows masking of by Ppant arm. (**A**) Fractional modification bar chart, *apo* in orange and *holo* in blue. The Ppant modified peptide is highlighted in pink and significance values for two peptide displayed. (**B**) Significance of masking from (A) mapped to a model of PsyACP1. Peptides with significant masking are shown in dark ($p < 0.05$) and light ($p < 0.10$) orange. The Ppant modified peptide is shown in pink with the conserved Ser57 shown as red sphere.

5.7 Conclusions and Outlook

Covalent footprinting techniques provide a facile yet nuanced insight into protein interactions, however, they are marred by complex data and laborious manual processing. High-throughput use of these techniques provides a complement to biochemical assays and traditional structural biology approaches. Effective tools are needed to alleviate these bottlenecks, particularly at the analysis stage. An open-source software tool has been developed to tackle this challenge, providing a streamlined and interactive medium, accessible to non-expert users. PepFoot provides a user-friendly and generic platform for investigating all covalent labelling techniques with consistent and shareable output. Data processed through PepFoot is highly similar to purely manual analysis for the set of proteins investigated, but achieved in a fraction of the time and, in some cases, covering regions previously overlooked.

PepFoot currently focuses on peptide-level analysis, with residue level interrogation underway. Future iterations of PepFoot will also allow interaction ensembles to be filtered on the basis of changes in solvent accessibility. In combination with crosslinking and HDX software available, full interrogation of protein interactions is possible on a purely MS-based platform — ideal for high-throughput analysis. These data can complement traditional structural techniques by filtering out otherwise expensive and time-consuming experiments.

References

1. Thomson, J. J. *Rays of Positive Electricity and Their Applications to Chemical Analysis*; Longmands Green, London: London, 1913.
2. Aston, F. LXXIV. A positive ray spectrograph . *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2009**, *38*, 707–714.
3. Ringer, J. M. Detection of nerve agents using proton transfer reaction mass spectrometry with ammonia as reagent gas. *Eur. J. Mass Spectrom.* **2013**, *19*, 175–185.
4. Uetrecht, C.; Barbu, I. M.; Shoemaker, G. K.; van Duijn, E.; Heck, A. J. R. Interrogating viral capsid assembly with ion mobility-mass spectrometry. *Nat. Chem.* **2011**, *3*, 126–132.
5. Hoffmann, E. D.; Stroobant, V. *Mass Spectrometry: Principles and Applications*, 3rd ed.; John Wiley & Sons, Ltd: Chichester, 2007.
6. Loo, J. a. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **1997**, *16*, 1–23.
7. Zeleny, J. The electrical discharge from liquid points, and a hydrostatic method of measuring the electric intensity at their surfaces. *Phys. Rev.* **1914**, *3*, 69–91.
8. Zeleny, J. Instability of Electrified Liquid Surfaces. *Phys. Rev.* **1917**, *10*, 1–6.
9. Taylor, G. Disintegration of Water Drops in an Electric Field. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1964**, *280*, 383–397.
10. Taylor, G. The force exerted by an electric field on a long cylindrical conductor. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **1966**, *291*, 145–158.
11. Taylor, G. Electrically Driven Jets. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1969**, *313*, 453–475.
12. Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. Molecular Beams of Macroions. *J. Chem. Phys.* **1968**, *49*, 2240–2249.
13. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-.).* **1989**, *246*, 64–71.
14. Rayleigh, L. XX. On the equilibrium of liquid conducting masses charged with electricity. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1882**, *14*, 184–186.
15. Gomez, A.; Tang, K. Charge and fission of droplets in electrostatic sprays. *Phys. Fluids* **1994**, *6*, 404–414.
16. Iribarne, J. V. On the evaporation of small ions from charged droplets. *J. Chem. Phys.* **1976**, *64*, 2287.
17. De La Mora, J. F. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Anal. Chim. Acta* **2000**, *406*, 93–104.
18. Wilm, M. S.; Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int. J. Mass Spectrom. Ion Process.* **1994**, *136*, 167–180.

19. Zampronio, C. G.; Giannakopulos, A. E.; Zeller, M.; Bitziou, E.; Macpherson, J. V.; Derrick, P. J. Production and properties of nanoelectrospray emitters used in fourier transform ion cyclotron resonance mass spectrometry: Implications for determination of association constants for noncovalent complexes. *Anal. Chem.* **2004**, *76*, 5172–5179.

20. Juraschek, R.; Dülcks, T.; Karas, M. Nanoelectrospray - More than just a minimized-flow electrospray ionization source. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 300–308.

21. Benesch, J. L.; Ruotolo, B. T.; Simmons, D. A.; Robinsons, C. V. Protein complexes in the gas phase: Technology for structural genomics and proteomics. *Chem. Rev.* **2007**, *107*, 3544–3567.

22. Paul, W.; Steinwedel, H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift für Naturforsch. A* **1953**, *8*, 448–450.

23. March, R. E. An introduction to quadrupole ion trap mass spectrometry. *J. Mass Spectrom.* **1997**, *32*, 351–369.

24. Sobott, F.; Hernández, H.; McCammon, M. G.; Tito, M. A.; Robinson, C. V. A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal. Chem.* **2002**, *74*, 1402–1407.

25. Van Den Heuvel, R. H.; Van Duijn, E.; Mazon, H.; Synowsky, S. A.; Lorenzen, K.; Versluis, C.; Brouns, S. J.; Langridge, D.; Van Der Oost, J.; Hoyes, J.; Heck, A. J. Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. *Anal. Chem.* **2006**, *78*, 7473–7483.

26. Stephens, W. E. A Pulsed Mass Spectrometer with Time Dispersion. *Proc. Am. Phys. Soc.* **1946**, *69*, 691–691.

27. Mamyrin, B. A.; Karataev, V. I.; Shmikk, D. V.; Zagulin, V. A. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Sov. Phys. - JETP* **1973**, *37*, 45–48.

28. Comisarow, M. B.; Marshall, A. G. Convolution Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **1979**, *63*, 515–518.

29. Sleno, L.; Volmer, D. A. Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **2004**, *39*, 1091–1112.

30. Jones, R. Letter to the editors. *J. Geogr. High. Educ.* **1990**, *14*, 106–107.

31. Khatun, J.; Ramkissoon, K.; Giddings, M. C. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem.* **2007**, *79*, 3032–3040.

32. Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R. Influence of Basic Residue Content on Fragment Ion Peak Intensities in Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.* **2004**, *76*, 1243–1248.

33. Van Den Heuvel, R. H. H.; Heck, A. J. R. Native protein mass spectrometry: From intact oligomers to functional machineries. *Curr. Opin. Chem. Biol.* **2004**, *8*, 519–526.

34. Ruotolo, B. T.; Robinson, C. V. Aspects of native proteins are retained in vacuum. *Curr. Opin. Chem. Biol.* **2006**, *10*, 402–408.

35. Chernushevich, I. V.; Thomson, B. A. Collisional Cooling of Large Ions in Electrospray Mass Spectrometry. *Anal. Chem.* **2004**, *76*, 1754–1760.

36. Hopper, J. T. S.; Oldham, N. J. Alkali metal cation-induced destabilization of gas-phase protein-ligand complexes: consequences and prevention. *Anal. Chem.* **2011**, *83*, 7472–7479.

37. Susa, A. C.; Xia, Z.; Williams, E. R. Small Emitter Tips for Native Mass Spectrometry of Proteins and Protein Complexes from Nonvolatile Buffers That Mimic the Intracellular Environment. *Anal. Chem.* **2017**, *89*, 3116–3122.

38. Nguyen, G. T.; Tran, T. N.; Podgorski, M. N.; Bell, S. G.; Supuran, C. T.; Donald, W. A. Nanoscale Ion Emitters in Native Mass Spectrometry for Measuring Ligand-Protein Binding Affinities. *ACS Cent. Sci.* **2019**, *5*, 308–318.

39. Tolić, L. P.; Bruce, J. E.; Lei, Q. P.; Anderson, G. A.; Smith, R. D. In-Trap Cleanup of Proteins from Electrospray Ionization Using Soft Sustained Off-Resonance Irradiation with Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **1998**, *70*, 405–408.

40. Smith, R. D.; Loa, J. A.; Barinaga, C. J.; Edmonds, C. G.; Udseth, H. R. Collisional activation and collision-activated dissociation of large multiply charged polypeptides and proteins produced by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **1990**, *1*, 53–65.

41. Benesch, J. L. P. Collisional Activation of Protein Complexes: Picking Up the Pieces. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 341–348.

42. Hopper, J. T. S.; Sokratous, K.; Oldham, N. J. Charge state and adduct reduction in electrospray ionization-mass spectrometry using solvent vapor exposure. *Anal. Biochem.* **2012**, *421*, 788–790.

43. Cassou, C. A.; Williams, E. R. Desalting protein ions in native mass spectrometry using supercharging reagents. *Analyst* **2014**, *139*, 4810–4819.

44. Douglass, K. A.; Venter, A. R. Investigating the role of adducts in protein supercharging with sulfolane. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 489–497.

45. Clarke, D. J.; Campopiano, D. J. Desalting large protein complexes during native electrospray mass spectrometry by addition of amino acids to the working solution. *Analyst* **2015**, *140*, 2679–2686.

46. Hernández, H.; Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* **2007**, *2*, 715–726.

47. Ebong, I.-o.; Morgner, N.; Zhou, M.; Saraiva, M. A.; Daturpalli, S.; Jackson, S. E.; Robinson, C. V. Heterogeneity and dynamics in the assembly of the Heat Shock Protein 90 chaperone complexes. *Proc. Natl. Acad. Sci.* **2011**, *108*, 17939–17944.

48. Liu, L.; Kitova, E. N.; Klassen, J. S. Quantifying protein-fatty acid interactions using electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 310–318.

49. Cubrilovic, D.; Biela, A.; Sielaff, F.; Steinmetzer, T.; Klebe, G.; Zenobi, R. Quantifying protein-ligand binding constants using electrospray ionization mass spectrometry: A systematic binding affinity study of a series of hydrophobically modified trypsin inhibitors. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1768–1777.

50. Tahallah, N.; Pinkse, M.; Maier, C. S.; Heck, A. J. The effect of the source pressure on the abundance of ions of noncovalent protein assemblies in an electrospray ionization orthogonal time-of-flight instrument. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 596–601.

51. Cumeras, R.; Figueras, E.; Davis, C. E.; Baumbach, J. I.; Gràcia, I. Review on ion mobility spectrometry. Part 1: current instrumentation. *Analyst* **2015**, *140*, 1376–1390.

52. Mason, E. A.; Schamp, H. W. Mobility of gaseous Ions in weak electric fields. *Ann. Phys. (N. Y).* **1958**, *4*, 233–270.

53. Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Naked Protein Conformations: Cytochrome c in the Gas Phase. *J. Am. Chem. Soc.* **1995**, *117*, 10141–10142.

54. Pringle, S. D.; Giles, K.; Wildgoose, J. L.; Williams, J. P.; Slade, S. E.; Thalassinos, K.; Bateman, R. H.; Bowers, M. T.; Scrivens, J. H. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass Spectrom.* **2007**, *261*, 1–12.

55. Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. Ion mobility-mass spectrometry analysis of large protein complexes. *Nat. Protoc.* **2008**, *3*, 1139–1152.

56. Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Protein structure in Vacuo: Gas-phase conformations of BPTI and cytochrome c. *J. Am. Chem. Soc.* **1997**, *119*, 2240–2248.

57. Hopper, J. T. S.; Oldham, N. J. Collision Induced Unfolding of Protein Ions in the Gas Phase Studied by Ion Mobility-Mass Spectrometry: The Effect of Ligand Binding on Conformational Stability. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1851–1858.

58. Dixit, S. M.; Polasky, D. A.; Ruotolo, B. T. Collision induced unfolding of isolated proteins in the gas phase: past, present, and future. *Curr. Opin. Chem. Biol.* **2018**, *42*, 93–100.

59. Hyung, S. J.; Robinson, C. V.; Ruotolo, B. T. Gas-Phase Unfolding and Disassembly Reveals Stability Differences in Ligand-Bound Multiprotein Complexes. *Chem. Biol.* **2009**, *16*, 382–390.

60. Rabuck, J. N.; Hyung, S. J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Activation state-selective kinase inhibitor assay based on ion mobility-mass spectrometry. *Anal. Chem.* **2013**, *85*, 6995–7002.

61. Niu, S.; Ruotolo, B. T. Collisional unfolding of multiprotein complexes reveals cooperative stabilization upon ligand binding. *Protein Sci.* **2015**, *24*, 1272–1281.

62. Laganowsky, A.; Reading, E.; Allison, T. M.; Ulmschneider, M. B.; Degiacomi, M. T.; Baldwin, A. J.; Robinson, C. V. Membrane proteins bind lipids selectively to modulate their structure and function. *Nature* **2014**, *510*, 172–175.

63. Hopper, J. T. S.; Rawlings, A.; Afonso, J. P.; Channing, D.; Layfield, R.; Oldham, N. J. Evidence for the preservation of native interand intra-molecular hydrogen bonds in the desolvated FK-binding protein·FK506 complex produced by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1757–1767.

64. Tian, Y.; Han, L.; Buckner, A. C.; Ruotolo, B. T. Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and Structures. *Anal. Chem.* **2015**, *87*, 11509–11515.

65. Konermann, L.; Vahidi, S.; Sowole, M. A. Mass spectrometry methods for studying structure and dynamics of biological macromolecules. *Anal. Chem.* **2014**, *86*, 213–232.

66. Konermann, L.; Pan, J.; Liu, Y. H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **2011**, *40*, 1224–1234.

67. Sheff, J. G.; Rey, M.; Schriemer, D. C. Peptide-column interactions and their influence on back exchange rates in hydrogen/deuterium exchange-MS. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1006–1015.

68. Ferguson, P. L.; Pan, J.; Wilson, D. J.; Dempsey, B.; Lajoie, G.; Shilton, B.; Konermann, L. Hydrogen/deuterium scrambling during quadrupole time-of-flight MS/MS analysis of a zinc-binding protein domain. *Anal. Chem.* **2007**, *79*, 153–160.

69. Maleknia, S. D.; Brenowitz, M.; Chance, M. R. Millisecond radiolytic modification of peptides by synchrotron X-rays identified by mass spectrometry. *Anal. Chem.* **1999**, *71*, 3965–3973.

70. Xu, G.; Chance, M. R. Hydroxyl Radical-Mediated Modification of Proteins as Probes for Structural Proteomics. *Chem. Rev.* **2007**, *107*, 3514–3543.

71. Hambly, D. M.; Gross, M. L. Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2057–2063.

72. Gau, B. C.; Sharp, J. S.; Rempel, D. L.; Gross, M. L. Fast photochemical oxidation of protein footprints faster than protein unfolding. *Anal. Chem.* **2009**, *81*, 6563–6571.

73. Cheng, M.; Zhang, B.; Cui, W.; Gross, M. L. Laser-Initiated Radical Trifluoromethylation of Peptides and Proteins: Application to Mass-Spectrometry-Based Protein Footprinting. *Angew. Chemie - Int. Ed.* **2017**, *56*, 14007–14010.

74. Richards, F. M.; Lamed, R.; Wynn, R.; Patel, D.; Olack, G. Methylene as a possible universal footprinting reagent that will include hydrophobic surface areas: Overview and feasibility: Properties of diazirine as a precursor. *Protein Sci.* **2009**, *9*, 2506–2517.

75. Gómez, G. E.; Monti, J. L. E.; Mundo, M. R.; Delfino, J. M. Solvent Mimicry with Methylene Carbene to Probe Protein Topography. *Anal. Chem.* **2015**, *87*, 10080–10087.

76. Jumper, C. C.; Schriemer, D. C. Mass spectrometry of laser-initiated carbene reactions for protein topographic analysis. *Anal. Chem.* **2011**, *83*, 2913–2920.

77. Platz, M. S.; Modarelli, D. A.; Morgan, S.; White, W. R.; Mullins, M.; Celebi, S.; Toscano, J. P. Lifetimes of Alkyl and Dialkyl Carbenes in Solution. *Prog. React. Kinet.* **1994**, *19*, 93–137.

78. Jumper, C. C.; Bomgarden, R.; Rogers, J.; Etienne, C.; Schriemer, D. C. High-resolution mapping of carbene-based protein footprints. *Anal. Chem.* **2012**, *84*, 4411–4418.

79. Manzi, L.; Barrow, A. S.; Scott, D.; Layfield, R.; Wright, T. G.; Moses, J. E.; Oldham, N. J. Carbene footprinting accurately maps binding sites in protein-ligand and protein-protein interactions. *Nat. Commun.* **2016**, *7*, 1–9.

80. Manzi, L.; Barrow, A. S.; Hopper, J. T.; Kaminska, R.; Kleanthous, C.; Robinson, C. V.; Moses, J. E.; Oldham, N. J. Carbene Footprinting Reveals Binding Interfaces of a Multimeric Membrane-Spanning Protein. *Angew. Chemie - Int. Ed.* **2017**, *56*, 14873–14877.

81. Jenner, M.; Kosol, S.; Griffiths, D.; Prasongpholchai, P.; Manzi, L.; Barrow, A. S.; Moses, J. E.; Oldham, N. J.; Lewandowski, J. R.; Challis, G. L. Mechanism of intersubunit ketosynthase-dehydratase interaction in polyketide synthases. *Nat. Chem. Biol.* **2018**, *14*, 270–275.

82. Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat. Prod. Rep.* **2010**, *27*, 996–1047.

83. Helfrich, E. J. N.; Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat. Prod. Rep.* **2016**, *33*, 231–316.

84. Staunton, J.; Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **2001**, *18*, 380–416.

85. Heath, R. J.; Rock, C. O. The Claisen condensation in biology. *Nat. Prod. Rep.* **2002**, *19*, 581–596.

86. Smith, S.; Tsai, S.-C. The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat. Prod. Rep.* **2007**, *24*, 1041–1072.

87. Keatinge-Clay, A. The structures of type I polyketide synthases. *Nat. Prod. Rep.* **2012**, *29*, 1050–1073.

88. Hertweck, C. The Biosynthetic Logic of Polyketide Diversity. *Angew. Chemie Int. Ed.* **2009**, *48*, 4688–4716.

89. Keatinge-Clay, A. Crystal Structure of the Erythromycin Polyketide Synthase Dehydratase. *J. Mol. Biol.* **2008**, *384*, 941–953.

90. Trindade-Silva, A. E.; Lim-Fong, G. E.; Sharp, K. H.; Haygood, M. G. Bryostatins: biological context and biotechnological prospects. *Curr. Opin. Biotechnol.* **2010**, *21*, 834–842.

91. Haapalainen, A. M.; Meriläinen, G.; Wierenga, R. K. The thiolase superfamily: Condensing enzymes with diverse reaction specificities. *Trends Biochem. Sci.* **2006**, *31*, 64–71.

92. Shen, B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr. Opin. Chem. Biol.* **2003**, *7*, 285–295.

93. Hopwood, D. A. Genetic Contributions to Understanding Polyketide Synthases. *Chem. Rev.* **1997**, *97*, 2465–2498.

94. Yadav, G.; Gokhale, R. S.; Mohanty, D. Towards Prediction of Metabolic Products of Polyketide Synthases: An In Silico Analysis. *PLoS Comput. Biol.* **2009**, *5*, 1–14.

95. Cane, D. E.; Walsh, C. T.; Khosla, C. Harnessing the biosynthetic code: Combinations, permutations, and mutations. *Science (80-.).* **1998**, *282*, 63–68.

96. Bayly, C.; Yadav, V. Towards Precision Engineering of Canonical Polyketide Synthase Domains: Recent Advances and Future Prospects. *Molecules* **2017**, *22*, 235.

97. Hagen, A.; Poust, S.; De Rond, T.; Fortman, J. L.; Katz, L.; Petzold, C. J.; Keasling, J. D. Engineering a Polyketide Synthase for in Vitro Production of Adipic Acid. *ACS Synth. Biol.* **2016**, *5*, 21–27.

98. Hertweck, C. Decoding and reprogramming complex polyketide assembly lines: prospects for synthetic biology. *Trends Biochem. Sci.* **2015**, *40*, 189–199.

99. Franke, J.; Hertweck, C. Biomimetic Thioesters as Probes for Enzymatic Assembly Lines: Synthesis, Applications, and Challenges. *Cell Chem. Biol.* **2016**, *23*, 1179–1192.

100. Hertweck, C.; Luzhetskyy, A.; Rebets, Y.; Bechthold, A. Type II polyketide synthases: Gaining a deeper insight into enzymatic teamwork. *Nat. Prod. Rep.* **2007**, *24*, 162–190.

101. Shimizu, Y.; Ogata, H.; Goto, S. Type III Polyketide Synthases: Functional Classification and Phylogenomics. *ChemBioChem* **2017**, *18*, 50–65.

102. Walsh, C. T.; O'Brien, R. V.; Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew. Chemie - Int. Ed.* **2013**, *52*, 7098–7124.

103. Finking, R.; Marahiel, M. A. Biosynthesis of Nonribosomal Peptides. *Annu. Rev. Microbiol.* **2004**, *58*, 453–488.

104. Fischbach, M. A.; Walsh, C. T. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. *Chem. Rev.* **2006**, *106*, 3468–3496.

105. O'Brien, R. V.; Davis, R. W.; Khosla, C.; Hillenmeyer, M. E. Computational identification and analysis of orphan assembly-line polyketide synthases. *J. Antibiot. (Tokyo).* **2014**, *67*, 89–97.

106. Nguyen, T. A.; Ishida, K.; Jenke-Kodama, H.; Dittmann, E.; Gurgui, C.; Hochmuth, T.; Taudien, S.; Platzer, M.; Hertweck, C.; Piel, J. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **2008**, *26*, 225–233.

107. Gay, D. C.; Gay, G.; Axelrod, A. J.; Jenner, M.; Kohlhaas, C.; Kampa, A.; Oldham, N. J.; Piel, J.; Keatinge-Clay, A. T. A Close Look at a Ketosynthase from a Trans-Acyltransferase Modular Polyketide Synthase. *Structure* **2014**, *22*, 444–451.

108. Fisch, K. M.; Gurgui, C.; Heycke, N.; van der Sar, S. A.; Anderson, S. A.; Webb, V. L.; Taudien, S.; Platzer, M.; Rubio, B. K.; Robinson, S. J.; Crews, P.; Piel, J. Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* **2009**, *5*, 494–501.

109. Kampa, A.; Gagunashvili, A. N.; Gulder, T. A. M.; Morinaka, B. I.; Daolio, C.; Godejohann, M.; Miao, V. P. W.; Piel, J.; Andresson, O. S. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proc. Natl. Acad. Sci.* **2013**, *110*, 3129–3137.

110. Helfrich, E. J. N.; Ueoka, R.; Dolev, A.; Rust, M.; Meoded, R. A.; Bhushan, A.; Califano, G.; Costa, R.; Gugger, M.; Steinbeck, C.; Moreno, P.; Piel, J. Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **2019**, *15*, 813–821.

111. Jenner, M.; Frank, S.; Kampa, A.; Kohlhaas, C.; Pöplau, P.; Briggs, G. S.; Piel, J.; Oldham, N. J. Substrate specificity in ketosynthase domains from trans-AT polyketide synthases. *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 1143–1147.

112. Jenner, M.; Afonso, J. P.; Bailey, H. R.; Frank, S.; Kampa, A.; Piel, J.; Oldham, N. J. Acyl-Chain Elongation Drives Ketosynthase Substrate Selectivity in trans -Acyltransferase Polyketide Synthases. *Angew. Chemie Int. Ed.* **2015**, *54*, 1817–1821.

113. Ginolhac, A.; Jarrin, C.; Robe, P.; Perrière, G.; Vogel, T. M.; Simonet, P.; Nalin, R. Type I polyketide synthases may have evolved through horizontal gene transfer. *J. Mol. Evol.* **2005**, *60*, 716–725.

114. Jenke-Kodama, H.; Börner, T.; Dittmann, E. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium Streptomyces avermitilis. *PLoS Comput. Biol.* **2006**, *2*, 1210–1218.

115. Dorrestein, P. C.; Bumpus, S. B.; Calderone, C. T.; Garneau-Tsodikova, S.; Aron, Z. D.; Straight, P. D.; Kolter, R.; Walsh, C. T.; Kelleher, N. L. Facile detection of acyl and peptidyl intermediates on thiotemplate carrier domains via phosphopantetheinyl elimination reactions during tandem mass spectrometry. *Biochemistry* **2006**, *45*, 12756–12766.

116. Crump, M. P.; Crosby, J.; Dempsey, C. E.; Parkinson, J. A.; Murray, M.; Hopwood, D. A.; Simpson, T. J. Solution structure of the actinorhodin polyketide synthase acyl carrier protein from Streptomyces coelicolor A3(2). *Biochemistry* **1997**, *36*, 6000–6008.

117. Alekseyev, V. Y.; Liu, C. W.; Cane, D. E.; Puglisi, J. D.; Khosla, C. Solution structure and proposed domain domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci* **2007**, *16*, 2093–2107.

118. Roujeinikova, A.; Simon, W. J.; Gilroy, J.; Rice, D. W.; Rafferty, J. B.; Slabas, A. R. Structural Studies of Fatty Acyl-(Acyl Carrier Protein) Thioesters Reveal a Hydrophobic Binding Cavity that Can Expand to Fit Longer Substrates. *J. Mol. Biol.* **2007**, *365*, 135–145.

119. Meier, J. L.; Burkart, M. D. The chemical biology of modular biosynthetic enzymes. *Chem. Soc. Rev.* **2009**, *38*, 2012–2045.

120. Gong, H.; Murphy, A.; Mcmaster, C. R.; Byers, D. M. Neutralization of Acidic Residues in Helix II Stabilizes the Folded Conformation of Acyl Carrier Protein and Variably Alters Its Function with Different Enzymes. *J. Biol. Chem.* **2007**, *282*, 4494–4503.

121. Chandran, S. S.; Menzella, H. G.; Carney, J. R.; Santi, D. V. Activating Hybrid Modular Interfaces in Synthetic Polyketide Synthases by Cassette Replacement of Ketosynthase Domains. *Chem. Biol.* **2006**, *13*, 469–474.

122. Beld, J.; Sonnenschein, E. C.; Vickery, C. R.; Noel, J. P.; Burkart, M. D. The phosphopantetheinyl transferases: Catalysis of a post-translational modification crucial for life. *Nat. Prod. Rep.* **2014**, *31*, 61–108.

123. Gu, L.; Geders, T. W.; Wang, B.; Gerwick, W. H.; Hakansson, K.; Smith, J. L.; Sherman, D. H.; Håkansson, K.; Smith, J. L.; Sherman, D. H. GNAT-Like Strategy for Polyketide Chain Initiation. *Science (80-.).* **2007**, *318*, 970–974.

124. Jenner, M.; Afonso, J. P.; Kohlhaas, C.; Karbaum, P.; Frank, S.; Piel, J.; Oldham, N. J. Acyl hydrolases from trans-AT polyketide synthases target acetyl units on acyl carrier proteins. *Chem. Commun.* **2016**, *52*, 5262–5265.

125. Skiba, M. A.; Sikkema, A. P.; Moss, N. A.; Tran, C. L.; Sturgis, R. M.; Gerwick, L.; Gerwick, W. H.; Sherman, D. H.; Smith, J. L. A Mononuclear Iron-Dependent Methyltransferase Catalyzes Initial Steps in Assembly of the Apratoxin A Polyketide Starter Unit. *ACS Chem. Biol.* **2017**, *12*, 3039–3048.

126. Chun, S. W.; Hinze, M. E.; Skiba, M. A.; Narayan, A. R. Chemistry of a Unique Polyketide-like Synthase. *J. Am. Chem. Soc.* **2018**, *140*, 2430–2433.

127. Haines, A. S. *et al.* A conserved motif flags acyl carrier proteins for β-branching in polyketide synthesis. *Nat. Chem. Biol.* **2013**, *9*, 685–692.

128. Charkoudian, L. K.; Liu, C. W.; Capone, S.; Kapur, S.; Cane, D. E.; Togni, A.; Seebach, D.; Khosla, C. Probing the interactions of an acyl carrier protein domain from the 6-deoxyerythronolide B synthase. *Protein Sci.* **2011**, *20*, 1244–1255.

129. Vance, S.; Tkachenko, O.; Thomas, B.; Bassuni, M.; Hong, H.; Nietlispach, D.; Broadhurst, W. Sticky swinging arm dynamics: studies of an acyl carrier protein domain from the mycolactone polyketide synthase. *Biochem. J.* **2016**, *473*, 1097–1110.

130. Long, P. F.; Wilkinson, C. J.; Bisang, C. P.; Cortés, J.; Dunster, N.; Oliynyk, M.; McCormick, E.; McArthur, H.; Mendez, C.; Salas, J. A.; Staunton, J.; Leadlay, P. F. Engineering specificity of starter unit selection by the erythromycin-producing polyketide synthase. *Mol. Microbiol.* **2002**, *43*, 1215–1225.

131. Kresze, G. â.; Steber, L.; Oesterhelt, D.; Lynen, F. Reaction of Yeast Fatty Acid Synthetase with Iodoacetamide: 3. MalonylâĂŘCoenzyme A Decarboxylase as Product of the Reaction of Fatty Acid Synthetase with Iodoacetamide. *Eur. J. Biochem.* **1977**, *79*, 191–199.

132. Tang, Y.; Kim, C.-Y.; Mathews, I. I.; Cane, D. E.; Khosla, C. The 2.7-A crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proc. Natl. Acad. Sci.* **2006**, *103*, 11124–11129.

133. Wong, F. T.; Jin, X.; Mathews, I. I.; Cane, D. E.; Khosla, C. Structure and mechanism of the trans -Acting acyltransferase from the disorazole synthase. *Biochemistry* **2011**, *50*, 6539–6548.

134. Yadav, G.; Gokhale, R. S.; Mohanty, D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol.* **2003**, *328*, 335–363.

135. Marsden, A. F. A.; Caffrey, P.; Aparicio, J. F.; Loughran, M. S.; Staunton, J.; Leadlay, P. F.; Mark, S. Stereospecific Acyl Transfers on the Polyketide Synthase Erythromycin-Producing. *Science (80-.).* **2011**, *263*, 378–380.

136. Keatinge-Clay, A. T.; Shelat, A. A.; Savage, D. F.; Tsai, S. C.; Miercke, L. J.; O'Connell, J. D.; Khosla, C.; Stroud, R. M. Catalysis, specificity, and ACP docking site of Streptomyces coelicolor malonyl-CoA:ACP Transacylase. *Structure* **2003**, *11*, 147–154.

137. Jensen, K.; Niederkrüger, H.; Zimmermann, K.; Vagstad, A. L.; Moldenhauer, J.; Brendel, N.; Frank, S.; Pöplau, P.; Kohlhaas, C.; Townsend, C. A.; Oldiges, M.; Hertweck, C.; Piel, J. Polyketide Proofreading by an Acyltransferase-like Enzyme. *Chem. Biol.* **2012**, *19*, 329–339.

138. Keatinge-Clay, A. T.; Stroud, R. M. The Structure of a Ketoreductase Determines the Organization of the β-Carbon Processing Enzymes of Modular Polyketide Synthases. *Structure* **2006**, *14*, 737–748.

139. Zheng, J.; Gay, D. C.; Demeler, B.; White, M. A.; Keatinge-Clay, A. T. Divergence of multimodular polyketide synthases revealed by a didomain structure. *Nat. Chem. Biol.* **2012**, *8*, 615–621.

140. Keatinge-Clay, A. T. A Tylosin Ketoreductase Reveals How Chirality Is Determined in Polyketides. *Chem. Biol.* **2007**, *14*, 898–908.

141. Kusebauch, B.; Busch, B.; Scherlach, K.; Roth, M.; Hertweck, C. Functionally distinct modules operate two consecutive α, β->β,γ Double-bond shifts in the rhizoxin polyketide assembly line. *Angew. Chemie - Int. Ed.* **2010**, *49*, 1460–1464.

142. Moldenhauer, J.; Götz, D. C.; Albert, C. R.; Bischof, S. K.; Schneider, K.; Süssmuth, R. D.; Engeser, M.; Gross, H.; Bringmann, G.; Piel, J. The final steps of bacillaene biosynthesis in bacillus amyloliquefaciens FZB42: Direct evidence for β,y dehydration by a iraiw-acyltransferase polyketide synthase. *Angew. Chemie - Int. Ed.* **2010**, *49*, 1465–1467.

143. Gay, D. C.; Spear, P. J.; Keatinge-Clay, A. T. A double-hotdog with a new trick: structure and mechanism of the trans-acyltransferase polyketide synthase enoyl-isomerase. *ACS Chem. Biol.* **2014**, *9*, 2374–2381.

144. Kwan, D. H.; Sun, Y.; Schulz, F.; Hong, H.; Popovic, B.; Sim-Stark, J. C.; Haydock, S. F.; Leadlay, P. F. Prediction and Manipulation of the Stereochemistry of Enoylreduction in Modular Polyketide Synthases. *Chem. Biol.* **2008**, *15*, 1231–1240.

145. Ames, B. D.; Nguyen, C.; Bruegger, J.; Smith, P.; Xu, W.; Ma, S.; Wong, E.; Wong, S.; Xie, X.; Li, J. W.-H.; Vederas, J. C.; Tang, Y.; Tsai, S.-C. Crystal structure and biochemical studies of the trans-acting polyketide enoyl reductase LovC from lovastatin biosynthesis. *Proc. Natl. Acad. Sci.* **2012**, *109*, 11144–11149.

146. Piel, J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc. Natl. Acad. Sci.* **2002**, *99*, 14002–14007.

147. Piel, J.; Hui, D.; Wen, G.; Butzke, D.; Platzer, M.; Fusetani, N.; Matsunaga, S. Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. *Proc. Natl. Acad. Sci.* **2004**, *101*, 16222–16227.

148. Piel, J.; Wen, G.; Platzer, M.; Hui, D. Unprecedented Diversity of Catalytic Domains in the First Four Modules of the Putative Pederin Polyketide Synthase. *ChemBioChem* **2004**, *5*, 93–98.

149. Nakabachi, A. *et al.* Defensive bacteriome symbiont with a drastically reduced genome. *Curr. Biol.* **2013**, *23*, 1478–1484.

150. Cichewicz, R. H.; Valeriote, F. A.; Crews, P. Psymberin, a potent sponge-derived cytotoxin from Psammocinia distantly related to the pederin family. *Org. Lett.* **2004**, *6*, 1951–1954.

151. Pang, B.; Valencia, L. E.; Wang, J.; Wan, Y.; Lal, R.; Zargar, A.; Keasling, J. D. Technical Advances to Accelerate Modular Type I Polyketide Synthase Engineering towards a Retro-biosynthetic Platform. *Biotechnol. Bioprocess Eng.* **2019**, *423*, 413–423.

152. Zhang, F.; Shi, T.; Ji, H.; Ali, I.; Huang, S.; Deng, Z.; Min, Q.; Bai, L.; Zhao, Y.; Zheng, J. Structural Insights into the Substrate Specificity of Acyltransferases from Salinomycin Polyketide Synthase. *Biochemistry* **2019**, acs.biochem.9b00305.

153. Agarwal, V.; Diethelm, S.; Ray, L.; Garg, N.; Awakawa, T.; Dorrestein, P. C.; Moore, B. S. Chemoenzymatic Synthesis of Acyl Coenzyme A Substrates Enables in Situ Labeling of Small Molecules and Proteins. *Org. Lett.* **2015**, *17*, 4452–4455.

154. Lowry, B.; Robbins, T.; Weng, C. H.; O'Brien, R. V.; Cane, D. E.; Khosla, C. In vitro reconstitution and analysis of the 6-deoxyerythronolide B synthase. *J. Am. Chem. Soc.* **2013**, *135*, 16809–16812.

155. Watanabe, K.; Wang, C. C. C.; Boddy, C. N.; Cane, D. E.; Khosla, C. Understanding Substrate Specificity of Polyketide Synthase Modules by Generating Hybrid Multimodular Synthases. *J. Biol. Chem.* **2003**, *278*, 42020–42026.

156. Valenzano, C. R.; You, Y.-O.; Garg, A.; Keatinge-Clay, A.; Khosla, C.; Cane, D. E. Stereospecificity of the Dehydratase Domain of the Erythromycin Polyketide Synthase. *J. Am. Chem. Soc.* **2010**, *132*, 14697–14699.

157. Worthington, A. S.; Rivera, H.; Torpey, J. W.; Alexander, M. D.; Burkart, M. D. Mechanism-Based Protein Cross-Linking Probes To Investigate Carrier Protein-Mediated Biosynthesis. *ACS Chem. Biol.* **2006**, *1*, 687–691.

158. Thiele, G. A.; Friedman, C. P.; Tsai, K. J.; Beld, J.; Londergan, C. H.; Charkoudian, L. K. Acyl Carrier Protein Cyanylation Delivers a Ketoacyl Synthase-Carrier Protein Cross-Link. *Biochemistry* **2017**, *56*, 2533–2536.

159. Gulick, A. M.; Aldrich, C. C. Trapping interactions between catalytic domains and carrier proteins of modular biosynthetic enzymes with chemical probes. *Nat. Prod. Rep.* **2018**, *35*, 1156–1184.

160. Sánchez, C.; Du, L.; Edwards, D. J.; Toney, M. D.; Shen, B. Cloning and characterization of a phosphopantetheinyl transferase from Streptomyces verticillus ATCC15003, the producer of the hybrid peptide–polyketide antitumor drug bleomycin. *Chem. Biol.* **2001**, *8*, 725–738.

161. Kelley, L. a.; Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **2009**, *4*, 363–371.

162. Jeong, J. Y.; Yim, H. S.; Ryu, J. Y.; Lee, H. S.; Lee, J. H.; Seen, D. S.; Kang, S. G. One-step sequence-and ligation-independent cloning as a rapid and versatile cloning method for functional genomics Studies. *Appl. Environ. Microbiol.* **2012**, *78*, 5440–5443.

163. Chan, W.; Verma, C. S.; Lane, D. P.; Gan, S. K. A comparison and optimization of methods and factors affecting the transformation of Escherichia coli. *Biosci. Rep.* **2013**, *33*, 931–937.

164. Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89*, 5669–5672.

165. Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. *Anal. Chem.* **2015**, *87*, 11516–11522.

166. Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91*, 3147–3155.

167. Degiacomi, M. T. On the Effect of Sphere-Overlap on Super Coarse-Grained Models of Protein Assemblies. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 113–117.

168. Marklund, E. G. Molecular self-occlusion as a means for accelerating collision cross-section calculations. *Int. J. Mass Spectrom.* **2015**, *386*, 54–55.

169. Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem. Sci.* **2019**, *10*, 983–993.

170. Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. Collision Cross Sections of Proteins and Their Complexes: A Calibration Framework and Database for Gas-Phase Structural Biology. *Anal. Chem.* **2010**, *82*, 9557–9565.

171. Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; Leduc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical software to analyze top-down mass spectrometry data. *Proteomics* **2015**, *15*, 1235–1238.

172. Shevchenko, A.; Tomas, H.; Havli, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1*, 2856–2860.

173. Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11*, 996–999.

174. Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33*, 22–24.

175. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **2015**, *11*, 625–631.

176. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.

177. Sievers, F.; Higgins, D. G. Clustal omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **2014**, *1079*, 105–116.

178. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321.

179. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 1–9.

180. Eddy, S. R. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLoS Comput. Biol.* **2008**, *4*, e1000069.

181. Eddy, S. R. A New Generation Of Homology Search Tools Based On Probabilistic Inference. *Genome Informatics* **2009**, *23*, 205–211.

182. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, 1–16.

183. Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188–90.

184. Nielsen, M.; Lundegaard, C.; Lund, O.; Petersen, T. N. CPHmodels-3.0—remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Res.* **2010**, *38*, W576–W581.

185. DeLano, W. L. The PyMOL Molecular Graphics System. *Schrödinger LLC wwwpymolorg* **2002**, *Version 1.*, http://www.pymol.org.

186. Dryden, M. D.; Fobel, R.; Fobel, C.; Wheeler, A. R. Upon the Shoulders of Giants: Open-Source Hardware and Software in Analytical Chemistry. *Anal. Chem.* **2017**, *89*, 4330–4338.

187. Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J. Proteome Res.* **2019**, *18*, 709–714.

188. Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics - A python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 301–304.

189. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 99–104.

190. Rose, A. S.; Hildebrand, P. W. NGL Viewer: A web application for molecular visualization. *Nucleic Acids Res.* **2015**, *43*, W576–W579.

191. Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlic, A.; Rose, P. W. NGL viewer: Web-based molecular graphics for large complexes. *Bioinformatics* **2018**, *34*, 3755–3758.

192. Martens, L.; Hermjakob, H.; Jones, P.; Adamsk, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: The proteomics identifications database. *Proteomics* **2005**, *5*, 3537–3545.

193. Vizcaíno, J. A.; Côté, R.; Reisinger, F.; Barsnes, H.; Foster, J. M.; Rameseder, J.; Hermjakob, H.; Martens, L. The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.* **2009**, *38*, D736–D742.

194. Holman, J. D.; Tabb, D. L.; Mallick, P. *Curr. Protoc. Bioinforma.*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014; Vol. 46; pp 13.24.1–13.24.9.

195. Klaus, M.; Ostrowski, M. P.; Austerjost, J.; Robbins, T.; Lowry, B.; Cane, D. E.; Khosla, C. Protein-protein interactions, not substrate recognition, dominate the turnover of chimeric assembly line polyketide synthases. *J. Biol. Chem.* **2016**, *291*, 16404–16415.

196. Pistorius, D.; Müller, R. Discovery of the rhizopodin biosynthetic gene cluster in stigmatella aurantiaca sg a15 by genome mining. *ChemBioChem* **2012**, *13*, 416–426.

197. Zeng, J.; Wagner, D. T.; Zhang, Z.; Moretto, L.; Addison, J. D.; Keatinge-Clay, A. T. Portability and Structure of the Four-Helix Bundle Docking Domains of trans-Acyltransferase Modular Polyketide Synthases. *ACS Chem. Biol.* **2016**, *11*, 2466–2474.

198. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.

199. Farmer, R.; Thomas, C. M.; Winn, P. J. Structure, function and dynamics in acyl carrier proteins. *PLoS One* **2019**, *14*, 1–17.

200. Vander Wood, D. A.; Keatinge-Clay, A. T. The modules of trans-acyltransferase assembly lines redefined with a central acyl carrier protein. *Proteins Struct. Funct. Bioinforma.* **2018**, *86*, 664–675.

201. Sievers, F.; Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **2018**, *27*, 135–145.

202. Hughes, A. J.; Tibby, M. R.; Wagner, D. T.; Brantley, J. N.; Keatinge-Clay, A. T. Investigating the reactivities of a polyketide synthase module through fluorescent click chemistry. *Chem. Commun.* **2014**, *50*, 5276–5278.

203. Hall, Z.; Politis, A.; Bush, M. F.; Smith, L. J.; Robinson, C. V. Charge-state dependent compaction and dissociation of protein complexes: Insights from ion mobility and molecular dynamics. *J. Am. Chem. Soc.* **2012**, *134*, 3429–3438.

204. Covey, T.; Douglas, D. J. Collision cross sections for protein ions. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 616–623.

205. Mark, K. J.; Douglas, D. J. Coulomb effects in binding of heme in gas-phase ions of myoglobin. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 111–117.

206. Zhong, Y.; Han, L.; Ruotolo, B. T. Collisional and coulombic unfolding of gas-phase proteins: High correlation to their domain structures in solution. *Angew. Chemie - Int. Ed.* **2014**, *53*, 9209–9212.

207. Politis, A.; Park, A. Y.; Hyung, S. J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. Integrating ion mobility mass spectrometry with molecular modelling to determine the architecture of multiprotein complexes. *PLoS One* **2010**, *5*, 1–11.

208. Eschweiler, J. D.; Frank, A. T.; Ruotolo, B. T. Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1991–2000.

209. Wysocki, V. H.; Kenttämaa, H. I.; Cooks, R. Internal energy distributions of isolated ions after activation by various methods. *Int. J. Mass Spectrom. Ion Process.* **1987**, *75*, 181–208.

210. Klaus, M.; D'Souza, A. D.; Nivina, A.; Khosla, C.; Grininger, M. Engineering of Chimeric Polyketide Synthases Using SYNZIP Docking Domains. *ACS Chem. Biol.* **2019**, *14*, 426–433.

211. Wang, L.; Chance, M. R. Protein Footprinting Comes of Age: Mass Spectrometry for Biophysical Structure Assessment. *Mol. Cell. Proteomics* **2017**, *16*, 706–716.

212. Kavan, D.; Man, P. MSTools - Web based application for visualization and presentation of HXMS data. *Int. J. Mass Spectrom.* **2011**, *302*, 53–58.

213. Liu, S.; Liu, L.; Uzuner, U.; Zhou, X.; Gu, M.; Shi, W.; Zhang, Y.; Dai, S. Y.; Yuan, J. S. HDX-Analyzer: a novel package for statistical analysis of protein structure dynamics. *BMC Bioinformatics* **2011**, *12*, S43.

214. Guttman, M.; Weis, D. D.; Engen, J. R.; Lee, K. K. Analysis of overlapped and noisy hydrogen/deuterium exchange mass spectra. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1906–1912.

215. Kan, Z.-Y.; Walters, B. T.; Mayne, L.; Englander, S. W. Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc. Natl. Acad. Sci.* **2013**, *110*, 16438–16443.

216. Rey, M.; Sarpe, V.; Burns, K. M.; Buse, J.; Baker, C. A.; Van Dijk, M.; Wordeman, L.; Bonvin, A. M.; Schriemer, D. C. Mass Spec Studio for integrative structural biology. *Structure* **2014**, *22*, 1538–1548.

217. Gallagher, E. S.; Hudgens, J. W. *Methods Enzymol.*, 1st ed.; Elsevier Inc., 2016; Vol. 566; pp 357–404.

218. Lau, A. M. C.; Ahdash, Z.; Martens, C.; Politis, A. Deuteros: software for rapid analysis and visualization of data from differential hydrogen deuterium exchange-mass spectrometry. *Bioinformatics* **2019**, 1–3.

219. Kaur, P.; Kiselar, J. G.; Chance, M. R. Integrated algorithms for high-throughput examination of covalently labeled biomolecules by structural mass spectrometry. *Anal. Chem.* **2009**, *81*, 8141–8149.

220. Bellamy-Carter, J.; Oldham, N. J. PepFoot: A Software Package for Semiautomated Processing of Protein Footprinting Data. *J. Proteome Res.* **2019**, *18*, 2925–2930.

221. Deutsch, E. mzML: A single, unifying data format for mass spectrometer output. *Proteomics* **2008**, *8*, 2776–2777.

222. Deutsch, E. W. Mass Spectrometer Output File Format mzML. *Curr. Protoc. Protein Sci.* **2010**, 319–331.

223. Wilhelm, M.; Kirchner, M.; Steen, J. A. J.; Steen, H. mz5 : Space- and Time-efficient Storage of Mass Spectrometry Data Sets. *Mol. Cell. Proteomics* **2012**, *11*, O111.011379.

224. Mayer, G.; Jones, A. R.; Binz, P. A.; Deutsch, E. W.; Orchard, S.; Montecchi-Palazzi, L.; Vizcaíno, J. A.; Hermjakob, H.; Oveillero, D.; Julian, R.; Stephan, C.; Meyer, H. E.; Eisenacher, M. Controlled vocabularies and ontologies in proteomics: Overview, principles and practice. *Biochim. Biophys. Acta - Proteins Proteomics* **2014**, *1844*, 98–107.

225. Mayer, G. *et al.* The HUPO proteomics standards initiativemass spectrometry controlled vocabulary. *Database* **2013**, *2013*, bat009.

226. Van Der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

227. Colaert, N.; Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **2009**, *6*, 786–787.

A Appendix

A.1 Sequences

Sequences for all proteins used, biophysical properties may be found in Table 2.3. Non-PKS coded residues are in lowercase. The Ser for Ppant post-translational modification are underlined and in red. The Arg→Gln mutant position in PedD is underlined and in blue.

PsyACP1

mgsshhhhhhssglvprgshmRQRRDGGTRSVNPTEPVIHGDFSPHLERLVRSLLLDET
AFAWDRPLMEMGLDSADLLQLGERVASAFGVSPDPAFFFTHNTCKKILATLAPSTK

PsyACP2

mgsshhhhhhssglvprgshmDVSSALYKLLRETLARELHMAVEDIDDDRPFLDMGLDSV
IGVTWVRKLNERFGLSITVTKVYAHPTVCAMGHFLLQEESVRK

PsyACP3

mgsshhhhhhssglvprgshmTSSGELATVVRTTLMQLLELPTVDDDEAFQNYGLDSISA
TIFSNRLEQVLGQPVLPHWLIDYPTVSALAQQLEAVCV

PsyACP4

mgsshhhhhhssglvprgsTSSPKGNLTDIKGRLVKIFTDRLGLAREEIAEDESFMAMGL
SSINVVEFLEHINQCFDLTLSTGLIFEHDTLHALAAHIEACIP

PedACP4

mgsshhhhhhssglvprgshmasmtggqqmgrgsGHDAASSSSGIAQVIVNTVMEVLKLK
KLDPTQPFQNYGLDSISAMVLATRLEKRLNQQVQPQWLIDFASVEALSAHLLSQSRRRTG
DRSAMQETAQ

EcACP

mketaaakferqhmdspdlgtlvprgsmadigsSTIEERVKKIIGEQLGVKQEEVTNNAS
FVEDLGADSLDTVELVMALEEEFDTEIPDEEAEKITTVQAAIDYINGHQalehhhhhh

PsyKS1

mkhhhhhhhhgglvprgshggsefDQSARAEKGVAIVGMACRLPGGITTPEALWTVLAEG
RDVVGTVPAGRWVWPQETGPEHGDPGIDCGGFLDDIARFDAKLFRISPREAKVMDPQQRL
LLELAWSAFEDAGYSKDAVEGTKTGVFVGASGSDYRLLLEQHRVNIEPVMGTGTAVAVLP
NRISYFFDLQGPSLLIDTACSSSLVAIHEAVQALRAGSCEQALVGGINIMCHPAMTLAYY
KAGMLSPDGRCKTFDAEANGYVRSEGAIVMMLKPLSAAQRDGDRIYAVVKGSACNHGGQA
GGLTVPNPQQQTALLRAAWASARVTPDQLGYLEAHGTGTSLGDPIEVKGMQDAFRADDNI
AAATTCYLGSVKSNLGHLEAAAGIAGLMKLALCLYHRQLVSSLHVHTVNPKLGLEQTPFQ
IAQQVMAWPTLKSGQPSLTGVSSFGSGGTNAHVVVEGVEQVGPARAERPVVIRLSAPNVE
QLAIYARCLRDYLQGLPERARPPLSALAYTLSRRQPMAVSASYWARDEASLVSGLADIAA
GLVTSVEEERGLSFGEGPVIALPGYPFAETSFWFDKPEAQAAPARPAKVALEDPVVIARR
GLGIVSDVLTRS

PsyACP-KS1

mgsshhhhhhssglvprgshmTRQRRDGGTRSVNPTEPVIHGDFSPHLERLVRSLLLDET
AFAWDRPLMEMGLDSADLLQLGERVASAFGVSPDPAFFFTHNTCKKILATLAPSTKGREV
IVDQSARAEKGVAIVGMACRLPGGITTPEALWTVLAEGRDVVGTVPAGRWVWPQETGPEH
GDPGIDCGGFLDDIARFDAKLFRISPREAKVMDPQQRLLLELAWSAFEDAGYSKDAVEGT
KTGVFVGASGSDYRLLLEQHRVNIEPVMGTGTAVAVLPNRISYFFDLQGPSLLIDTACSS
SLVAIHEAVQALRAGSCEQALVGGINIMCHPAMTLAYYKAGMLSPDGRCKTFDAEANGYV
RSEGAIVMMLKPLSAAQRDGDRIYAVVKGSACNHGGQAGGLTVPNPQQQTALLRAAWASA
RVTPDQLGYLEAHGTGTSLGDPIEVKGMQDAFRADDNIAAATTCYLGSVKSNLGHLEAAA
GIAGLMKLALCLYHRQLVSSLHVHTVNPKLGLEQTPFQIAQQVMAWPTLKSGQPSLTGVS
SFGSGGTNAHVVVEGVEQVGPARAERPVVIRLSAPNVEQLAIYARCLRDYLQGLPERARP
PLSALAYTLSRRQPMAVSASYWARDEASLVSGLADIAAGLVTSVEEERGLSFGEGPVIAL
PGYPFAETSFWFDKPEAQAAPARPAKVALEDPVVIARRGLGIVSDVLTRS

PksM3

mgsshhhhhhssglvprgshmasmtggqqmgrgsAADFEPVAIVGISGRFPGAMDIDEFW
KNLEEGKDSITEVPKDRWDWREHYGNPDTDVNKTDIKWGGFIDGVAEFDPLFFGISPREA
DYVDPQQRLLMTYVWKALEDAGCSPQSLSGTGTGIFIGTGNTGYKDLFHRANLPIEGHAA
TGHMIPSVGPNRMSYFLNIHGPSEPVETACSSSLVAIHRAVTAMQNGDCEMAIAGGVNTI
LTEEAHISYSKAGMLSTDGRCKTFSADANGYVRGEGVGMVMLKKLEDAERDGNHIYGVIR
GTAENHGGRANTLTSPNPKAQADLLVRAYRQADIDPSTVTYIEAHGTGTELGDPIEINGL
KAAFKELSNMRGESQPDVPDHRCGIGSVKSNIGHLELAAGISGLIKVLLQMKHKTLVKSL
HCETLNPYLQLTDSPFYIVQEKQEWKSVTDRDGNELPRRAGISSFGIGGVNAHIVIEEYM
PKANSEHTATEQPNVIVLSAKNKSRLIDRASQLLEVIRNKKYTDQDLHRIAYTLQVGREE
MDERLACVAGTMQELEEKLQAFVDGKEETDEFFRGQSHRNKETQTIFTADEDMALALDAW
IRKRKYAKLADLWVKGVSIQWNTLYGETKPRLISLPSYPFAKDHYWVPAKEHSERDKKEL
VNAIEDRAACFLTKQWSLSPIGSAVPGTRTVAILCCQETADLAAEVSSYFPNHLLIDVSR
IENDQSDIDWKEFDGLVDVIGCGWDDEGRLDWIEWVQRLVEFGHKEGLRLLCVTKGLESF
QNTSVRMAGASRAGLYRMLQCEYSHLISRHMDAEEVTDHRRLAKLIADEFYSDSYDAEVC
YRDGLRYQAFLKAHPETGKATEQSAVFPKDHVLLITGGTRGIGLLCARHFAECYGVKKLV
LTGREQLPPREEWARFKTSNTSLAEKIQAVRELEAKGVQVEMLSLTLSDDAQVEQTLQHI

KRTLGPIGGVIHCAGLTDMDTLAFIRKTSDDIQRVLEPKVSGLTTLYRHVCNEPLQFFVL
FSSVSAIIPELSAGQADYAMANSYMDYFAEAHQKHAPIISVQWPNWKETGMGEVTNQAYR
DSGLLSITNSEGLRFLDQIVSKKFGPVVLPAMANQTNWEPELLMKRRKPHEGGLQEAALQ
SPPARDIEEADEVSKCDGLLSETQSWLIDLFTEELRIDREDFEIDGLFQDYGVDSIILAQ
VLQRINRKLEAALDPSILYEYPTIQRFADWLIGSYSERLSALFGGRISDASAP

PsyGNAT

mgsshhhhhhssglvprgshmDETGACHASVTHFVVRPYQMRYVQIEDLAALDRLEALCW
DTAIRTPRTGLEARLRNHPQDHVVLTAEGAVVGVIYSQRLAQVDALQGACAADVSERRHA
GGPVVQLLAVNIDPASQARRWGDELLEFMLQRCALLTGVHTVVGVTLCQRFHRQALPMET
YIHARTEDGGLVDPVLWFHELHGAQIIRPMPGYRPADQLNQGYGVLVHYGLA

PsyGNAT-ACP1

mgsshhhhhhssglvprgshmDETGACHASVTHFVVRPYQMRYVQIEDLAALDRLEALCW
DTAIRTPRTGLEARLRNHPQDHVVLTAEGAVVGVIYSQRLAQVDALQGACAADVSERRHA
GGPVVQLLAVNIDPASQARRWGDELLEFMLQRCALLTGVHTVVGVTLCQRFHRQALPMET
YIHARTEDGGLVDPVLWFHELHGAQIIRPMPGYRPADQLNQGYGVLVHYGLATRQRRDGG
TRSVNPTEPVIHGDFSPHLERLVRSLLLDETAFAWDRPLMEMGLDSADLLQLGERVASAF
GVSPDPAFFFTHNTCKKILATLAPSTK

PsyAR-GNAT-ACP1

mgsshhhhhhssglvprgshmMKDVKKLENATLLMEQYARAYTAGSAVLAAKRVGLLDLL
ASQKRVARSELQACCEGREFLDPCLEALHATGYLRSIGEQYVTLAMGGHIGDCPKEAAKL
YDLNPVDWVQKPKLQRVMAAWLMSRDDWLSAATRFCLLLAVAVEQSAHAFKAWEKPLRKA
FVAMCEQEGWGVKKGTSFSFTPLGHQLLDAAKKYVPFIEIGPHLAQWDRLMQGEGREVAA
GRRLGAARRSASWFNPADEAFKAAVLCYFAETDQDHPRYLLHVDGDGEVLSDLYRLIETK
VPEAASIELLACCPDAANAEMARKHLKPLPHRVWVDGDEAPAALLAAQGVDAQQVLFVGS
PSARWAQVLDDHPFGLVALQYHGLSAGWGEAPETLSAWRIAATLMGQRDAESALMAAARI
GLFPRQQPRTFSDETGACHASVTHFVVRPYQMRYVQIEDLAALDRLEALCWDTAIRTPRT
GLEARLRNHPQDHVVLTAEGAVVGVIYSQRLAQVDALQGACAADVSERRHAGGPVVQLLA
VNIDPASQARRWGDELLEFMLQRCALLTGVHTVVGVTLCQRFHRQALPMETYIHARTEDG
GLVDPVLWFHELHGAQIIRPMPGYRPADQLNQGYGVLVHYGLATRQRRDGGTRSVNPTEP
VIHGDFSPHLERLVRSLLLDETAFAWDRPLMEMGLDSADLLQLGERVASAFGVSPDPAFF
FTHNTCKKILATLAPSTK

PedC

MKDLQNIQNTHPVVWMFSGQGSQYFQMGRQLYEQDETFHAWMKSLDDNVRDYIGQSLLDI
IYDTGHERSLPFDRLIHTHPALFMVQYALAKSLLARGLPAPDFLIGASLGEFIAISLAGD
THVENILFNLIKQARLFDEYCNAGAMLLVIDHIDTFSTTPAFSKDCELAGINFDHCFVVS
GPRTGILQTRKSLTKQNIACQLLPVSIAFHSSWMDEVHEIFIQQFPEQICRRLHTPVISC
ALPVPEQLTRFSSTYWWHVIRQPIAFHLAINTFHQSSPNAVYLDLGPAGNMAAATKYNLP
SSIHYRILPTMTPFGRDLENIEIARLRLLELDQRlehhhhhh

PedD

mkhhhhhhhhgglvprgshggsefksylfpgqgsqhlgMGEQLFDRFPNIIEAANDILGY
SIKTLCLEDPQRQLRLTQYTQVALYVVNALTYRQHLQQGGGLPDFVAGHSLGEYNALESA
GVFSFEDGLRLVQKRGDLMSQAPRGAMAAILGISADSVAGILAEQGLTRIDIANYNAPTQ

TIISGLEADIRDAQAVFESCQAMYVPLNTSGAFHSRYMQSARDEFAQFLEAFEFRDPQIP
VVANVTAKPYVGTEVVRTLADQLTGSVRWLDSMRFLLDQGVTEFRELGPGDVLSKLVESI
RSSAMSKPVSEFAAENSQQLVDEWNRTCPIGSRVRVKGYDDILVTKSRAVLLFGHRAAIY
MENYQGYFALSEVEPLIEQQPLVEKVW

Svp

MIAALLPSWAVTEHAFTDAPDDPVSLLFPEEAAHVARAVPKRLHEFATVRVCARAALGRL
GLPPGPLLPGRRGAPSWPDGVVGSMTHCQGFRGAAVARAADAASLGIDAEPNGPLPDGVL
AMVSLPSEREWLAGLAARRPDVHWDRLLFSAKESVFKAWYPLTGLELDFDEAELAVDPDA
GTFTARLLVPGPVVGGRRLDGFEGRWAAGEGLVVTAIAVAAPAGTAEESAEGAGKEATAD
DRTAVPrshhhhhh

A.2 ACP Sequence Analysis

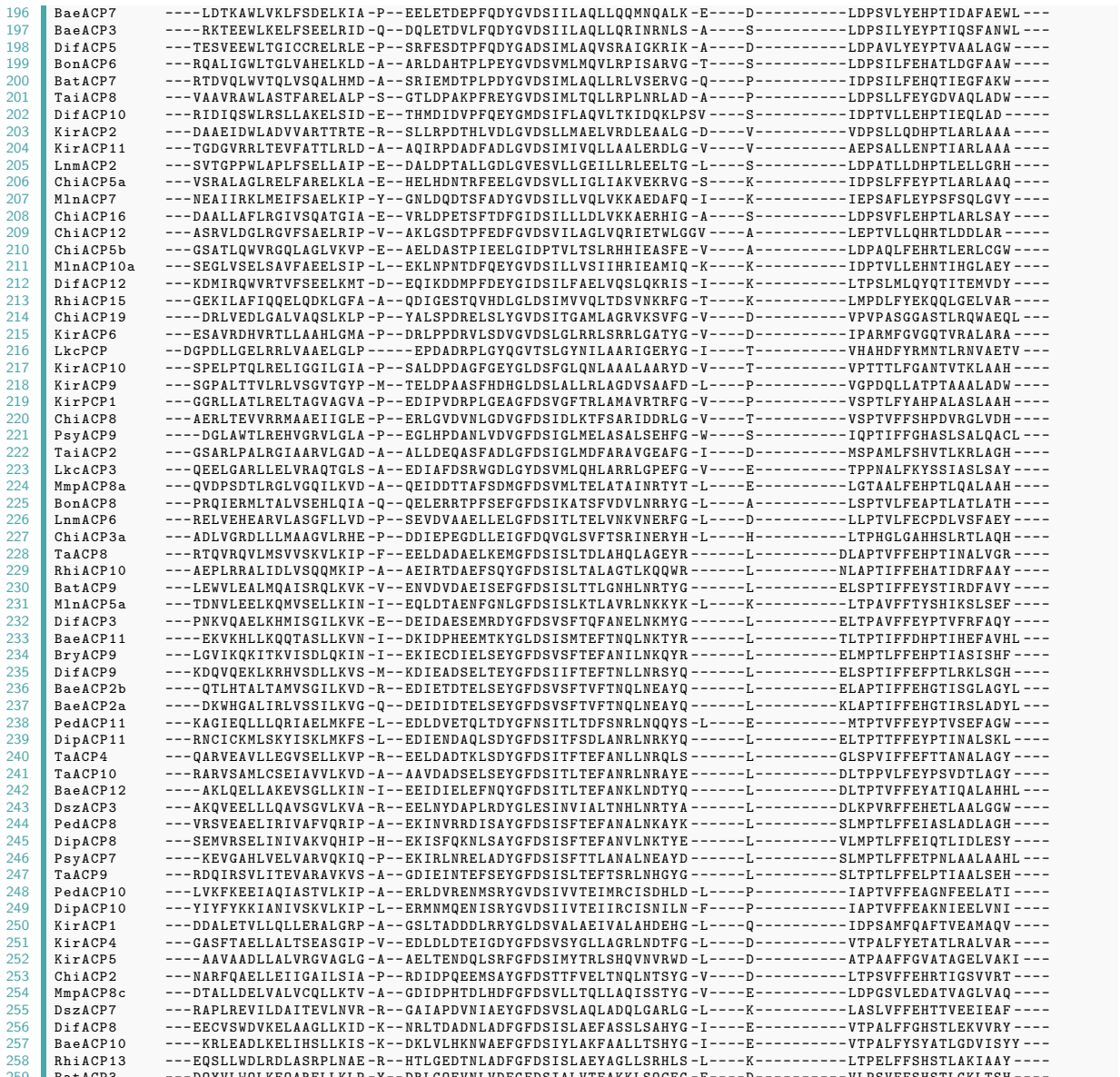
Domain	Clade	Homologue	%
Bat KS1	VI	Onn KS1	61
Bat KS2	XVI	Bae KS10	62
Bat KS3	II	Ped KS9	63
Bat KS4	VIII	GU KS10	68
Bat KS5	I	Ta KS8	66
Bat KS6	IX	GU KS11	70
Bat KS7	XIV	GU KS2	65
Bat KS8	XI	Bae KS8	68
Bat KS9	V	Ped KS11	73
Bat KS10	I	Ta KS8	64
Bat KS11	I	Ta KS8	66
Bon KS1		Lnm KS7/Onn KS1	49
Bon KS2	I	Bae KS6	66
Bon KS3	I	Dif KS14	63
Bon KS4	II	Dif KS8	62
Bon KS5	IX	Bae KS5	68
Bon KS6	XIV	GU KS2	63
Bon KS7	IX	GU KS11	68
Bon KS8	V	Ta KS9	64
Bon KS9	IX	GU KS11	71
Bon KS10	I	Dif KS14	64
Bon KS11	IX	Bae KS5	69
Bon KS12		Ped KS13/GU KS10	43
Bry KS1		Dsz KS5/Dif KS7	51
Bry KS2	II	Ped KS9	64
Bry KS3	VIII	GU KS10	69
Bry KS4	I	Ta KS8	67
Bry KS5	I	Dif KS14	66
Bry KS6	IX	Ped KS12	62
Bry KS7	II	Ped KS9	66
Bry KS8	II	Ped KS9	66
Bry KS9	IV	Ped KS8	61
Bry KS10	I	Dif KS14	66
Bry KS11	VIII	Ped KS6	49
Bry KS12	II	Ped KS9	64
Bry KS13	VIII	Onn KS6	61
Bry KS14		Onn KS1/Mln KS6	51
Bry KS15	IX	Ped KS12	61
Bry KS16	I	Ped KS3	53
Dip KS1	VI	Ped KS1	62
Dip KS2	VIII	Ped KS6	62
Dip KS3	I	Ped KS3	70
Dip KS4	III	Ped KS4	65
Dip KS5	XVI	Ped KS5	64
Dip KS6	VIII	Onn KS6	65
Dip KS7	I	Ped KS7	71
Dip KS8	IV	Ped KS8	68
Dip KS9	II	Ped KS9	70
Dip KS10	V	Ped KS10	75
Dip KS11	V	Ped KS11	74
Ena KS1		Ped KS1/Ery KS4	49
Ena KS2	I	Ery KS4/Bae KS6/Tai KS13	46
Ena KS3	VI	GU KS1	45
Ena KS4		Ery KS4/Mln KS6	48
Ena KS5		Ery KS4/Lnm KS5	47
Ena KS6	*cis*	Ery KS4/Lnm KS4	47
Ena KS7	IV	Ery KS4/Ped KS1/Onn KS1	52
Ena KS8		Ery KS4/Tai KS7	47
Ena KS9		Ery KS4/Tai KS13	50
Ena KS10		Ery KS4/Dsz KS6	49
Ena KS11	X	Ped KS13/GU KS11	58
Kir KS1	IV	Chi KS2	53
Kir KS2	XV	Mln KS8	57
Kir KS3	IX	Dsz KS7	61
Kir KS4	IV	Chi KS2	56
Kir KS5	IV	Chi KS2	58
Kir KS6	IV	Dsz KS4	53
Kir KS7	IX	Dsz KS7	60
Kir KS8	IX	Mln KS7	58
Kir KS9		Mln KS3/Ped KS11	53
Kir KS10	IV	Mln KS6	53
Kir KS11	XV	Chi KS16	63
Kir KS12	IX	Tai KS6	60
Kir KS13	*cis*	Ery KS4/Ped KS11	65
Kir KS14	*cis*	Ery KS4/Ped KS11	65
Nsp KS1	VI	Onn KS1	63
Nsp KS2	VIII	Onn KS6	66
Nsp KS3	I	Onn KS3	68
Nsp KS4	III	Onn KS4	64
Nsp KS5	XVI	Onn KS5	65
Nsp KS6	X	Bae KS13	61
Nsp KS7		Dsz KS6/Bae KS5	47
Nsp KS8	VIII	Ped KS6	50
Nsp KS9	XVI	Bae KS1	46
Rhi KS1	III	Bae KS14	47
Rhi KS2	X	Dsz KS9	49
Rhi KS3	IX	GU KS11	61
Rhi KS4	VII	Bae KS9	66
Rhi KS5	IX	Ped KS12	63
Rhi KS6	VII	Bae KS9	62
Rhi KS7	I	Ped KS7	66
Rhi KS8	VIII	GU KS10	67
Rhi KS9	I	Ped KS7	67
Rhi KS10	V	GU KS8	67
Rhi KS11	X	Bae KS13	55
Rhi KS12	IX	Bae KS5	47
Rhi KS13	II	Ped KS9	68
Rhi KS14	IX	GU KS11	49
Rhi KS15		Chi KS2/Bae KS2	54
Rhi KS16		Ped KS13/GU KS11	59

Table A.1 | Assignment of clades as described in Section 2.5.1.1. The nearest homologue is shown along with its sequence identity, where the two nearest homologue differ on clade none has been assigned and both homologues are listed. Domains with a *cis*-AT have been noted.

Listing A.1 | Aligned sequences for ACPs used in Section 3.5. Sequences were aligned using `clustalo` and sorted by order in a tree generated by PhyML.

```
BryACP10    ---KASVKQLLVEQLSQSLKLD-M--NEIHPDESFADYGVDSITGASFIQQLNDTLT-L----T----------LKTVCLFDHSSVNRLTAY----
BryACP5     ---KASVKQLLVEQLSQSLKLD-M--NEIHPDESFADYGVDSITGASFIQQLNDTLT-L----T----------LKTVCLFDHSSVNRLTAY----
MmpACP1     ---RPLLLASVRELVASSLQFE-L--CDIRDDEPFADYGVDSITGVALVRQLNARLA-I----D----------LHTTCLFDYPSINRLVDY----
MmpACP8b    ---RAQVAQGVREVVAEALKVR-L--EDIGDDDPWSDYGMDSVSSVQMTGLLNERFD-I----Q----------LAADTFQAFGNVVELTTA----
MmpACP3     ---FTQVRQEVMASVAKALKVA-L--EVIDPQESFSDYGLDSITGVNLVRVLNERLG-I----D----------LGTTALFDFSTATRLARH----
TaiACP5     ---HARIEAVIRDALAQALGVA-P--DAFELAVPLGEYGADSMLDLHLATRIEETLG-V----Q----------LSVRELFAHRTLGALRDH----
TaiACP4     ---ATRTAALLRSLLGEALKVE-P--VSIEGDGAFGDYGLDSIIGMGFVDAINRALD-L----S----------LDVTAVFEHNTVDALAAY----
BryACP14    ----QSIQKILIDFLVDITNFS-R--QDINPGRMPGDYGVDSLLGMRFLNRINSTFN-I----E----------ADALLLTEGTINSISHKVH--
RhiACP7     ---AGKLQQLIIDSLAKTLKAD-A--QSIALRTPFSDYGVDSILGVGFVKHLSQQLG-I----K----------LNSAILFEQATVAKLRDY----
PsyACP8     ----AHVEPVIVEALCRSLKLQ-S--RDVNRDLPFSDYGIDSILGIRFVKEVSATLG-I----R----------LNTAVIFDHTSVSRLASHI---
OnnACP7b    ---GELVHRLVVSHLAKVLDVA-E--STIEGDVPFSDYGLDSILGVNFITQINDDLG-L----E----------MNTTVIFDHTSVNALADH----
OnnACP7a    ---REQVTQLVLSQLAQTLEVP-E--NCIDLDVPFSDYGVDSILGVNFITQINDHMG-I----E----------MNTTVIFDHTTVHDLTAH----
DipACP7     ----NFIQNSILNALSESLIIQ-I--DIIKLNIPFSDYGIDSILGVNFIQKINNDLK-L----K----------LNTTILFDYTTVDSLSQYI---
PedACP7     ---AARVRELILDSLAQALSIG-R--EQIEQDIPFSDYGIDSILGVGFVQRLNDELG-L----S----------LNTTLLFDYTTVQRLAEH----
PsyACP6     ----DQIASHVRDALTRVLRME-N--GRIDPHVAFSDYGVDSILGADFIDRVNNALG-L----A----------LNTTILFEYTTLANLTQHI---
DszACP2     ---LQAVEDKIIDSLASTLQID-R--SKLSSDVPFTTFGVDSIFAVEVAGVIGRELS-I----D----------LRTTALFNYPTARALAEH----
ChiACP4     ---GRDVERAIVACLGEVLQTD-R--QRIDRDAPFTELGVDSLLSIRIVDRINKQLG-I----T----------LRTTDAYSYPSVRELAGH----
RhiACP9b    ---QEAAAALVVQGLASALQLD-A--AQIDNEQPFSDYGLDSILGASLVDHLNEALG-I----E----------LSAAILFDYPTVTTLSTY----
RhiACP9a    ---RDALTAVALNGLAKTLQLE-P--ERIMTDQPFADYGIDSILGVSFVGDLGEALG-I----E----------LNTALLFDYPTLDSLVDY----
TaACP12     ---GGELRTVVRDCLAACVEMA-P--AAVANDRRFTEYGVDSIIAVKLINAIGQRLG-L----V----------LETTLLFDHGTVDELADA----
DifACP14    ---ASHVRSVIKTSIADVLKMK-E--EDIQEELNFSEYGIDSILAVNVINEINKRLN-L----T----------LQTTILFDYYHVKLLTGY----
TaACP11     ---EERVRDVIKACVAKSLKVP-E--SQIQVDYSFADYGVDSIIAVNLSNAINDACG-T----S----------LPTTVLFDYNSVAALSQF----
BonACP10    ---RAVAKQAIRASLAATLKLD-D--AQLLDDQAFSSFGVDSITGVSLVNEINERLG-L----A----------LPTTVLFDYATVDQLSAY----
BonACP3     ---QLTARAVIRESIGQALKLG-E--SQLDDDEAFSNYGVDSITGVALVETINARLG-L----D----------LPTTVLFDYVSIEQLSRH----
ChiACP14    ---GHEVEARVTACVAATLDLP-V--ASIDPSAPFAEYGVDSILGVDLVRLLNEAFD-V----K----------LRTTVIFDWTSVRALARF----
ChiACP3b    ---KERIAHRVAAEVALALGST-V--AEVPTSAPFVEYGVDSIIGVDLVNRLNGAFG-T----A----------LRATVLYDYPDVNALSEL----
NspACP3b    ---SPSILNILIESLAEELFME-A--TEIDEEAKFVDMGMDSITAVTWMRKINQLFG-S----S----------LPATDVYKYPTLSDFARN----
NspACP3a    ---NQLVIQTLRETIAAELFLE-P--DEIDEDAKFVDLGLDSISAVMWIKRINAKFD-L----S----------LPATKVYSYPTLREFANY----
PedACP3b    ----RELLAWLKSSLSTELLLE-D--QPLDEEARFVDVGLDSITAVTWIKAINQRYG-L----S----------IGATKVYSYSCLAEFSQYV---
PedACP3a    ----QKVYAVIRDSLAEELQMD-G--ADIDPDTAFVEMGLDSIVAVTWVRKLNKAFG-L----S----------IGATKVYSYPDLTQFARFV---
DipACP3     ---NSRVFSILKNSLIEELQIK-S--EQFDDDAIFIDMGLDSIIAVTWIRKINSQLD-L----S----------INSTKIYDYPTLNKFFKF----
PsyACP2     ---SSALYKLLRETLARELHMA-V--EDIDDDRPFLDMGLDSVIGVTWVRKLNERFG-L----S----------ITVTKVYAHPTVCAMGHF----
OnnACP3c    ---AVAMMAELRQLLAEELHMA-A--EAIEDDVNFVEMGLDYVMAGSWVQKLNQAYG-L----S----------LEATVIYTYTNLLDLAGH----
OnnACP3a    ---MQQVRQELRRFLAEELHMT-P--DMVEEDIEFVKMGLDSIIAVSWVQKINQAFG-L----A----------LGATIVYTYTNLLDLLQH----
OnnACP3b    ---LKDVHVKLRQLLAEELHMT-P--EAVEDDVSFVEMGLDSIIAVSWIQKINQAYG-L----S----------LEATVVYTYTTLLDLAQH----
BryACP8     ---LQKIRSLLRESLALELDLD-A--DELDESISFTELGLDSINGVIWIRKINSHYE-L----S----------ITVSKVYDYPNIIELAEF----
BryACP4     ---LLQIQTMLRESLEFELDIE-P--GMLDELKPFTDLGLDSINGVTWIRKINSHYG-L----S----------MTATKVYDYPNIIELAEF----
BryACP17    ---VDMIRHSLRDTLADELYVN-A--EQLKDDMPFVDMGLDSIVGVTWIRKLSQKYH-V----S----------IDATKVYQYSSLRKMANY----
DifACP15b   ---DGTLRKELKDSLADILFLK-P--EDIDEHEAFIEMGLDSIIGVEWVQSINKTYQ-A----S----------ITANLVYEYPTIATLAGY----
DifACP15a   ---NRTLLEELRTSLADILFLS-P--EDIDADEPFIDMGLDSIIGVEWIQSVNKTYH-T----E----------VTANKVYEHPTLEELAEY----
TaiACP13b   ---PAALEAALRETLADALFAD-A--HDIEPDATFQDLGVDSIIGVEWVQAINRRYG-T----S----------IPAPQIYQYPTLRAFGGL----
TaiACP13a   ---PEAVVPALRALLADALYVD-A--QTIDANAEFVSLGVDSIIGVEWIDAVNRRFG-V----A----------LSATTIYDHPTLSSFARH----
BatACP12    ---LSKLEAELCEMLSSTLFMP-V--TAADCEKPFSELGVDSIIGLEWIQAVNRRYQ-T----S----------LTASMLYQHATLKRFTLF----
TaiACP7b    ---LDAIRAHLVDSLAQALYVE-P--AEIGVDQPFAELGLDSIVGVEWITAVNRRFG-T----A----------LPAVAIYDHPSVVALARF----
TaiACP7a    ---LLDIEAFLAGSLAAALMAT-T--DEIDREQTFNALGVDSIVGVEWVRAINDRYG-T----A----------LPATVIYDHPSVRAMARH----
BonACP12b   ---LGDLQRELAEELGEMLMLD-G--AAIDADTPFIDLGLDSITGVEWVRKINGRHG-L----S----------LTAMQMYEHPTLRALAKL----
BonACP12a   ---LGALQRELAEDLGAMLMLD-S--AAIDADTPFIDLGLDSIAGVEWVRRINARQG-V----S----------LTVMQMYEHATLRKLAAF----
BonACP2     ---LRALERELAASLAEALYLD-S--SEVDVDRPFAELGLDSIIGVEWMRAINRRHG-L----A----------LNATLVYEHPTVRRMAAR----
TaACP5      ---SSKLREELATSLAQALYID-R--AQVNAESTFVELGLDSIVGVEWIHAINKQYG-L----S----------LPVTKVYDHPNLSLFADY----
BatACP11    ---VADLRMELAESLAKVLYLE-N--DDVDMDTAFSELGLDSVTGVGWIRELNQRYG-T----S----------LGASQLYDYPTLKVLASH----
BatACP5c    ---HEELSAELVQGLAKVLYLE-H--DDIDRDTAFSELGLDSITGVEWVRDLNRRYQ-T----S----------LGASRLYDYSTIKMLASH----
BatACP5b    ---LDDFRSELAEGLAKVLYLE-G--QDVEMDTAFSELGLDSITGVEWVRNLNKRYD-L----A----------LVATRLYDYPTITRLADY----
BatACP5a    ---VDALRLALIEGLAKVLYLE-S--SNLDIDTAFSELGLDSITGVEWIRTLNQQYG-T----T----------LAATRLYDHPTVRALADY----
BaeACP6b    ----EAVHDELSKALADVLYME-R--HEVDIDEAFIDLGMDSITGLEWIKAVNKRYG-T----D----------CNVTKVYDYPTIRQFADFL---
BaeACP6a    ----EKVSAVLTESLADVLYMD-A--DDIDADDTFIDIGMDSITGLEWIKSVNKAYG-T----S----------LTVTKVYDYPTIRQFAAFL---
TaACP7      ---LEQLQAELTVSLARALFLQ-P--SEVDLDVGFRHIGLDSIISVEWIQSVNKSYG-I----N----------VAATRVYDYPNIREFSKY----
MmpACP7b    ---RESIQDYLKQSLGELLFLD-P--GQLRSGAQFLDLGMDSVTGTQWMRGVSRHFS-I----Q----------LAADAIYTWPTLKSLADE----
MmpACP7a    ---DDECAQFLRQSLAAMLYCE-P--GQIRDGSRFLELGLDSVIAAQWIREINKHYQ-L----K----------IPADGIYTYPVFKAFTQW----
BatACP10b   ---MSQVKEGLRRTLSTVLGLE-P--GDIADDDVFVDIGLDSIMAIQWVRVINSQFS-L----D----------LEAVRVYDYPTLDGLTDY----
BatACP10a   ---ASRVHKTLKQTLSSVLGID-I--LEIGDDDAFVDLGLDSIVGVQWVRIINDRFG-I----D----------IDAVRIYDYPSLTALVNY----
LnmACP7a    ---ASTVVGLLRGELSKILGMP-S--EEIENDAPFGELGLDSIYRMDLVRTLNEAFG-L----D----------LKATELYNYDTIGKLTEF----
TaiACP1     ---MPDIQRWCADYVADALGID-A--RDVNPDARIVDLGLGSLLACGMRCELNRLLP-D----------TASFDWGDPGWTLRTLARA
EnaACP8     ----EIARRHLVGIVRRIMALD-A-ARPLAQNKSFHELGLDSLMAIELKRALQEGFA-A----R----------VPATVMFDYPDIDSLAHW----
EnaACP5     ---RRALRARLREIVAACIGRD-A--AAITDTDGFAEIGIDSLHATVLHRQLEREFG-A----A----------LPATIAFDHPTVAAVADC----
EnaACP9     --RRRELGEYLEQTAGSLLRRP----GAIDGQASLFDQGLDSLLAIDLRGTLERRFE-Q----R----------FESTLLFDHPSVAALTEF----
EnaACP11    --RAARLKRHLEAAIRKLLNRA----DTLDDRASMFDLGLDSLLSIDLRMQLEKDLA-C----S----------LSTTVLHDHPTIEALAGF----
EnaACP3     ----RQMQAHAEAIVRKVLAID-A-GDAIDPARSLLELGMDSLLSVELRNRFAAQWG-L----S----------LPVSLMFDCPSVAAVSRR----
PsyACP12    ----RRVQVYVQEKLSALLRLP-L--EDMDITLSTLAMGMDSLVGLEFRQLISDELD-V----E----------LPLSLLLEGPSVEALAADV----
DszACP8     ---ARALEGYLCARLESTLGLD-Q--GEISARASLRRLGLDSILAAKLKVTLEGELA-M----T----------IPMEVLSGDKSVAELGDY----
RhiACP2     ------QPQGFRQILAGLLGCE-PAALDGCESRSLASLGLNSLAAVGLKAKLEQQLQ-L----S----------VSLAQLSPYLSLGEVETRL---
DifACP1     ---MKCINEFLRTVIAEKLGAE-V--TEAEDKTAFFTLGVTSLISEEIMTILHKEFT------G----------LSETLLFEYPNIEKLSEFL---
BryACP12a   ----IAISQRIRTILDKLLYLD-D-DDEIDEITNFVDIGITSMNLVEFISQLNREFS-L----D----------MKETVLFDYPNINTLSNM----
PsyACP11    -----AMETWIADQVAACLGFDRA--EDVDLTSGFFELGMESIQAQKLVQDCRAQTG-L----D----------LPDTLLFDYPTTRGLAEHL---
TaiPCP      -GLVGGLEGEVLALWRETLKVG-----DIGPTDGFFDAGGDSILAVALAARIEQRFG-V----T----------FSATTLFKYACVRDIAGY----
OnnPCP      -QPQSAIEESLLEIWQDILKID-----NISPTDGFFALGGNSVLAVRLAERIAQTLG-M----P----------FNPTALFKHASIREIGHY----
PsyPCP      --PQGELDRKILAIWQDLLKIE-----KLSTTDGFFEVGGSSVLAVVLADRICQAFD-V----A----------FTAPELFKYSSVQAIAAHL---
DipPCP      -ILKSKIQYELLNIWYNLLNIK-----GIKINDGFFAIGGNSVLSVILAQKISKNFN-I----K----------FHTSDLFKYSTIEDISLY----
PedPCP1     -LPCSDIEREVLDIWRSLLAVE-----GIGVSVGFFEVGGNSILSVMLAQQISEAFG-I----R----------FAATDLFKYPTIRDISLL----
NspPCP1     -TAKTETEAILADIWKNVLNVT-----DISINDGFFAAGGDSVLSVILADRIARRFN-Q----E----------FPVTSLFKYSTIRDIAAY----
TaPCP       -HARSPVEATLVELWKNVLQVN-----EVGVEDRFFEVGGDSVLAAVLVEEMNRRFD-T----R----------LAVTDLFKYVNIRDMARH----
BaePCP2     --PLSEAEAKVTAIWEDLLNTA-----GIKIEDGFFDVGGDSLLAVAAAERIKQELE-C----D----------FTVTDLFEHSTIQAISRFI---
BaePCP1     --SSSHIQQAVLAIWQDVLKSK-----DIELDDRFFDAGGDSLLAVTVTDRMTQELD-C----E----------FSVTELFEYATVKDISRYI---
BatPCP      -ASRTSIELRVRRIWQSVLITN-----NLGSDEGFFDAGGDSMSAVIVAARIKADFD-C----Q----------FDATDMFRFGTIRRIAEH----
NspPCP2     -MPKTEVEKIIAEVWQKALQVQ-----KVGICDNFFELGGHSLLLVQINQQLQAVLS-S----E----------LSILDMFKYPTIQTLSQY----
EnaACP12    -VSAAGIEPDLTAIWQALFALP-----AVGRHQDFFALGGDSQLGLRMLAQLRERHG-V----D----------LPLRCLYEAPTVARLAET----
PedPCP2     -APRTSMEIRLAEIWRRVLDID-----RVGVHDSFFDLGGHSLLALRLMSAIKQGLG-Y----E----------LPISSLFQAPTLTAQAEL----
NspACP10    -VPRNKLEQIIAGIWEDVLGFE-----QVGIHDDFFMLGGDSLVAVRVMAQIKNIFQ-V----E----------LPSATLLQKPTIAQLSQL----
KirPCP2     -VGGDALTATLTKLFAELLDQ-----PDLDADSHFFTSGGHSLLAAALAQRIKDETG-V----R----------LPLTDVFENPTPAALAGR----
ChiPCPb     -FAGQRLVDQLVGFFREAMNVA-----HVDPHQDLFDLGATSFTLVRIAQRLREQLD-L----S----------LSVDVLLKRPTVAGIAAF----
ChiPCPa     -AVKEKILRDIVALFELALGRK-----GLAPEADVFDLGATSLTLVRVSQKVREAFG-V----S----------LSVETILKHPTMAGIAAS----
DszPCP      -VSREAAEEELQRILGKALHLT-----RLDPGADLFELGATSLTIVQASQHIEERFG-V----G----------LPVEVVLAEPTLDAIARH----
RhiPCP      -INSQSLTAWLQQYFIEALGID-----ELAADDDLFDQGATSFTLVQAVHSIQQQHQ-V----T----------LPVDVFLEHPTIAAVTAY----
LnmPCP      -PSRDELCAEIADLFAQALGVE-----SVDADTDLWDQGATSFTMVRVSGSLQRSYK-Q----R----------FPVSALLDNPSVSAIAGW----
TaiACP17    ---AASVAASVEDAFEASLGLR-R--ADLERADDLRALGIGSIQAVMLAEAINERLD-L----A----------LPSRFVLEHATFGALVRA----
PsyACP4     ----TDIKGRLVKIFTDRLGLA-R--EEIAEDESFMAMGLSSINVVEFLEHINQCFD-L----T----------LSTGLIFEHDTLHALAAHI---
ChiACP10    ---LDAITTYLQGTIGDFLGLA----GSLPADQALVDFGVGSIAMAKLCYHLEQRLR-R----P----------FPNAVLAQAGTIDSLAGMI---
```

```
97  EnaACP7     ----AETGALLRESIAEVLDLPDP--GAIGAHDTLHALGMDSITLVELRDQLVRRLG-R----E----------LPSRLLFDFPQVGQLARY----
98  RhiACP12    ---REGIEDYLLAQLQALGAKQ-I--GKKQRHRNLMELGLESLQLVNLTSRLESAAA-I----E----------LEPTIFFEYPSVAELATF----
99  NspACP7     ---ATDVLRYLIDKINITAQIS-I--AETEVNTNLMQLGLGSTELVAIAASIEQDTD-L----E----------LNATLFFEYPSLMEVVNF----
100 BaeACP14    ----DDIRSYVLGKLAKTMDEP-A--PSLRSDANIMELGLDSVSLIGLTKEIGQEAD-I----D----------VNPTLFFEYPTAEELTRHF---
101 LkcACP2     -GLLVGTTRWLRELIVRHSGLR-G----LRDDENLLEQGLDSVGSIRVSRDIESCLG-IAGSSR----------LSRAVLFEYPTVAA--------
102 DszACP1     ---IARIEEDLRRLVSARIEAP-S--QAVDAEESFFSLGVDSVALQEITETLERTYG------S----------LPPTLLFENPNIRQLARYL---
103 DifACP13    ---VQQMEETLYRIFSTYVGRP-V--TRNDAHTVFFELGLESSQLLAIIKDIEGAFD-L----R----------LNPTLLFECSSISELLED----
104 KirACP12    -PAAQDADRIVRDLIEARTGR-----ADFDETAGFYELGLDSTALLGIAADLERVFG-V----D----------FSPTLLFEHNTFTLLAAH----
105 LnmACP7b   QQSSASLEDLVQDVIERELGR------TADPAKSFVDNGFGSFDMLRVVASLERVFG------A----------LRKTLLFDHPTIGALAAHL---
106 TaiACP9     ---GARAERFVRDAIAARLGVA-P--AAIDADAGFYDLGLDSGMLLELAETIGAAIG-A----S----------LAPTLLFEHANARELGAW----
107 BonACP7     ---RVGMEQFLRRIIGARLRVA-P--QRLDAQTSYYELGIDSAGMLELVQTIEARLG-M----T----------LSPTLLFEHVSIAELAAH----
108 BatACP8     ---YSDIERFLQGLIADKLGVA-P--EQIDRDSGYYELGVKSAGLLELVDDIEQKID-A----S----------LPPTLLFEYVTIGKLAAY----
109 BaeACP8     ----KDIILFLKKLLADKLGQP-W--ETLDVLAGYYELGLDSSSLLEIVQDISKKTG------A----------DLAPTLLFEYTNIKELAAYL---
110 BaeACP4     ----EEAERYVMNLMAEKIGKP-L--EQIDSQVGYYEMGLTSSGLLDVVETISEKIG-E----T----------LSPTLLFEYTTAAELAAFL---
111 DifACP6     ---TEQCILYLKELLAARLNKP-V--KEIDPDIGYYEMGLDSPGLLDMVKEIEQKIG------T----------QLLPTLLFEYTTIAELAGY----
112 MlnACP11b   ---HKTLTDELSSIIGEATNTP-P--NKADKHQSFYELGLDSKQLLHISKMLEERLD-T----Q----------LYPTLLFDYNNIHDLAAY----
113 MlnACP11a   ---KEMIRNNLIQEIAGMIEKP-A--ETVKNDSSFYEMGLDSKQLLDLSKKLELMLD-T----K----------LYPTLLFDYSTIGELTEY----
114 LnmACP3     ---PDAVRAHLRELVGTLLGRA-P--HAIRTDAGFYDLGLDSGHMLDISRRLEEYVC-A----P----------LYPTLLFEFSDIDSLAAH----
115 ChiACP6b    ---TAPLEDALRRKIAGLLGVP-P--AQVRSDSGFYELGLESTDLLSLARELEALLG-E----Q----------LYPTVLFEHPTIDALVAH----
116 ChiACP6a    ---AAALKARLRQLVAGRLGRD-P--ESVPTERGFYELGLESTDLLDLVRELEALLG-E----Q----------LYPTLLFEHATIDELAAY----
117 MlnACP8b    ---ALITEGDIQQLIADIKGTA-P--ETLSVVKGFYEQGLDSVTLMQAVKALEEKTG------R----------ELYPTLLFEYPTIQKLANY----
118 MlnACP8a    ---SESIETHLKELISGVSNID-L--RDIRTDIGFYEQGLDSASLLAIVKELENKTG------E----------QLYPTLLFEQKTIRELAAY----
119 KirACP3     ----ASLRAEIVELVRQVAADP-D-LLEVEGDRGFYDLGLESTDLLALVRSFEERWN-I----T----------LYPTLLFEYPTVDAVTEH----
120 ChiACP17    ---RALVERELIDMVASASGRP-A--DEISLEDGFYDQGLESTNLLQMVRDLERRLG-K----R----------LYPTLLFEYKSVRELTDH----
121 ChiACP13    ---RGAIVAELRAMLGGIHGEG-A--EEVPLDVGFYEQGLDSADLLRLVRTLEKRLQ-T----Q----------LYPTLLFEHTTVADLAAY----
122 DifACP11    ---EQNMVTYIKSEIARLINRP-A--QDIDVRKGFYDMGLDSAHLLKLVKILEQTFQ-Q----Q----------FYPTLLFEYSSVRRLADY----
123 TaiACP10    ---AAAAAAAVADILQSVTGLG-P--DEWRDDTTFDALGLDSLMIGEFTRRIEAMTG------E----------RDTTLLFRFRDLTALAAHL---
124 DszACP5     ---AQAVEDYLKGHFAAVFKMD-A--AQIDPQTSFDDYGIDSLVIVELHARLDKDMT----------------PLPRTTFFELRTVRAVADHL---
125 ChiACP11    ---LDAVERWLVQLFASVAEIP-A--RDVRAKTPLSDYGLNSLMAMGLYRLLSAEGL-A----G----------LPKTLFFQHRDLREVARH----
126 ChiACP15    ---LESTRTYVKGIFSDVLKID-P--ESIQDHRNFDAYGVDSLVVLNVTKRFEQRVG----------------ALPSTLLFEKLTVEDLAVFF---
127 ChiACP7     ---RERFVAYAKAIFAEVLQIP-P--GKIDPAATFETYGVDSIANLDLLRRFEVDLG----------------PLAPTLLFEHLTVDALAAYF---
128 LnmACP5b    ---PESVRSYVTGVFAEVLKYR-A--EDLDPAVTLENFGVDSLVSLNIVDRLEQDLG----------------DLPQTLLFEYTSIDSIAEYL---
129 LnmACP5a    ---AEAVRSYVAGVFAEVLKYE-S--AALDPEATFETFGVDSLVSLNIVDRFEQDLG----------------DLPQTLLFEYMTIDQVAGYF---
130 ChiACP18    ---QERCEEYVKGVIARVLKIA-G--SELDRDAGFDAYGVDSLLAMELADAFAKDLG----------------DMPSTLVFERRTVAEVATYL---
131 KirACP8     ---VEAARDYVKDAFVQVLEVP-R--DRLWLDETYENFGVDSLTVPRIADLLAERMG----------------DLPPTLLFEHPTIREVADHL---
132 BryACP1     ----NDLKKIIINVFAKVLARR-I--DENDLNSNLENYGLDSYAIINIVVELSKTFE----------------NVHYTILFEHRTINNIIDYLT--
133 TaACP6     ---QDAVSDHLKTVFGRVLKVP-R--QRLHDDVSLGNYGIDSINMLQVISELERDFG------P----------LSKTLLFEHRTLGELTAHL---
134 ChiACP9     ---ARWIRDELRGVLARTLKLD-P--ASVEDEVPLDALGVDSLVMMTLKRELSRDYP------D----------LPATLFLEHSTLAELTGYF---
135 DszACP9     ---RKATLGLLSSCFAEVAEIP-R--RSLDPEVPLDRYGLNSMLIAQLSARLEALLG------E----------LPKTLLFEHHTLAALTDWL---
136 DszACP6     ---APRAVGFLKRVFSEQWQLP-I--HRIDAEQSLDQYGLDSIMAMSLTRRLETFFG------E----------LPKTLLFEHQTIAALAGYL---
137 BatACP2     ---LEETLDRVIRRMSPVIGMH-M--EKIDVDTDFMVYGIDSILSMKVLHELELEFS------I---------QLSNSMLFEHRTIADLSTA----
138 RhiACP14    ---KKRIEAFVTRIIADVVQID-T--AKIHPSEKFEVYGFDSIMAVGVTNELEKYLG----------------SLPKTLLFEHVNIAGMVDHF---
139 KirACP7     ---TAAAEEFLRGVLAEVLQYP-Q--EEIEADTAFEQYGIDSLLIMDLTRRLEDHFG------S----------LPKTLFFEYQELRGLAGWF---
140 PedACP13    ---VRAVEGLLVRHLSELLKLP-E--HRIETDVPVEHYGIDSVGMMRLTVELEETFG------S----------LSKTLFFEYQDVQSLAAYL---
141 TaiACP12    ---VDASLALLTERLADVIKLP-A--PRIDPDAELTTYGIDSVVVMQVTTQLEKQLG------P----------LSKTLFFEYGTLRAIAARV---
142 MmpACP6     ---LEHALQVLKRVLSPVVQWP-E--DRLDSDEPLERYGLDSMMVMTITAALEAQFG------A----------LPKTLFFEYSTLRALAAYL---
143 TaiACP6     ---TEHALALLKRLLSTALHTP-A--SRLDAHAPLERYGIDSIVVVSMNGELEKAFG------S----------LSKTLFFEYRTLHELACYF---
144 PedACP12    ---SKRAIDYFKALLSSTLKFP-V--EEIAPDETMDAYGIDSIMVAELTSTLESHFG----------------PLSKTLFFEYQTLGELVDYF---
145 DipACP12    ----EKTIEYFKKIFSSLLKYP-I--DELDPNENIHSYGINSISIMELNIILEKKFG----------------TLSKTLFFEFNTLNSLSNYFI--
146 RhiACP16    ---TEPAIRYFKQLLSTTLKRP-V--EKIDSDGSFERYGVDSIILTELTTKLEQSFG----------------SLPNTLLYEYQTVRELSEYF---
147 TaACP1     ----EKVETKLKALFSEVTRYE-E--RRIDARQPMERYGIDSIIITQMNQALEGPYN------A----------LSKTLFFEYRTLAEVSGYLA--
148 RhiACP3     ---RQALENHIKRHFSKVSAIP-E--HNLDCEDTFDRYGLTSLMITTLNQRLAQEFG------E----------LSATLFFEYQSIAALAGYF---
149 BatACP4     ---QSRSLQRLRQLFSDVTKLG-A--ERIDVDEPLTAYGIDSLMILHLNRKLDEVFS------K----------LSRTLFFEYPTLRKLGEYL---
150 RhiACP8     ---IATTLERLKAVFCEVTKLP-P--ARVESDALLEQYGIDSIMITRLNKALEDAFS------RMTSLQPAERQLSKTLFFEY-------------
151 TaiACP11    ---AAAVAERLKHVLSEATGIA-V--SRLDADEPFDAYGVDSVVVMTVNRALDDVFD------A----------LSQTLLYEYGTLGALAGHL---
152 DifACP4     ---YEQTLLQLKTLFGLTTKIA-V--SNIDTEEPLETYGIDSVTVTKLNRELSLLFG------D----------LSKTVFYEYKTLGSLAAHF---
153 PsyACP5     ----AQVLDQLRRLFADVMRLS-V--DDVDVQAPLESFGLDSVVVTYLNQRLRDAFG------E----------VSATLFYEYRTLADLADFLA--
154 BryACP13    ---RQKVQRQFKGLLAEVIKLP-L--ERMDTQAPLERYGIDSLIVIQVNQALAAIFD------A----------LPKTLLFEYQTIDAVVAYL---
155 BryACP3     ---RQKVQRQFKGLLAEVIKLP-L--ERMDTQAPLERYGIDSLIVIQVNQALAAIFD------A----------LPKTLLFEYQTIDAVVAYL---
156 BryACP11    ---QDQVLQKFKELLSEHIQVP-A--ERLGSQQKFESFGIDSLIINQLNKDLSLVFG------E----------ISKTLFYEYRTLIELVDYL---
157 OnnACP6     ---ATRTLQEMKTLLGSVIGLV-P--DEIDAQKPLENYGLDSIAIIQLNEKLDGVFA------D----------LSKTLFYEYQTLDELVDSL---
158 PedACP6     ---KDKTIYQLKRLLAQVIGRA-V--ERIESCEPMDRYGLDSIAITQLNRKLEEQFG------G----------LSKTLFYQYQTVEALAEYL---
159 DipACP6     ---KQKIIFQLKILLSKILKTP-V--EKIQSTELMEKYGVDSIAITQINHELEKIFG------N----------ISKTLFYQYQTIDSLSNYL---
160 MlnACP1     ---LDQTVLFLTNIISSEVGAE-R--EEIDQDVHLAEYGMDSVMIHQCNTKLEQSLG------A----------ISKTLLFEFHIIRDLADHL---
161 MlnACP9     ---RQNIISYLKNIMSEELKLP-V--SLIDEKQYLEKYGIDSVMIMRLTKKLEQSIG------R----------LSKTLFFEYHSISQLAEYF---
162 DifACP2b    ---RERGITYLTELIGKTLKIP-S--GRMDPSVSLTAYGLDSILVVRLTNAFRNVFD----------------QMTSTLFFEYQTIEEIADYL---
163 DifACP2a    ---REKSVELLKKVIGEALKVP-A--HRLLAHEPLTAYGLDSILVVRLTNALKDVFG----------------SVSSTLFFEYQTIDELADYF---
164 RhiACP6     ---RRRVSDYLRKLVASTLKIP-A--DKMSLSRPLEDYGLDSILVVRLTNAFNEQFE----------------AIDSALFFEMPTLAALTEYF---
165 TaiACP15    ---RDAASAWLRARIGAALDIR-A--DDIDSRETLERYGIDSITVLQLTKRLRDAFD----------------DVPNTLLFEYNTVDALVDHW---
166 BaeACP9     ----EKSTEYLKTLIGETLKIP-P--AQIDPKAPLEKYGIDSIVVVQLTNALRNVLD----------------QVSSTLFFEYQTIEALVAHLI--
167 RhiACP4     ---KDQTVHYLKQMIGGLFKIR-P--EQIDAGEPLETYGIDSILVVQLTNLLRDKLS----------------GISSTLFFEYQTIDALADHF---
168 DifACP7b    ---LEKSTEYLKQIVSAILKMP-I--EDVQVHTPLEEYGIDSILIVQLTNRVREDVP----------------DISSTLFFEYPTIGGLAEYL---
169 LnmACP1     ---ADRAQEHVAACFAEVSGIP-A--EQLHPRVPLEHYGLSSRLVARFNERLRQDVQ------G----------VSSTVLFEYPDLAGVAAHL---
170 DifACP7a    ---SEKTAAFLIKAAAKVLKMP-A--EEINPLAPLDEYGIDSILIVQLTNELRDLVP----------------DISSTLFEYRTIEELAAFL---
171 OnnACP2     ---REACTHYFQTLIGKVLRIE-S--QRIDPAESLEAYGIDSILIVQIIEALQEIFE----------------DVSSTLFFEVQTVDALVTHF---
172 NspACP8     ---KEKGTTYFKQLIGEVLMMD-G--NRIDSSEPLAVYGVDSIIIAQLTNALNKNFE----------------GVKAGLLFEASTIDALVNHF---
173 NspACP2     ---RQKATQFFQQLVAKILRMN-P--EQIEADEPFQSYGIDSILVVQIAGALQEKFG----------------GIHSTLLFEVQTLGELVNYF---
174 PsyACP10    ----DRATVYFQRLFSEVLRIE-A--DKIDPQEPLGSYGLDSILITGVNQRLHDDFG----------------DVGSALLFKFQTLGALVHYFM--
175 PedACP2     ----QASYVHLKKIVARILRMD-S--GELAVREPLETYGIDSILIVQITNALREVFP----------------EIPSTLLFECQTLDALSDYLI--
176 DipACP2     ---LKSITLYLKKIIAKVLHMD-I--SKIDAQELLGNYGIDSILIIQITNKLRKDFS----------------NISSTLLFEYKNISELANYL---
177 DszACP4     ---LAQCRDYLRNLIAQATKLP-V--DKIHADRDLEDYGINSLMIMELNSMLDRDFD------A----------LPRTLFFEYKNVAELAAFF---
178 BaeACP5     ----EKTEHFLKQVLSEVTKLP-V--GKIDAKAPMEDYGIDSIMIMHLTGRLEKSFG------S----------LSKTLFFEYQNIESLTQYFM--
179 MlnACP6     ---KKQTVRLLTEIVGAETGRP-A--DSILPEDTFDRYGIDSLTILGLNSRLEKYFG------R----------VSKTLFFEYSSVEELSEYF---
180 MlnACP4     ---KAETSDFIIKVIAAEFRMP-A--DKIDAEEALEKYGLDSVMIINMTNELEKHFG----------------ALSKTLLFEYQTAAELSQYF---
181 MlnACP3     ---KEKVNDYLIKIISGVLKIQ-A--HKINVKDSFEKFGINSLMIMSLISKLEDTFG------D----------LPKTIFFEYQNIEELSNFF---
182 BaeACP13    ----DRTVQFIKETLAKSIKLA-P--ERIQPETSFEKYGIDSILQVNFIRELETLTG----------------ELPKTVLFEHTNTKELAEYLL--
183 NspACP6     ---QERTTEFVRQTLSEATRTP-A--NKVRLDTPFEQYGVDSIVQVGLIQKLEQVTG----------------SLPNTILFEYSTTRELVDYL---
184 RhiACP11    ---AKQCRQLITEVLSKLVNIA-P--ERLQRDVAFDRYGIDSILQISLIQALEKATG----------------ELSRTILFEYNTINQLTDYL---
185 BryACP16    ---EERVFQYFKTLISSTLKRP-R--HRIQLDEALENYGIDSLLSIEMIIELEKVFG----------------SLPKTLFEYQTIRELGHYF---
186 BryACP6     ---EQKTLDYLKNVFATTTQIP-E--KNIYVHETLDKYGVDSLLVMKMTNQLEKVFG----------------KLSKTLFFEYQTIRELGDYF---
187 BonACP9     ---LARAQHYLVGLLSHSLRMP-P--QRIDAQAPLEQYGIDSVLAVTLTQDLEQVFG------P----------LSKTLFFEYQTIAALTRFF---
188 BonACP5     ---LEPAQRYLVGLLSRSLKLP-P--QRIDAQAPLEQYGIDSVLAITLTQDLEQVFG------P----------LSKTLFFEYQTIAALTRYF---
189 BonACP11    ---STRARRYLTGLMAGGLKLS-P--ERIDAEEPLERYGIDSVLAVKLVQDLEQVFG------P----------LSKTLFFEYQTIAALTQYF---
190 BatACP6     ----QEKTLRFLTTLLSAPLKLA-A--DRIEPDVGLEKYGIDSILALQLVSELERPFG------P----------LPKTLMFECQSVAELAGYF---
191 RhiACP5b    ---CGSLNAQLRAMLASTLGIA-L--EQIDDAQCFEQYGLDSLKIMEMTSVLEKRFG------A----------LPRTLLFEHSSLQQLAHYL---
192 RhiACP5a    ---GDDVLMGLRRLLADTIKID-E--SRLDNDEPFEKYGIDSLMVMDMTAALEKRFG------P----------LPKTLFFEYPSLQSLADHL---
193 MmpACP2     ---RDFIIGLLRTVLVTVTRVS-P--AQLQDDRHFEHYGIDSLMVIQMTAELEKSFG------S----------LSKTLFFEHSSIKELADYF---
194 TaiACP3     ---RARVGARLSALLADVLKVP-V--ARLEPDAPFERYGIDSLTVVALNESLGRHVD------A----------LPKTLFFEYRTLAELTDYF---
195 TaiACP14    ---RAPVLARLKAMFADAFKMS-P--DDVLETAEFADYGIDSIVVLRMTRQLEKHFG------P----------LSKTLLFEQRRLDDLADYF---
```



```
EnaACP2     ---AALARRLVAEQIALVTGIA-A--ATIEPAAPLSILGMDSLMSVALSDALAHCLG-I----A----------ASATLLFDHPTLDALALH----
LkcACP1     ---RGAVEGFVREVIAGTMRLD-V--SEVECDRPFRDMGIDSLISMELLKPLRERVG-Y--------------VPATVWFEYPTVDRLVEFL---
EnaACP10    ---QDLVARLVRELLVRFLKID-A--GAVSDERPFFELGMDSVSALEFSDELGACFA-L----D----------LHVDTIFDYPSVASLSAY----
EnaACP6     ----SMLDELLHALLREVLGLS-G-NFAAYAGTGFHDLGMDSLLTLSFAEKLGARVG-L----P----------VSSVDVFDNANPARLRGW----
PsyACP1     ---HGDFSPHLERLVRSLLLDE----TAFAWDRPLMEMGLDSADLLQLGERVASAFG-V----S----------PDPAFFFTHNTCKKILATL---
BatACP1     ----GHVGQYVSRLIRSRLQLD-N-DVPLDLDQPLMEMGLDSADLLELATSISGEFN-V----T----------LTAMFFFENNTCNKVIVA----
NspACP1     ----ELATAFLEATVQSCLGDE-K-RSGYKFDRPLMEMGLNSVDLLILSEQISDRFG-L----I----------LDPAFFFSYTTPEKVLMF----
PedACP1     -----RICSQVCASISDALGEA-R-KSVFATDRPLMEMGLDSADLLGLKEAISSHFK-L----A----------LEPTFFFRCNTADKIIDYL---
OnnACP1     ----EAISDYLEEAVKTILGPE-Q-ECLFSKETPLMEMGLDSADFLGLREQIAVRFQ-T----A----------LEPTFFFQSKTPQDIISY----
DipACP1     -----EISDYIKKIIYKCLSFE-F-KKFFSIDTPLMDMGLDSSDLLFLSKNLSSHYQ-L----N----------IDSTFLFYNNTVNKIISYF---
RhiACP1     ----VSIDDAVRQGVNRCLGQS-E-DSRLSSRHSLMELGLDSADLLALGEQLGPMFG-L----T----------LEPTFFFRHNSIDKIVAA----
TaACP13     ---STETEYYLRQVVARIVECD-A--AAVRLDLPLMELGVDSVGLMELAEHIGRRFE-L----K----------LEETFFFEHNTLDRIGAF----
BonACP1     ----GSLARAVRETIEQVLGEA-R-VDSYGPHTPLMSMGLSSFDLAELRKRLGARLG-M----T----------LDATFLLRHGTPARLLEA----
BryACP15    ----TSLHDSIEKIFYKILGDK-R-KLGFSPKRPLMELGLDSIELLELRSLLGKHFS-I----K----------LEPTFLFQYETMAAVIEY----
EnaACP1     --RERDTLEVVSESVRRIMRVP----DSFAADCPLRELGLDSMGLMELRLLLGAAFS-I----E----------FDPAAFFSYPTARAIAGH----
LnmACP4     ---PQEVLWAVVDAVRTRLYLE-R--DEVDHRLSFNEMGVDSVGAVEIVEQLGARFA-L----E----------MDPVTLFDHPTVPRLAEH----
ChiACP1     ---KAALVGALRRAAAELLGRA-K--EDIDPGVPFRTLGLDSVGAIALLRRLSAEID-R----A----------LPLTLLWSYPTLESLATH----
KirACP14    ----RVLLDLVLGHAAAVLGYASP--DMIDPDRAFKDAGFGSLSAVQLRNRLAGAAG-M----R----------LPATLVFDHPNPAALARH----
KirACP13    ----ARLSGLVRGEAAAVLGYAGA--AAIDPGRSFQELGFDSLSSVEFRNRLGAALG-L----T----------LEATLVFDHPTPADLVEH----
KirACP15    ----AEVLTGVRGLAAAVLGHDSG--AGIDPADDFFDLGLSSVTALELRDHLGTLTG-L----E----------WPADVLYECPTPQALADL----
EryACP4     ----ENLLELVANAVAEVLGHESA--AEINVRRAFSELGLDSLNAMALRKRLSASTG-L----R----------LPASLVFDHPTVTALAQH----
```

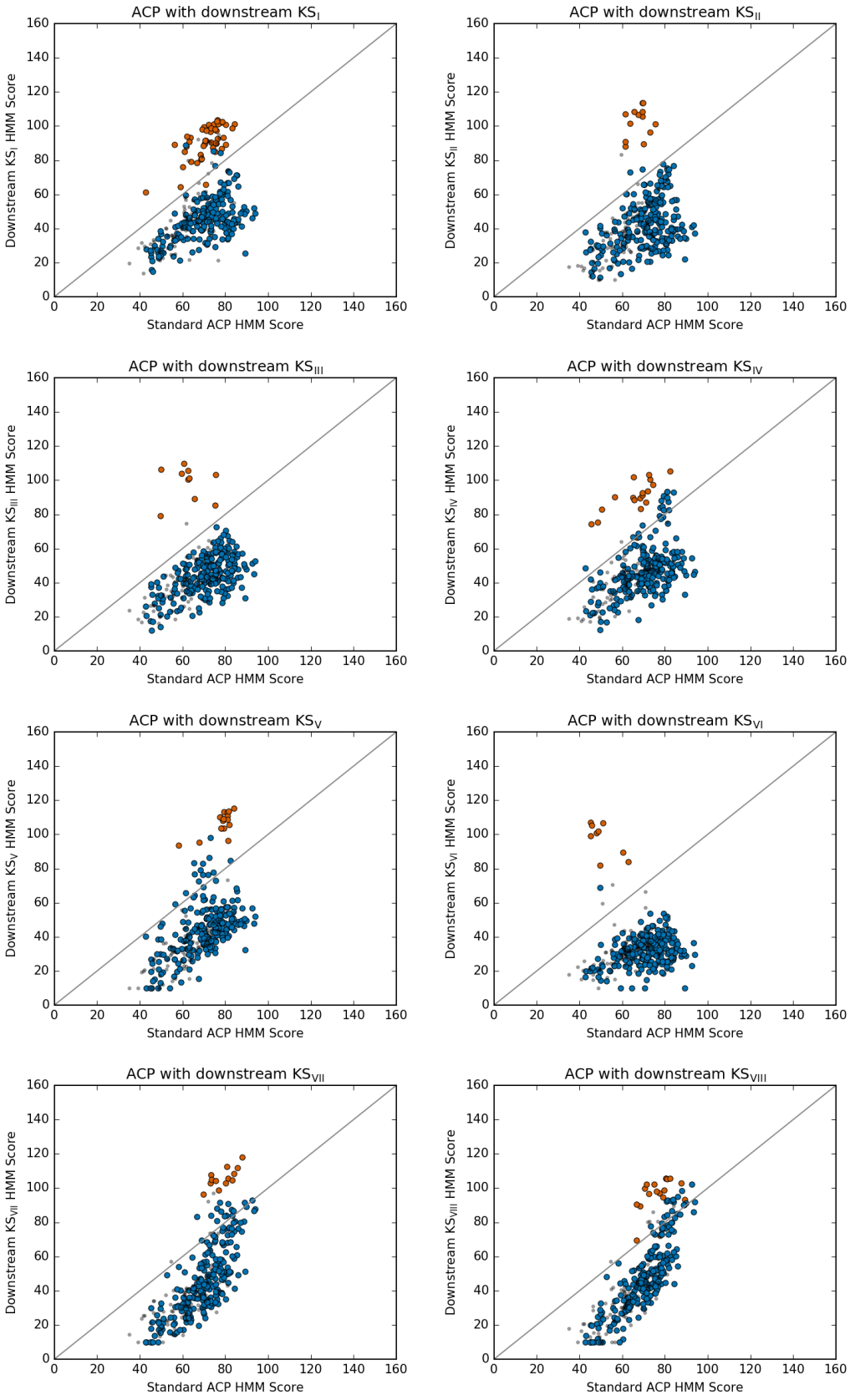


Figure A.1 | Plot of bit scores for ACPs scored against **downstream** KS clade grouped HMM profiles and a standard ACP HMM profile. ACPs from the group used to generate each profile are shown as orange dots. ACPs from other groups are shown as blue dots. ACPs with no assigned group are shown as grey dots ($n = 73$).

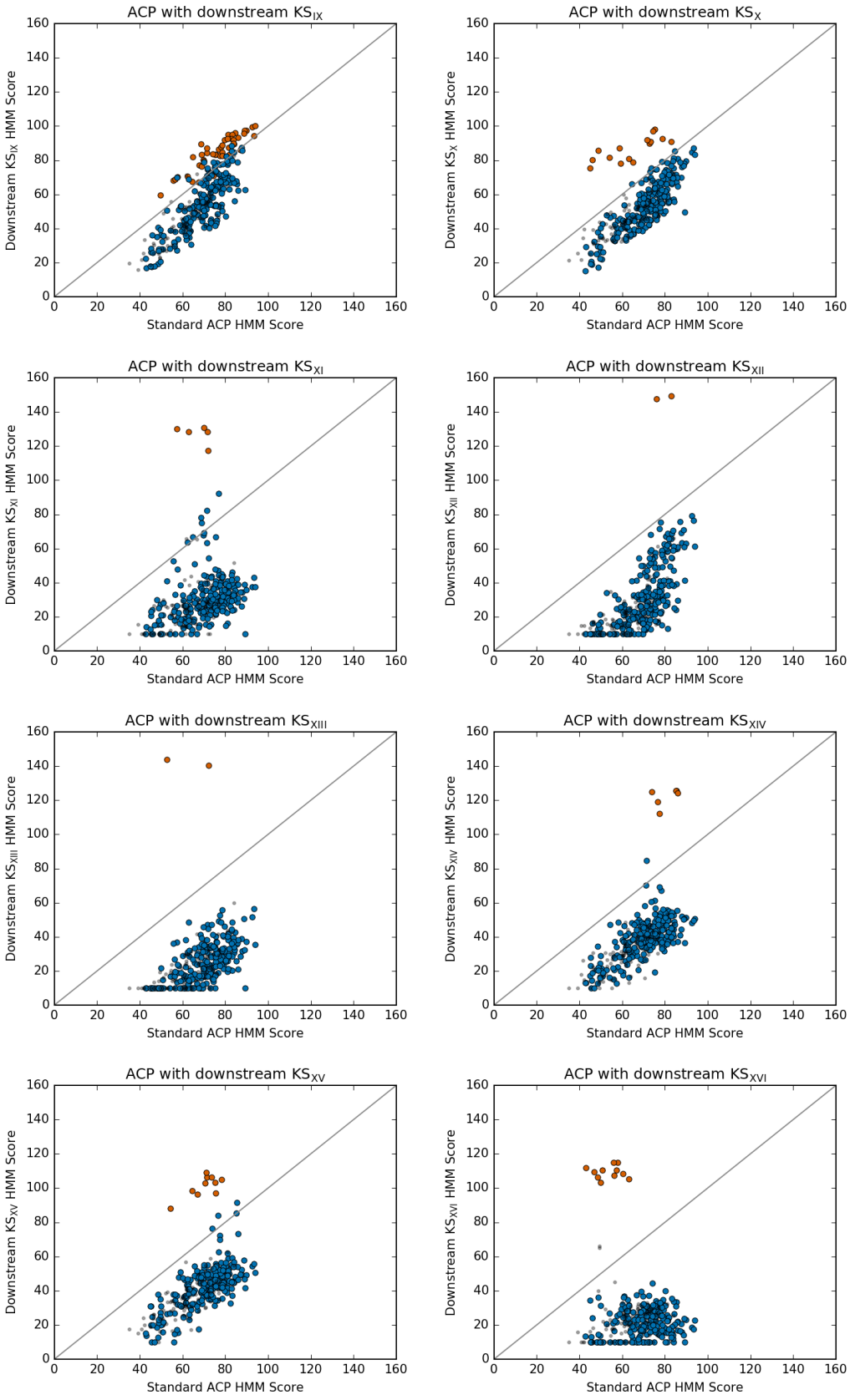


Figure A.1 | **(continued)** Plot of bit scores for ACPs scored against **downstream** KS clade grouped HMM profiles and a standard ACP HMM profile. ACPs from the group used to generate each profile are shown as orange dots. ACPs from other groups are shown as blue dots. ACPs with no assigned group are shown as grey dots ($n = 73$).

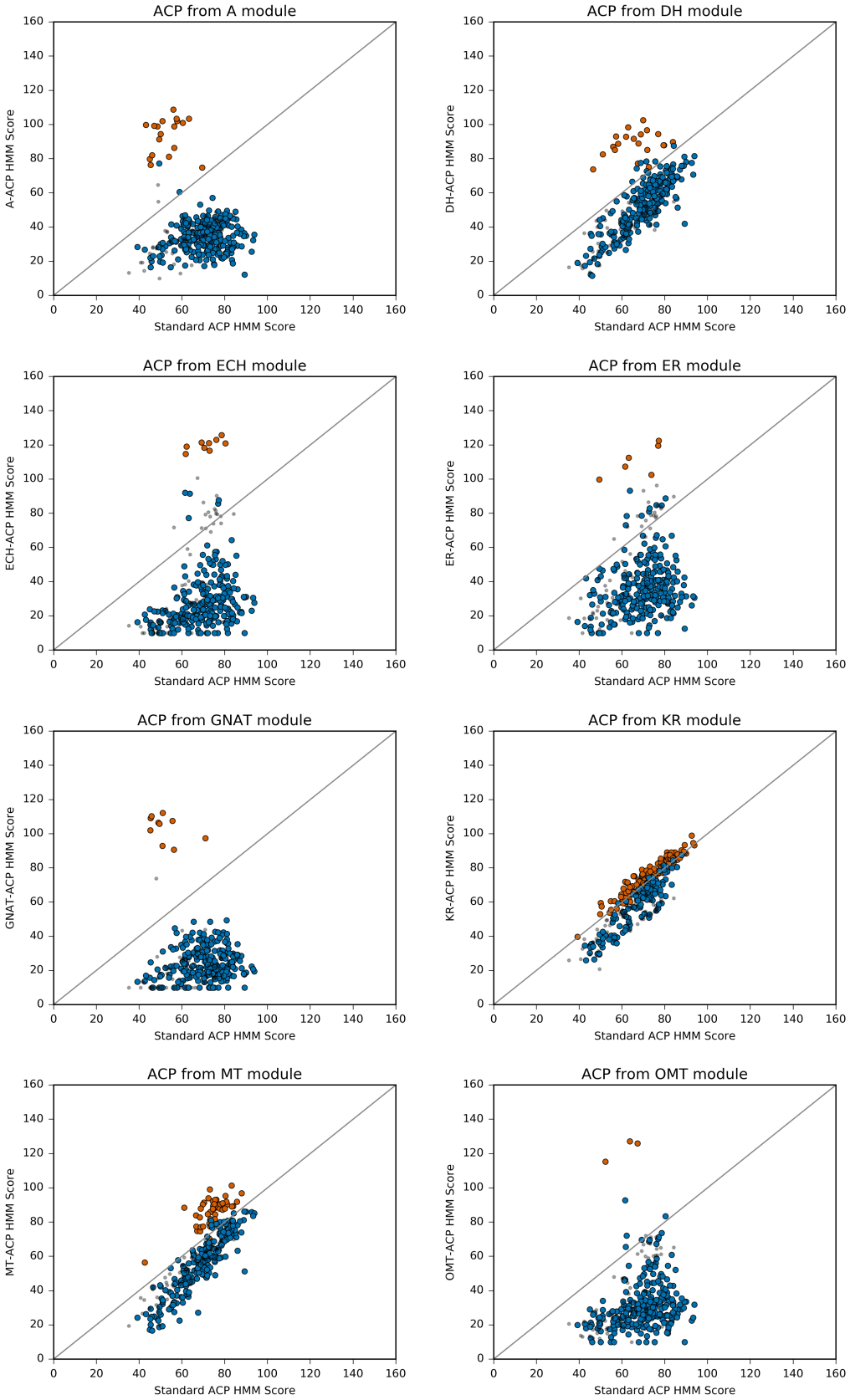


Figure A.2 | Plot of bit scores for ACPs scored against module-terminating-domain grouped HMM profiles and a standard ACP HMM profile. ACPs from the group used to generate each profile are shown as orange dots. ACPs from other groups are shown as blue dots. ACPs with no assigned group are shown as grey dots ($n = 64$).

Figure A.3 | Differential logos for the ACPs grouped by downstream KS clade against all ACPs. Generated using iceLogo[227] (Percentage change, p-value = 0.05). Positions are relative to the conserved Ser for Ppant attachment, indicated by a vertical red line. Using same alignment as in Figure 3.19.

A.3 Selected PepFoot Data

Selected Python code written for PepFoot are shown here. The complete source code is available at github.com/jbellamycarter/pepfoot. For descriptions of usage and rationale, see Chapter 5.

Listing A.2 | Python Code for `mz5Reader.py`. Contains only one class, `mz5`, which allows interaction with a .mz5 datafile. The latest version of this file is available at github.com/jbellamycarter/pepfoot/blob/master/mz5Reader.py

```
import os
import numpy as np
from scipy import signal
import h5py

#  mz5 Keys:
# 'CVParam',
# 'CVReference',
# 'ChomatogramTime',
# 'ChromatogramIndex',
# 'ChromatogramIntensity',
# 'ChromatogramList',
# 'ChromatogramListBinaryData',
# 'ControlledVocabulary',
# 'DataProcessing',
# 'FileContent',
# 'FileInformation',
# 'InstrumentConfiguration',
# 'ParamGroups',
# 'RefParam',
# 'Run',
# 'Software',
# 'SourceFiles',
# 'SpectrumIndex',
# 'SpectrumIntensity',
# 'SpectrumListBinaryData',
# 'SpectrumMZ',
# 'SpectrumMetaData'


class mz5():
    """A class object for accessing data from an mz5 file"""

    def __init__(self, filename=None, in_memory=True):
        if not os.path.splitext(filename)[1] == '.mz5':
            raise ValueError('Incorrect file extension, must be .mz5')
        else:
            self.filename = filename
        self.file = h5py.File(self.filename, 'r')
        self.in_memory = in_memory
        self.num_scan = self.file['ChromatogramIndex'][0]
        self.scan_lookup = np.zeros((self.num_scan),
                                    dtype=[('time', 'float32'),
                                           ('start', 'uint32'),
                                           ('end', 'uint32'),
                                           ('ms level', 'uint8'),
                                           ('precursor', 'float64'),
                                           ('min mz', 'uint32'),
                                           ('max mz', 'uint32')])
        self.time_ref = int(np.where(self.file['CVReference']['name'] ==
                                     b'scan start time')[0])
        self.msn_ref = int(np.where(self.file['CVReference']['name'] ==
                                    b'ms level')[0])
        self.min_mz_ref = int(np.where(self.file['CVReference']['name'] ==
                                       b'scan window lower limit')[0])
        self.max_mz_ref = int(np.where(self.file['CVReference']['name'] ==
                                       b'scan window upper limit')[0])

        try:
            self.precursor_ref = int(
                np.where(self.file['CVReference']['name'] == b'selected ion m/z')[0])
        except TypeError:
            print("No MSn spectra in this file.")

        self.fill_lookup()
        self.get_limits()
        if in_memory:
            self.mzs = self.get_all_mzs()

    def fill_lookup(self):
        """Extract ssential scan information from CVParam into
        self.scan_lookup"""

        self.scan_lookup['start'][1:] = self.file['SpectrumIndex'][:-1]
        self.scan_lookup['end'] = self.file['SpectrumIndex']
        meta = self.file["SpectrumMetaData"]["params"]["cvstart"]
        param_value = self.file['CVParam']['value'].astype('U')
        param_id = self.file['CVParam']['cvRefID']

        for scan in range(self.num_scan):
            start = meta[scan]
            if scan == self.num_scan-1:
                end = -1
```

```
            else:
                end = meta[scan+1]
            idx = np.where(param_id[start:end] == self.time_ref)[0]
            self.scan_lookup[scan]['time'] = param_value[start+idx][0]
            idx = np.where(param_id[start:end] == self.msn_ref)[0]
            self.scan_lookup[scan]['ms level'] = param_value[start+idx][0]
            if self.scan_lookup[scan]['ms level'] > 1:
                idx = np.where(param_id[start:end] == self.precursor_ref)[0]
                self.scan_lookup[scan]['precursor'] = param_value[start+idx][0]
            idx = np.where(param_id[start:end] == self.min_mz_ref)[0]
            self.scan_lookup[scan]['min mz'] = float(param_value[start+idx][0])
            idx = np.where(param_id[start:end] == self.max_mz_ref)[0]
            self.scan_lookup[scan]['max mz'] = float(param_value[start+idx][0])

    def get_limits(self):
        self.time_range = (self.scan_lookup['time'][0],
                           self.scan_lookup['time'][-1])
        scan_list = np.where(self.scan_lookup['ms level'] == 1)[0]
        self.ms1_range = (self.scan_lookup['min mz'][scan_list].min(),
                          self.scan_lookup['max mz'][scan_list].max())

    def get_mzs(self, scan):
        """Generate m/z array for a given scan"""

        if self.in_memory:
            return self.mzs[scan]
        else:
            return self.file['SpectrumMZ'][self.scan_lookup['start'][scan]:
                                           self.scan_lookup['end'][scan]].cumsum()

    def get_all_mzs(self):
        """Generate ragged array of all m/z arrays for data file.
        Only runs when in_memory == True
        """

        mzs = []
        for scan in range(self.num_scan):
            mzs.append(self.file['SpectrumMZ'][self.scan_lookup['start'][scan]:
                                           self.scan_lookup['end'][scan]].cumsum())

        return mzs

    def get_ints(self, scan, mz_idx=None):
        """Generate int array for a given scan and mz_idx if specified"""

        if mz_idx is not None:
            return self.file['SpectrumIntensity'][self.scan_lookup['start'][scan] + mz_idx[0]:
                                                  self.scan_lookup['start'][scan] + mz_idx[1]]
        else:
            return self.file['SpectrumIntensity'][self.scan_lookup['start'][scan]:
                                                  self.scan_lookup['end'][scan]]

    def get_scan_from_time(self, min_time, max_time):
        """Get scan numbers from a tuple of scan times"""

        return np.searchsorted(self.scan_lookup['time'], (min_time, max_time))

    def chromatogram(self, min_mz, max_mz, ms_level=1):
        """Generate 1xN array of ion current for m/z range by scan time.
           Returns an array of scan times and an array of intensities"""

        scan_list = np.where(self.scan_lookup['ms level'] == ms_level)[0]
        xic = np.zeros_like(scan_list, dtype='uint64')

        for i, scan in enumerate(scan_list):
            tmp_mz = self.get_mzs(scan)
            idx = np.where((tmp_mz >= min_mz) & (tmp_mz < max_mz))[0]
            if idx.any():
                xic[i] = self.get_ints(scan, (idx[0], idx[-1])).sum()

        return self.scan_lookup['time'][scan_list], xic

    def spectrum(self, min_time, max_time, ms_level=1, mz_range=None, precursor=None, tolerance=0.1):
        """Generate 1xN array of total ion current by m/z.
        ms_level allows selection for MS1 or MS2 """

        min_scan, max_scan = self.get_scan_from_time(min_time, max_time)

        if ms_level == 1:
            scan_list = np.where(self.scan_lookup['ms level'][min_scan:max_scan] == 1)[
                0] + min_scan
        elif ms_level == 2:
            if precursor is not None:
                scan_list = np.where((self.scan_lookup['ms level'][min_scan:max_scan] == 2) &
                                     (precursor - tolerance <=
                                      self.scan_lookup['precursor'][min_scan:max_scan]) &
                                     (precursor + tolerance >
                                      self.scan_lookup['precursor'][min_scan:max_scan]))[0] + min_scan
            else:
                scan_list = np.where(self.scan_lookup['ms level'][min_scan:max_scan] == 2)[0] + min_scan

        merge_mz = self.get_mzs(scan_list[0])
        merge_int = self.get_ints(scan_list[0])

        if mz_range:
            idx = np.where((merge_mz >= mz_range[0]) & (merge_mz < mz_range[1]))[0]
            merge_mz = merge_mz[idx]
            merge_int = merge_int[idx]

```

```
        for scan in scan_list[1:]:
            tmp_mz = self.get_mzs(scan)
            tmp_int = self.get_ints(scan)
            if mz_range:
                idx = np.where((tmp_mz >= mz_range[0]) & (tmp_mz < mz_range[1]))[0]
                tmp_mz = tmp_mz[idx]
                tmp_int = tmp_int[idx]

            mz1_mask = np.in1d(merge_mz, tmp_mz)
            mz2_mask = np.in1d(tmp_mz, merge_mz, invert=True)
            # Generate unique tmp_mz and tmp_int if multiple mz are the same
            if tmp_mz[~mz2_mask].size != merge_int[mz1_mask].size:
                tmp_mz, unique_idx = np.unique(tmp_mz, return_index=True)
                tmp_int = tmp_int[unique_idx]
                mz2_mask = np.in1d(tmp_mz, merge_mz, invert=True)
            ins = np.searchsorted(merge_mz, tmp_mz[mz2_mask])
            merge_int[mz1_mask] += tmp_int[~mz2_mask]
            merge_int = np.insert(merge_int, ins, tmp_int[mz2_mask])
            merge_mz = np.insert(merge_mz, ins, tmp_mz[mz2_mask])

        return merge_mz, merge_int

    def get_area(self, rt_range, mz_range, ms_level=1):
        """Calculates the sum of intensities for given m/z and time ranges"""

        min_scan, max_scan = self.get_scan_from_time(*rt_range)
        min_mz, max_mz = mz_range
        scan_list = np.where(
            self.scan_lookup['ms level'][min_scan:max_scan] == ms_level)[0] + min_scan

        area = np.zeros(len(scan_list))
        for i, scan in enumerate(scan_list):
            tmp_mz = self.get_mzs(scan)
            idx = np.where((tmp_mz >= min_mz) & (tmp_mz < max_mz))[0]
            if idx.any():
                area[i] = self.get_ints(scan, (idx[0], idx[-1])).sum()

        return int(np.trapz(area, x=self.scan_lookup['time'][scan_list]*60))

    def detect_peaks(self, spectrum, order=3, threshold=500, cwt_width=0.1, cwt_minwidth=0.01):
        """Peak picking from a given spectrum using the relative maxima
        algorithm using a window of size order.

        Only peaks above the specified threshold are returned"""

        peaks = signal.argrelmax(spectrum[1], order=order)[0]

        return peaks[spectrum[1][peaks] > threshold]
```

Listing A.3 | Example .pfoot file for OmpF analysis.

```
{
 "areas": [
  [
   [3170837, 5843793, 0, 10307944, 409497439, 774518956, 11375957, 0, 898592311, 579488895, 1472254,
        2508948945, 659091969, 435509968, 16384989, 14991875, 3073868, 2439090, 2258413651, 0, 3684021,
        16533456, 1424585109, 1168401670, 172449914, 1183823459, 961269458],
   [0, 10692484, 1794512, 0, 0, 0, 14246946, 39837377, 123373197, 0, 38619390, 153386645, 101685436,
        62924177, 0, 68202444, 29185373, 38957259, 226751858, 21912786, 0, 0, 423656328, 66322838, 0,
        15771390, 0]
  ],
  [
   [1760793, 2795658, 0, 4213123, 92735265, 903033222, 5115125, 0, 498901265, 1514849787, 1456733,
        3096908980, 462968684, 530484297, 8143751, 8834054, 2418302, 235290, 1066112788, 0, 192657,
        11395333, 2729866895, 1147103771, 535067968, 1968583134, 2242147882],
   [0, 5805312, 516032, 0, 0, 0, 12808167, 28215354, 79109620, 0, 20017224, 170591147, 115204046,
        115475278, 0, 42437109, 23117811, 18302164, 137036113, 13317098, 0, 0, 716660039, 94269073, 0,
        51522347, 0]
  ],
  [
   [2978335, 4394504, 0, 7864862, 343236443, 1141578119, 6296906, 0, 682788344, 1032698980, 1388608,
        2630902827, 833960643, 576596683, 13252931, 12526941, 2629764, 837423, 1779936961, 0, 1047066,
        15385700, 3055394177, 1524229041, 337348449, 1560798454, 1939614278],
   [0, 7000604, 1059264, 0, 0, 0, 13886940, 44347066, 107715332, 0, 29603955, 165591068, 159754200,
        95951892, 0, 54503090, 25591995, 42205943, 207733865, 16267519, 0, 0, 672960443, 123002398, 0,
        34031732, 0]
  ]
 ],
 "charge array": [2, 2, 4, 2, 4, 3, 1, 3, 2, 3, 2, 2, 3, 3, 2, 2, 2, 3, 2, 3, 3, 2, 3, 3, 3, 2, 2],
 "charge range": [1, 4],
 "creation date": "08 Feb 2018  05:46pm",
 "data files": ["LM_AD_OmpF_1_040816.mz5", "LM_AD_OmpF_2_040816.mz5", "LM_AD_OmpF_3_040816.mz5"],
 "differential mod": "Aryldiazarine-TDBA",
 "enzyme": "Trypsin",
 "fixed mods": [],
 "fractional mod": [
  [0, 0.646607697730269, 1, 0, 0, 0, 0.5560238822275524, 1, 0.12072148818548972, 0, 0.9632777842684626,
       0.05761356516291021, 0.13365990542266434, 0.12624371269749185, 0, 0.8197968902179487,
       0.904713567191491, 0.9410795865113611, 0.09124215557427487, 1, 0, 0, 0.22922131249673958,
       0.05371468499271094, 0, 0.013147263855915407, 0],
  [0, 0.6749601498435641, 1, 0, 0, 0, 0.7146101843344403, 1, 0.13686527719975378, 0, 0.9321628053925972,
       0.05220845917965591, 0.19925541282446857, 0.17876548698887232, 0, 0.827699363870486,
       0.9052987429997666, 0.9873073184699473, 0.11389788320140767, 1, 0, 0, 0.2079368746346202,
       0.07593937104040598, 0, 0.025504780559525642, 0],
  [0, 0.6143517025025125, 1, 0, 0, 0, 0.6880224908572925, 1, 0.1362616459230735, 0, 0.9551954447910617,
       0.05921381351701467, 0.16076463094553978, 0.1426690882513579, 0, 0.8131144978882674,
       0.9068178563922965, 0.9805446674407388, 0.10451120089036312, 1, 0, 0, 0.18049797071073673,
       0.07467220154240876, 0, 0.02133878095532862, 0]
```

```
 ],
 "length range": [5, 40],
 "m/z array": [
  [
   [554.765295498925, 554.8052954989249], [511.2569093674501, 511.29690936745004], [], [881.837914157195,
        881.8779141571949], [1035.9530821734334, 1035.9930821734333], [1232.1844639451888,
        1232.2244639451887], [719.4250373217102, 719.4650373217102], [], [625.2637005689222,
        625.29150428777785], [1186.1777839939268, 1186.2177839939268], [411.70705234681515,
        411.7470523468151], [1102.5441975016477, 1102.5841975016476], [975.7673073675394,
        975.8073073675394], [923.7336150731977, 923.7736150731977], [685.3371475547101, 685.3771475547101],
         [566.2914804619401, 566.33148046194], [502.24399895494, 502.28399895494], [945.1053807658315,
        945.1453807658314], [924.41157208554, 924.45157208554], [], [642.9763109065277, 643.0163109065277],
         [543.744363092935, 543.784363092935], [925.4711295066111, 925.5111295066111], [976.457337169346,
        976.497337169346], [910.0799831904194, 910.1199831904194], [1067.4753006851836,
        1067.5153006851835], [870.42615952052, 870.4661595205199]
  ],
  [
   [], [612.26900936745, 612.30900936745], [742.8015156557359, 742.8415156557359], [], [], [],
        [921.4492373217101, 921.4892373217101], [718.3190773016512, 718.3590773016512], [726.273399808019,
        726.3128627030812], [], [512.7191523468151, 512.7591523468151], [1203.5562975016476,
        1203.5962975016475], [1043.1087073675394, 1043.1487073675394], [991.0750150731976,
        991.1150150731976], [], [667.3035804619401, 667.3435804619401], [603.2560989549401,
        603.29609895494], [1012.4467807658314, 1012.4867807658313], [1025.42367208554, 1025.4636720855399],
         [964.0958525144355, 964.1358525144354], [], [], [992.812529506611, 992.852529506611],
        [1043.798737169346, 1043.838737169346], [], [1168.4874006851835, 1168.5274006851835], []
  ],
 ],
 "missed cleave": 1,
 "name": "OmpF (demo).pfoot",
 "pdb file": "OmpF Trimer (demo).pdb",
 "peptides": [
  [8, 17], [18, 26], [18, 43], [27, 43], [44, 81], [48, 81], [84, 90], [84, 101], [91, 101], [102, 133],
       [134, 141], [142, 161], [169, 197], [170, 197], [198, 210], [211, 220], [212, 220], [212, 236],
       [221, 236], [221, 244], [237, 254], [245, 254], [255, 278], [281, 306], [283, 306], [307, 324],
       [325, 341]
 ],
 "rt array": [
  [
   [16.362394023481464, 23.823238408795063], [22.845978786207986, 24.09727347014446], [],
        [18.302820603528918, 24.054484853936536], [29.860025419321147, 45.49849638623358],
        [28.599812909791734, 45.526232185897431], [25.281883862277347, 26.228035317120808], [],
        [23.064917753454804, 30.293931519865744], [37.45014689152233, 38.640452284107596],
        [13.930264603914111, 21.79836064551885], [39.234208692163016, 40.94667326987187],
        [30.512024320320386, 36.97934865716541], [31.001493624299442, 44.59456548502013],
        [18.628471289664468, 23.88821801641314], [20.599188850695754, 23.897233531717333],
        [22.46878582346003, 23.8418306745051], [30.382110900357986, 31.051376759316653],
        [25.827885447183398, 30.145738666409331], [], [23.397228057258133, 23.997810601431546],
        [16.875643626741486, 24.025453760177953], [34.490337134177544, 38.701571082859594],
        [36.6946415749693, 38.103331638169855], [38.338355033443214, 42.8953890916743], [36.76304247300345,
         37.708475220413085], [33.41136565172683, 49.51950591415649]
  ],
  [
   [], [28.998232454195744, 32.375969740409715], [27.242793661307523, 28.230881085858012], [], [], [],
        [32.57499459217631, 33.93460386268896], [33.994853991076326, 35.59225061728683],
        [28.28526850737825, 35.037345029864284], [], [31.39463246285967, 31.93363020319046],
        [37.4402975604521, 43.99272768824636], [31.992395357810416, 41.196362921876485],
        [31.960370786759334, 45.18727729073817], [], [25.025505536450396, 32.67488380976391],
        [29.26652534401559, 33.645873126125295], [31.36134364120239, 36.04609420718228],
        [28.719705865717092, 35.38842920482914], [27.553102432668364, 30.959545921073337], [], [],
        [34.97319870589767, 43.67547056238419], [37.76499744576963, 43.94675151703888], [],
        [30.470554582800318, 40.69720739985104], []
  ]
 ],
 "sequence": "GAEIYNKDGNKVDLYGKAVGLHYFSKGNGENSYGGNGDMTYARLGFKGETQINSDLTGYGQWEYNFQGNNSEGADAQTGNKTRLAFAGLKY
              ADVGSFDYGRNYGVVYDALGYTDMLPEFGGDTAYSDDFFVGRVGGVATYRNSNFFGLVDGLNFAVQYLGKNERDTARRSNGDGVGGSISYE
              YEGFGIVGAYGAADRTNLQEAQPLGNGKKAEQWATGLKYDANNIYLAANYGETRNATPITNKFTNTSGFANKTQDVLLVAQYQFDFGLRPS
              IAYTKSKAKDVEGIGDVDLVNYFEVGATYYFNKNMSTYVDYIINQIDSDNKLGVGSDDTVAVGIVYQF",
 "treatment": []
}
```


Listing A.4 | Python Code for isotope pattern prediction. `get_isotopes` function. Relies on a custom dictionary of relative abundances for stable isotopes. The latest version of this is available at github.com/jbellamycarter/pepfoot/blob/master/utility.py

```python
def get_isotopes(composition, threshold=0.001, tolerance=0.01, mass_dict=_rel_mass):
    """Calculate isotope distribution from composition object using
    the binomial method.

    input
    -----
    composition: Composition object
    threshold: minimum abundance for isotope
    tolerance: tolerance for combining adjacent peaks in daltons
    mass_dict: dictionary containing relative masses

    returns
    -------
    masses: list of isotope masses
    abun: list of isotope relative abundances
    """

    masses = []
    abun = []

    for element in composition:
        ele_name, ele_num = re.match(r'^([A-Z][a-z+]*)(?:\[(\d+)\])?$', element).groups()
        count = composition[element]
```

```
        if ele_num:
            ele_num = int(ele_num)
            if not ele_num in mass_dict[ele_name]:
                raise AssertionError('Can not produce isotope pattern, specified isotope {} not abundant'.
                    format(element))
                continue
            isotopes = {0: mass_dict[ele_name][ele_num]}
        else:
            isotopes = mass_dict[element]

        for i in range(count):

            if len(masses) == 0:
                masses = [isotopes[iso][0] for iso in isotopes]
                abun = [isotopes[iso][1] for iso in isotopes]
                continue

            _masses = []
            _abun = []

            for peak in range(len(masses)):

                if abun[peak] < threshold:
                    continue

                for iso in isotopes:
                    _masses.append(masses[peak] + isotopes[iso][0])
                    _abun.append(abun[peak] * isotopes[iso][1])

            _masses, _abun = (list(buf) for buf in zip(*sorted(zip(_masses, _abun))))
            masses = [_masses[0]]
            abun = [_abun[0]]

            for peak in range(1, len(_masses)):
                if _masses[peak] <= (masses[-1] + tolerance):
                    _abundance = abun[-1] + _abun[peak]
                    masses[-1] = (masses[-1]*abun[-1] + _masses[peak]*_abun[peak]) / _abundance
                    abun[-1] = _abundance
                else:
                    masses.append(_masses[peak])
                    abun.append(_abun[peak])
            _max = max(abun)
            abun = [i/_max for i in abun]

    masses, abun = (list(buf) for buf in zip(*[(masses[i], ab) for i, ab in enumerate(abun) if ab >=
        threshold]))

    return np.array(masses), np.array(abun)
```

Listing A.5 | Python Code for PDB class. Allows parsing and writing of .pdb files. The latest version of this is available at github.com/jbellamycarter/pepfoot/blob/master/utility.py

```
class PDB():
    """
    Class for parsing PDB files that follow the accepted format
    https://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM

    Extracts ATOM and HETATM records.
    HETATM with HOH resname are removed.

    Dictionary format:

    Structure (list)>
        Model (odict)>
            Chain (odict)>
                Residue (list)>
                    Atom (dict)
    """

    def __init__(self, filename=None):
        if not os.path.splitext(filename)[1] == '.pdb':
            raise ValueError('Incorrect file extension, must be .pdb')
        else:
            self.filename = filename

        self.AA_3to1    = {'ALA':'A','ARG':'R','ASN':'N','ASP':'D','CYS':'C','GLU':'E',
                           'GLN':'Q','GLY':'G','HIS':'H','ILE':'I','LEU':'L','LYS':'K',
                           'MET':'M','PHE':'F','PRO':'P','SER':'S','THR':'T','TRP':'W',
                           'TYR':'Y','VAL':'V'}

        self.ATOM_STRING = "{}{:5d} {:4}{:.1}{:.3} {:.1}{:>4d}{:.1}   {:8.3f}{:8.3f}{:8.3f}{:6.2f}{:6.2f}
                    {:.2}  \n"

        self.structure = [None]
        self.num_models = 0
        self.num_chains = 0
        self.chains = {}
        self._parse()

# FIXME: Add ability to handle PDBs with gaps
    def _parse(self):
        """Parse the PDB file as a series of record entries into a dictionary"""

        with open(self.filename, 'r') as entries:
            model = 0
            open_model = False
            chain = None
```

```
            resnum = None
            for entry in entries:
                record_type = entry[0:6]

                if record_type == 'ATOM  ' or record_type == 'HETATM' and not entry[17:20] == 'HOH':
                    if not open_model:
                        model += 1
                        open_model = True
                        self.structure.append(OrderedDict())
                    if not entry[21] == chain:
                        chain = entry[21]
                        if chain == ' ':
                            chain = 'A'
                        if not chain in self.structure[model]:
                            self.structure[model][chain] = OrderedDict()
                    if not int(entry[22:26]) == resnum:
                        resnum = int(entry[22:26])
                        self.structure[model][chain][resnum] = []
                    self.structure[model][chain][resnum].append({'type': record_type,
                                                                 'serial': int(entry[6:11]),
                                                                 'name': entry[12:16],
                                                                 'altLoc': entry[16],
                                                                 'resname': entry[17:20],
                                                                 'icode': entry[26],
                                                                 'x': float(entry[30:38]),
                                                                 'y': float(entry[38:46]),
                                                                 'z': float(entry[46:54]),
                                                                 'occupancy': float(entry[54:60]),
                                                                 'bfactor': -1.00,
                                                                 'element': entry[76:78]
                                                                 })
                elif record_type == 'MODEL':
                    open_model = True
                    model = int(entry[10:14])
                elif record_type == 'ENDMDL':
                    open_model = False
                    chain = None
                    resnum = None

            self.num_models = model
            self.chains = {chain:self._get_sequence(self.structure[1][chain]) for chain in self.structure
                 [1]}
            self.num_chains = len(self.chains)

    def _get_sequence(self, chain):
        """Parse single character amino acid code from chain"""
        _sequence = ''

        for residue in chain:
            resname = chain[residue][0]['resname']
            if chain[residue][0]['type'] == 'ATOM  ':
                _sequence += self.AA_3to1[resname]

        return _sequence


# FIXME: Add ability to check for alignment with sequence and colour when gaps present
    def bfactor_by_residue(self, sequence, bfactors):
        """
        Modify b-factor of atoms in self.structure by residue.
        Takes input sequence to compare against self.chains and a
        list of corresponding b-factors
        """
        matches = []
        for chain in self.chains:
            if self.chains[chain] == sequence:
                matches.append(chain)

        if matches == []:
            raise ValueError('No chains match input sequence, please check that sequence matches PDB.')
            return

        for match in matches:
            for model in self.structure[1:]:
                for residue in model[match].items():
                    if residue[1][0]['type'] == 'ATOM  ': #Only count ATOM entries
                        bfactor = bfactors[residue[0]-1] #Account for one-indexing
                        for atom in residue[1]:
                            atom['bfactor'] = bfactor

    def _to_string(self):
        """Convert self.structure into string with PDB formatting"""

        _PDB_string = ''
        # Add REMARK for bookkeeping purposes
        _PDB_string += 'REMARK     Output from pepFoot.\n'
        _PDB_string += 'REMARK     {}\n'.format(os.path.split(self.filename)[1])
        _PDB_string += 'REMARK     {}\n'.format(datetime.today().strftime('%d %b %Y  %I:%M%p'))
        for i, model in enumerate(self.structure[1:]):
            _PDB_string += 'MODEL     {}\n'.format(i+1)
            for chain in model.items():
                for residue in chain[1].items():
                    for atom in residue[1]:
                        _line = self.ATOM_STRING.format(atom['type'], atom['serial'], atom['name'],
                                                        atom['altLoc'], atom['resname'],chain[0],
                                                        residue[0], atom['icode'], atom['x'],
                                                        atom['y'], atom['z'], atom['occupancy'],
                                                        atom['bfactor'], atom['element'])
                        _PDB_string += _line
```

```
143
144                 _PDB_string += 'ENDMDL\n'
145             _PDB_string += 'END\n'
146
147             return _PDB_string
148
149         def write(self, filename):
150             """Write self.structure to file"""
151
152             with open(filename, 'w') as out_file:
153                 lines = self._to_string()
154                 out_file.write(lines)
```

Figure A.4 | Screenshot of Analysis tab with *Difference Plot* enabled. Compare with Figure 5.5. The bar plot displays E_m instead of f_m (Equation 5.2).

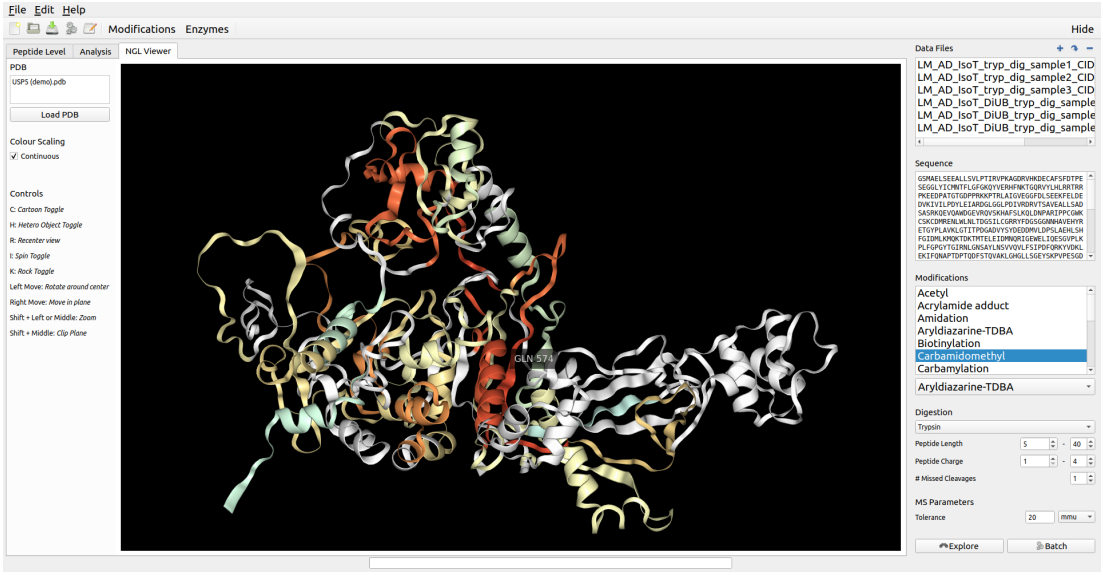


Figure A.5 | Screenshot of NGL Viewer tab with *Continuous* enabled. Compare with Figure 5.6. The PDB b-factor is written as $-E_m$ instead of f_m (Equation 5.2).