Structural basis of ribosomal peptide macrocyclization in plants

Joel Haywood[1,2], Jason W. Schmidberger[1,2], Amy M. James[1,2], Samuel G. Nonis[1,2], Kirill V. Sukhoverkov[1,2], Mikael Elias[3], Charles S. Bond[1], Joshua S. Mylne[1,2]

The University of Western Australia, [1] School of Molecular Sciences & [2] The ARC Centre of Excellence in Plant Energy Biology, 35 Stirling Highway, Crawley, Perth 6009, Australia; [3] University of Minnesota, Department of Biochemistry, Molecular Biology and Biophysics & Biotechnology Institute, St. Paul, MN 55108, USA

Constrained, cyclic peptides encoded by plant genes represent a new generation of drug leads. Evolution has repeatedly recruited the Cys-protease asparaginyl endopeptidase (AEP) to perform their head-to-tail ligation. These macrocyclization reactions use the substrates amino terminus instead of water to deacylate, so a peptide bond is formed. How solvent-exposed plant AEPs macrocyclize is poorly understood. Here we present the crystal structure of an active plant AEP from the common sunflower, *Helianthus annuus*. The active site contained electron density for a tetrahedral intermediate with partial occupancy that predicted a binding mode for peptide macrocyclization. By substituting catalytic residues we could alter the ratio of cyclic to acyclic products. Moreover, we showed AEPs from other species lacking cyclic peptides can perform macrocyclization under favorable pH conditions. This structural characterization of AEP presents a logical framework for engineering superior enzymes that generate macrocyclic peptide drug leads.

Introduction

Asparaginyl endopeptidases (AEPs) are a group of asparagine/aspartic acid (Asx) specific proteases that have been classified as belonging to the C13 family of cysteine proteases based on the presence of a His-Gly-spacer-Ala-Cys motif (Hara-Nishimura et al., 1993; Chen et al., 1997; Mathieu et al., 2002; Shafee et al., 2015). First described in plants as vacuolar processing enzymes based on their propensity for processing seed proteins stored in vacuoles, AEPs have since been described in a variety of organisms and shown to be involved in a wide range of processes including, cell death, antigen processing and hemoglobin degradation (Hara-Nishimura et al., 1993; Manoury et al., 1998; Hatsugai et al., 2004; Kuroyanagi et al., 2005; Yamada et al., 2005; Sojka et al., 2007). In addition to the proteolytic function observed in these processes, AEP has become well known for its curious ligation reactions (Min and Jones, 1994; Sheldon et al., 1996; Mylne et al., 2012; Nguyen et al., 2014; Zhao et al., 2014; Dall et al., 2015).

The ability of endoproteases to perform ligation reactions was first observed by Bergmann and Fruton in 1938 with chymotrypsin (Bergmann and Fruton, 1938). Later, *in vitro* ligation reactions were performed with AEP from jack bean (*Canavalia ensiformis*) seeds (Bowles et al., 1986; Min and Jones, 1994). The recent discovery that evolutionarily distinct plant families have repeatedly recruited AEPs to catalyze the formation of ribosomally synthesized and post-translationally modified peptides (RiPPs), through the macrocyclization of linear precursor sequences, has caught the attention of drug designers keen to overcome the current inefficiencies in native chemical ligation that limit the therapeutic use of cyclic peptides (Pattabiraman and Bode, 2011; Mylne et al., 2012; Arnison et al., 2013). Such therapeutic cyclic

peptides are viewed by many to have the potential to capitalize on a niche in the current pharmaceutical market by virtue of their intermediate size between small molecule drugs and large protein structures, and their unique capacity to combine favorable bioavailability and stability characteristics with high target specificity facilitated by tolerance to site-directed mutagenesis (Clark et al., 2005; Clark et al., 2010; Gould et al., 2011; Ji et al., 2013; Poth et al., 2013; Truman, 2016). Moreover, as computational techniques for the discovery of RiPPs improve and the number of cyclic peptides described continues to expand, an ever-wider array of scaffolds might be exploited to tailor molecules to specific drug targets (Bhardwaj et al., 2016; Truman, 2016; Hetrick and van der Donk, 2017) (**Figure 1).**

Sunflower trypsin inhibitor-1 (SFTI-1) is a 14-residue, bicyclic peptide with a cyclic backbone and an internal disulfide bond (Luckett et al., 1999). Its biosynthesis is rather unusual as it sequence is buried within a precursor that also encodes seed storage albumin. Seed storage albumins are a major class of seed storage protein that constitute over 50% of total seed protein and become a source of nitrogen and sulfur during seed germination (Youle and Huang, 1978; Shewry and Halford, 2002). The common sunflower (*Helianthus annuus*) has many genes encoding precursors for these napin-type or 2S seed storage albumins that are synthesized in the rough endoplasmic reticulum before undergoing cleavage maturation by AEP and localizing to storage vacuoles (Bollini and Chrispeels, 1979; Franke et al., 2016; Jayasena et al., 2016). Along with an adjacent albumin, SFTI-1 is post-translationally processed by AEP from within a unique seed storage albumin precursor called Preproalbumin with SFTI-1 (PawS1) (Mylne et al., 2011). SFTI-1 is a potent inhibitor of serine proteases, and its intrinsic stability and cellular penetration

capabilities have led to its application as a bioactive scaffold (White and Craik, 2016; Swedberg et al., 2017).

AEP-catalyzed macrocyclization of SFTI-1 is hypothesized to proceed via a cleavage coupled intramolecular transpeptidation reaction whereby catalysis begins with the deprotonation of the active site Cys by a localized His of the catalytic center, facilitating the nucleophilic attack on the carbonyl carbon of the Asp by the activated Cys thiol of AEP. This attack culminates in the formation of a thioacyl intermediate between the substrate and AEP, and the removal of the C-terminus. The reaction is proposed to be subsequently concluded by the nucleophilic attack of this intermediate by a Gly at the N-terminus of the substrate, resulting in the macrocyclization of SFTI-1 in a head-to-tail manner (Mylne et al., 2011; Bernath-Levin et al., 2015). Notably, this reaction proceeds in competition with nucleophilic attack upon the thioacyl intermediate by any nearby water molecules which would produce hydrolyzed, acyclic-SFTI. *In vitro* studies revealed the ratio of acyclic to cyclic SFTI-1 to be in the order of 5.8:1, with the less stable acyclic products hypothesized to be quickly degraded *in vivo* (Bernath-Levin et al., 2015). Evidence for AEP-mediated and hydrolysis-independent transpeptidation was demonstrated through the exclusion of a heavy atom O^{18} in the cyclic SFTI-1 product from an *in vitro* jack bean AEP (CeAEP1) catalyzed reaction (Bernath-Levin et al., 2015). The formation of a much larger cyclic peptide by an AEP from *Oldenlandia affinis* (OaAEP1) was also shown to lack O^{18} incorporation, suggesting a conserved mechanism of macrocyclization despite differences in substrate sequences (Harris et al., 2015). However, the suggestion that an AEP from butterfly pea (*Clitoria ternatea*) termed butelase 1 functions only as a ligase (Nguyen et al., 2014), combined with the proposal of a succinimide-driven, cleavage-independent ligation event

based on the crystal structure of human AEP (hAEP) and its ability to ligate substrate in the absence of the catalytic Cys, has cast uncertainty on the mechanism of AEP-catalyzed macrocyclization (Dall et al., 2015).

In order to elucidate a structural explanation why plant AEPs have been recruited by distinct plant lineages to perform macrocyclization and to understand the catalytic and structural nuances that might allow preferences towards cleavage or ligation reactions, we sought the crystal structure of a sunflower AEP.

Herein, we describe the first structure of an active plant AEP; one capable of performing peptide macrocyclization. This AEP, the most abundant AEP of five AEPs in the common sunflower (HaAEP1), displays structural similarity to previously published active AEPs from mammals, with subtle differences at residues involved in substrate recognition. Our characterization by site-directed mutagenesis of HaAEP1 residues integral to macrocyclization will facilitate the bioengineering of plant AEPs for improved macrocyclization efficiency, diversifying the scaffolds usable as cyclic therapeutic leads.

Results

Catalytic domain of HaAEP1 purified at pH 4

AEPs are synthesized as inactive precursors that have been shown to undergo irreversible auto-activation into their mature form on exposure to a low pH environment that resembles the acidic pH in the vesicles/vacuole where these proteins are active *in vivo* (Dall and Brandstetter, 2013; Shafee et al., 2015). In order to obtain an active form of a plant AEP, a ~51 kDa pro-

HaAEP1 (residues 28-491) lacking an endoplasmic reticulum signal sequence and including a N-terminal His-tag was expressed in *Escherichia coli* and purified by nickel affinity chromatography before being activated at pH 4.0 overnight. The activated form of HaAEP1 was then further purified by size exclusion chromatography, enabling separation of the core domain from the 'cap' domain (Zhao et al., 2014), and crystal trials were undertaken (**Figure 2—figure supplement 1**). SDS-PAGE analysis of the ~38 kDa core domain peak revealed the disassociation of the core domain from the cap domain but also showed several bands of HaAEP1, suggesting the presence of several cleavage sites at the termini of the core domain, as seen previously for several AEPs (Hara-Nishimura et al., 1998; Nguyen et al., 2014; Zhao et al., 2014; Harris et al., 2015). Indeed sequence comparison reveals the conservation of several of these predicted cleavage sites but notably lacks a previously described C-terminal di-Asp motif (Hiraiwa et al., 1999) (**Figure 2—figure supplement 2**). This led us to hypothesize that Asp52, Asn57, Asn338, Asp356, and Asp358 might represent the dominant autocatalytic cleavage sites in HaAEP1.

HaAEP1 structure reveals subtle differences to dictate substrate specificity

Crystallization trials of the ~38 kDa activated HaAEP1 yielded diffraction quality crystals that diffracted to a resolution of 1.8 Å (**Table 1**). The crystal structure was solved by molecular replacement yielding a single monomer in the asymmetric unit, and revealed an active monomeric HaAEP1 (residues 58-338 with weak electron density for Asn338) that forms a canonical C13 caspase structure, with a central six-stranded β-sheet region confined by five α-helices (Hara-Nishimura et al., 1993; Yamada et al., 2005) (**Figure 2A, Figure 2—figure supplement 3**). The structure of HaAEP1 lacks the C-terminal cap domain and N-terminal His-Tag, displaying dimensions of approximately 44 Å x 42 Å x 39 Å. Sequence analysis of this core

domain suggests that the aforementioned pro-domains are likely to have been auto-catalytically processed during maturation as the previously predicted Asn cleavage sites precede and follow the defined active structure.

HaAEP1 displays structural similarity to hAEP and a recently published inactive structure of OaAEP1 (PDB ID: 4N6O and 5H0I) with an r.m.s.d. value of 1.0 and 0.7 Å over 262 and 267 carbon alpha residues, respectively (Holm and Rosenström, 2010; Yang et al., 2017) (**Figure 2B**). Due to such topological conservation it is expected that subtle differences around the substrate binding pocket will define substrate specificity and catalytic efficiency. Indeed, comparison of these three structures reveals HaAEP1 exhibits a unique flexible extension, reflected by weak electron density, in the α5-β6 loop and differences between the residues that are local to the catalytic His and Cys and those that delineate the S3-S5 pockets (following the protease nomenclature defined by Schechter and Berger where the cleavage site residue is termed P1 and residues prior to and following the cleavage site are labelled P5-P2 and P1'-P2', respectively, and where the corresponding binding sites on the protease are described as S5-S2') (Schechter and Berger, 1967). Specifically, differences in the substrate pocket include residues YGT 249-251 in HaAEP1 (hAEP: YAC 217-219, OaAEP1: WCY 246-248), a bulky Trp232 in hAEP versus Leu271 in HaAEP1 and Leu268 in OaAEP1, and the presence of an additional proline prior to the β_{IV}-β_{V} polyproline loop which orients E257 away from the S4 region in HaAEP1 (**Figure 2B**). In OaAEP1, the C247 residue at the entrance to the S4 pocket was recently proposed to function as a 'gate keeper for ligation' with large bulky side chains inhibiting ligation (Yang et al., 2017). Differences local to the active site include residues P181, Q245, N247 in the β_{I} sheet and β5-β_{IV} loop (OaAEP1: A178, T242, S244, hAEP: T151, R213, S215) and

148 G185, E189, H191 in the β_{II}-β_{III} region (OaAEP1: G182, K186, Y188, hAEP: V155, N158, D160)
149 (**Figure 2B**).

150 A 6-residue insertion in the α5-β6 loop results in the disruption of a potential N-linked
151 glycosylation site that was hypothesized to affect substrate binding upstream to P5 in hAEP
152 (Fuentes-Prior and Salvesen, 2004). Interestingly, none of the four conserved potential N-linked
153 glycosylation sites in hAEP and mouse AEP (mAEP) are found in HaAEP1 (**Figure 2—figure**
154 **supplement 2**). Given that these glycosylation sites have been predicted to protect AEP from
155 non-specific protease activation it is intriguing to find that the only two potential N-linked
156 glycosylation sites in HaAEP1 (N138 and N143) are located on the opposite side of the protein
157 to the activation peptide in mammalian AEPs, at the beginning and center of the α2 helix,
158 respectively (Dall and Brandstetter, 2016). Moreover, only N138 is prominently surface
159 exposed, suggesting that N-linked glycosylation in HaAEP1 is not utilized for mitigation against
160 non-specific premature activation.

161 **HaAEP1 displays a tetrahedral intermediate in the active site**

162 The 1.8 Å resolution of our HaAEP1 structure allowed us to observe a succinimide moiety (Snn)
163 below the catalytic His in a location identical to human AEP (**Figure 3A**) that was hypothesized
164 to play a role in peptide ligation (Dall et al., 2015). We distinguished dual conformations of the
165 catalytic His178 and Cys220 residues which we hypothesize represent conformational changes
166 that occur during catalysis and correspond to substrate free and reaction intermediate states
167 (**Figure 3A**). In the intermediate state the catalytic Cys Sδ is oriented ~95° towards the Nδ of the
168 catalytic His imidazole ring which is orientated ~3.8 Å closer to this residue in the corresponding

intermediate state. Moreover, flexibility of His in a relatively open pocket free of steric hindrance (**Figure 3—figure supplement 1**) suggests an additional role for conformational shifts of the His in catalysis as seen with a range of proteases (McLuskey et al., 2012; Clark, 2016; Chekan et al., 2017). In the proposed alternate resting state the catalytic Cys Sδ is oriented towards the backbone amine of the highly conserved Gly179, reducing the distance between them from 5.4 Å to 3.6 Å. In this orientation the backbone amine might function in stabilizing proton abstraction from the Cys thiol.

Close examination of the electron density in the active site of HaAEP1 suggested a small peptide chain is intermittently bound to the intermediate conformation of Cys220 (**Figure 3B**). Given AEPs specificity for Asx and the nature of the electron density, we built and refined a tetrahedral complex of a three residue peptide ligand (AAN) bound to the active site Cys of HaAEP1 with partial occupancy (**Figure 3B**). Residues Ala-Ala were modelled upstream of the P1 Asn due to the weak electron density away from the peptide backbone at these residues. The AAN peptide ligand allowed characterization of interactions likely to exist between HaAEP1 and P1 Asn and main chain of residues P2-P3 during substrate recognition. Due to the high sequence conservation of the HaAEP1 active site with mammalian AEPs and conserved substrate orientation, many of the interactions with the substrate match those observed for inhibitors of hAEPs (Dall et al., 2015) (**Figure 3C**). The presence of this unexpected substrate in the HaAEP1 active site is likely to be a product of auto-activation. The continuous electron density between Cys220 and the P1 Asn supports the interpretation of the formation of a tetrahedral intermediate that is stabilized by the presence of an oxyanion hole formed by His178, Gly179 and the backbone amine of Cys220 that is more congruous to the electron

density than an acyl intermediate or free peptide (**Figure 3B — figure supplement 2**). Moreover, the observed short distance between the tetrahedral intermediate carbon atom and catalytic cysteine sulfur atom is incompatible with the absence of a covalent interaction. Main chain amino groups of Gly residues have previously been proposed to function in the creation of an oxyanion hole stabilizing the formation of a tetrahedral intermediate with the substrate in a wide range of cysteine proteases (Fuentes-Prior and Salvesen, 2004). Furthermore, previous observations of tetrahedral intermediates and enzyme-product complexes with serine proteases have been shown to exhibit a pH-dependent equilibrium (Wilmouth et al., 2001; Radisky et al., 2006; Lee and James, 2008). Similarly, the trapping of this intermediate state may have been fortuitously facilitated the activation of HaAEP1 at pH 4, followed by crystallization at pH 7. Although alternative conformations of the active site Cys in hAEP have been described previously, this tetrahedral intermediate state has not been described before (Dall et al., 2015).

Modulation of pH enables HaAEP1 to perform macrocyclization

The similarity of the HaAEP1 active site to OaAEP1 and the revelation it contained a reactive succinimide prompted us to test an enzyme preparation similarly taken to pH 4.0 against the modified SFTI-1 precursor substrate SFTI(D14N)-GLDN substrate. HaAEP1 had previously been unable to create a macrocyclic product from SFTI(D14N)-GLDN, but had been shown to efficiently cleave it at a rate, k_{cat}/K_m value of 610 M^{-1} S^{-1}, similar to rates published for other AEPs (Bernath-Levin et al., 2015). To our surprise, our new preparations of HaAEP1 taken to pH 4.0 produced cyclic SFTI(D14N) when the reactions were conducted at pH 6.5 (**Figure 4B** WT, **Figure 4—figure supplement 1A** WT). Previously, HaAEP1 had been purified at pH 8, activated at pH 5 and then used in reactions at pH 5. The HaAEP1 preparations that were able to

macrocyclize were similarly purified at pH 8, but activated at pH 4 then returned to pH 6.5. Activation at lower pH has been shown to be more effective at removing the cap domain, which in mammalian AEPs had a propensity to re-ligate (Zhao et al., 2014). Higher pH could also favor macrocyclization by facilitating the deprotonation of the Gly N-terminus at the active site, priming it to attack its C-terminus, which is held in the thioacyl intermediate at the active site.

Residues local to the catalytic dyad influence product formation

To investigate the mechanism of macrocyclization by HaAEP1 we identified several residues for site directed mutagenesis. Firstly, we hoped to clarify the roles of Cys220 and Snn177 in macrocyclization through Ser and Ala mutations, respectively. Furthermore, we also mutated Snn177 to Gly, as the C13 protease GPI8 has also been proposed to carry out an intramolecular transpeptidation reaction yet displays a Gly residue at a location equivalent to Snn177 (Zacks and Garg, 2006) (**Figure 2—figure supplement 2**). Secondly, we hoped to alter the ability of HaAEP1 to macrocyclize its native substrate SFTI-1 by altering the residues that fine tune this catalysis. By modelling the binding of an NMR structure of PawS1 (Franke et al., 2017) and the N-terminally cleaved SFTI-1 precursor (SFTI-GLDN) to HaAEP1 we were able to hypothesize the location of a hydrophobic S2' binding region encompassing strand β_{II} with G185 located at the center of this region (**Figure 4—figure supplement 2**). In addition, previous studies have also implicated that Asn, Glu and Asp residues proximal to the catalytic Cys and His function in catalysis (Dall and Brandstetter, 2013; Zhao et al., 2014). Inspection of the HaAEP1 structure revealed that several of these residues (including Asn73 and Glu221) are conserved in HaAEP1. Although, Glu221 is oriented away from Cys220, in a direction similar to that seen in hAEP

234 bound to human cystatin E, it might assume alternate conformations due to its solvent
235 exposure and high B-factor.

236 The mutant HaAEP1 proteins were expressed in *E.coli* and analyzed by circular dichroism
237 (**Figure 4—figure supplement 3**), with wild-type (WT) HaAEP1 displaying a melting temperature
238 of ~52°C similar to previous reports for hAEP (Dall and Brandstetter, 2013), and mutants
239 displaying similar spectra to WT.

240 Incubation of WT HaAEP1 with a fluorophore labelled (BODIPY) activity-based probe (Lu et al.,
241 2015) illustrated its heterogeneity in size following activation at pH 4 and incubation with
242 substrate at pH 6.5, which has previously been observed for AEP proteins both *in vivo* and *in*
243 *vitro* and speculated to be the result of processing of non-glycosylated forms (Zhao et al., 2014)
244 (**Figure 4a**). This probe also revealed a C220S mutation in HaAEP1 to result in the enzyme
245 becoming incapable of pH-dependent activation (**Figure 4a**). Further analysis of AEP mutants
246 activity using the BODIPY probe illustrates the substantial heterogeneity in size between active
247 WT and N73A, N73D, D177G, D177A or E221K mutant AEPs, likely due to autocatalytic
248 processing, yet indicates that they remain active (**Figure 4—figure supplement 1D**).

249 The use of a seleno-modified synthetic SFTI(D14N)-GLDN substrate, which we have previously
250 shown to be processed by HaAEP1 and CeAEP1(Bernath-Levin et al., 2015), allowed for a
251 comparison of activity profiles of HaAEP1 mutants through the quantification of distinctive
252 isotopic cyclic, acyclic and unprocessed peak areas by MALDI-MS (**Figure 4B-C**, **Figure 4—figure**
253 **supplement 1A**). This comparison reveals subtle changes between mutants in the ratio of cyclic,
254 acyclic and unprocessed peptides and confirmed that the HaAEP1-C220S mutant is unable to

cleave as evidenced by the lack of a peak at mass 1608 or 1626 (**Figure 4**, **Figure 4—figure supplement 1**). As expected, mutation of the second residue of the catalytic dyad (H178A) also results in a drastic reduction in activity, based on SFTI(D14N)-GLDN processing and activity based BODIPY probe results, confirming the significance of C220 and H178 in AEP activity (**Figure 4**, **Figure 4—figure supplement 1**). Interestingly, the H178A mutation does not abolish HaAEP1 activity, as previously seen with mAEP, and suggests a third residue could facilitate proton transfer at the active site (Zhao et al., 2014). Similar results to the H178A mutation were also observed for G185S that was directed at altering the hydrophobic S2' binding region. The effect of the G185S mutation suggests G185 has a role in substrate recognition and could sterically alter the conformation of H178 and Snn177 (**Figure 4**, **Figure 3—figure supplement 1**). The mutation E221K, which has previously been shown to increase endopeptidase activity in hAEP, resulted in a loss of cyclic product as shown by an absence of a peak of mass 1608 (**Figure 4B, Figure 4—figure supplement 1A)**. Mutation of N73A leads to a higher ratio of cyclic to acyclic product as shown by an increased peak area relative to WT of mass 1608 and a reduced acyclic product peak of mass 1626. N73D and D177G mutants appear to process SFTI(D14N)-GLDN in a manner similar to WT. Whereas the large fraction of unprocessed SFTI(D14N)-GLDN, mass 2025, after incubation with D177A illustrates a reduction in processing efficiency, as previously noted for the mAEP D149A mutant (Zhao et al., 2014).

The effect of the N73A, E221K and D177G mutations on AEP catalyzed ligation was probed further by investigating each mutant's ability to revert to its inactive form via re-ligation of its cap domain upon shift to neutral pH. As described for mAEP and equivalent mutants (Zhao et al., 2014), incubation of HaAEP1 at pH 4 led to an irreversible dissociation of its cap domain

(**Figure 4—figure supplement 4**). However, following activation at pH 5.5 and 6.5, upon shifting to pH 8 the WT, N73A, E221K and D177G mutants were able to re-ligate the cap onto the core domain resulting in the formation of the inactive pro-enzyme with a Mw ~52 kDa as evidenced by SDS-PAGE and activity based probes (**Figure 4—figure supplement 4**).

These results confirm the importance of the catalytic dyad in AEP function and show that HaAEP1, like mAEP and the closely related C13 family member GPI8, is able to perform its ligation reaction in the absence of a Snn residue. Moreover, subtle mutations affecting the stability of the catalytic dyad might favor either hydrolysis or macrocyclization.

Active HaAEP1 exhibits an open surface amenable to macrocyclization

In order to discern the structural determinants that favored AEPs to be recruited independently by evolution multiple times for macrocyclization we compared the structure of HaAEP1 and its predicted binding mode with the crystal structures of closely related cysteine proteases: sortase A, papain and metacaspase MCA2 (Suree et al., 2009; Chu et al., 2011; McLuskey et al., 2012) (**Figure 5**).

Sortase A is a *Staphylococcus aureus* cysteine protease which catalyzes a similar transpeptidation reaction to AEP, ligating proteins bearing a LPXTG motif to peptidoglycan precursors in the bacterial cell wall (Mazmanian et al., 1999). NMR studies have shown that the resolution of the transpeptidation thioacyl intermediate reaction occurs through nucleophilic attack of a lipid terminal amine in a steep valley between the β7-β8 and β4-β5 loops (Suree et al., 2009). In comparison to HaAEP1, the sortase A β7-β8 loop protrudes much further from the active site and the aromatic residues F122, Y128, W194 in these loops orient over the catalytic

298 Cys (**Figure 5B**). Together these loops, despite reported flexibility, would likely impart
299 considerable steric hindrance for macrocyclization by inhibiting both the resolution of the
300 intermediate by the N-terminal amine group and binding at the S2' substrate 'tail'.

301 Papain from the melon tree *Carica papaya* is the archetypal plant C1 family cysteine protease
302 and has been found to bind to a cystatin homologue, tarocystatin, in a manner analogous to
303 that of hAEP binding cystatin (Otto and Schirmeister, 1997; Chu et al., 2011; Dall et al., 2015).
304 However, inspection of the papain active site reveals a topology that is not conducive for
305 macrocyclization with the catalytic triad buried deep within the protein and steric hindrance for
306 peptide N-terminal attack likely to be imparted by W177, D158 and the extended alpha helical
307 loops of the α3 and α7 regions (**Figure 5C**).

308 Metacaspases are expressed in plant, fungi and protozoa and display a C14 caspase domain
309 that is structurally homologous to human caspases (Tsiatsiani et al., 2011). Metacaspases reside
310 within the same CD clan as AEP, but exhibit a strict specificity for a cleavage following Arg or Lys
311 (Vercammen et al., 2004). Currently no crystal structure is available for a plant metacaspase,
312 however the crystal structure of the protozoan metacaspase MCA2 reveals an architecture that
313 like papain would likely be unfavorable for macrocyclization due to steric hindrance around the
314 active site from several prominent loops; including the β1-α1, βA-βB and 280-loop (McLuskey et
315 al., 2012) (**Figure 5D**).

316 The crystal structure of active HaAEP1 suggests that the combination of a relatively open
317 reaction interface with space around the active site allowing for catalytic residue flexibility has

resulted in the convergence upon AEPs for macrocyclization ahead of the other 30 families of cysteine proteases in plants (Rawlings et al., 2016).

AEPs have an intrinsic ability to perform peptide macrocyclization

Given the sequence similarity between AEPs and the conservation of residues involved in catalysis we hypothesized that the ability to macrocyclize peptides might be inherent to AEPs. To test this hypothesis we recombinantly expressed two AEPs from species which are currently not thought to contain cyclic peptides of any kind; *Arabidopsis thaliana* (AtAEP2) and *Ricinus communis* (RcAEP1), respectively. These AEPs were purified and activated as described for HaAEP1 (pH 4) and then incubated with non-native substrates; SFTI-GLDN and SFTI(D14N)-GLDN, at a pH that favors ligation (pH 6.5). Under these conditions RcAEP1 was able to macrocyclize both SFTI-GLDN and SFTI(D14N)-GLDN substrates whereas AtAEP2 was able to macrocyclize only SFTI(D14N)-GLDN (**Figure 6**). These findings further supports our hypothesis that the structural features of AEPs described above have allowed for the convergence upon AEP for peptide macrocyclization reactions.

Discussion

Herein, we have described the structure of an active plant AEP containing a peptide ligand with partial occupancy bound to the active site catalytic Cys as a tetrahedral intermediate, illustrating conformational flexibility in the AEP catalytic dyad. This structure has enabled us to predict a model for SFTI-1 macrocyclization by HaAEP1 where the GLDN tail of the SFTI-1 precursor orients in a manner analogous to cystatin binding to hAEP, over the catalytic His to a hydrophobic patch on the β_{II} region, and where the N-terminus attack occurs between

339 diminutive β_I sheet and $\beta5$-β_{IV} loop in a manner analogous to the attack of thioacyl
340 intermediates by peptidoglycan precursors in sortase-catalyzed transpeptidation reactions
341 (Suree et al., 2009; Clancy et al., 2010). This model complements the requirement for a small
342 amino acid followed by a hydrophobic residue at P1' and P2' due to their orientation over the
343 catalytic His and towards the hydrophobic β_{II}-β_{III} region, respectively.

344 Several Cys proteases, including AEPs, have previously described a minor role for a third
345 Asx/Arg residue in catalysis (Vernet et al., 1995; Dodson and Wlodawer, 1998; Fuentes-Prior
346 and Salvesen, 2004; Clancy et al., 2010; Buller and Townsend, 2013; Dall and Brandstetter,
347 2013). This third residue has been suggested to function in stabilization and orientation of the
348 His imidazole ring. Interestingly, functional analysis of HaAEP1 catalytic triad residue N73
349 revealed that an N73A mutation resulted in the production of a higher ratio of cyclic SFTI-1. The
350 observation of potential conformational shifts in the crystal structure of HaAEP1 during
351 catalysis suggests that the loss of an orientating N73 side chain allows for further
352 conformational flexibility of H178. This increased flexibility of His in a relatively open pocket
353 that might accommodate a range of rotamers could reduce steric hindrance of an N-terminus
354 entry towards the acyl intermediate and also facilitate deprotonation of the substrate N-
355 terminus.

356 The requirement for space and flexibility between the catalytic dyad favors the convergence for
357 peptide macrocyclization upon Cys proteases over Ser proteases due to the close proximity
358 between Ser-His residues in these proteases, an inherent requirement based upon the reduced
359 nucleophilic properties of Ser (Buller and Townsend, 2013). Indeed, a recent structure of a

macrocyclizing Ser protease, PCY1, suggested that a shift in the catalytic His away from Ser is required for macrocyclization, based on a comparison to their hydrolytic relatives, and suggested a role for the catalytic His in deprotonation of the attacking peptide N-terminus (Chekan et al., 2017). Furthermore, comparison with other Cys proteases suggests that AEPs have been converged upon for macrocyclization based on their relatively flat and open catalytic site. In contrast to a recent hypothesis for efficient macrocyclization based on a comparison of AEP structures focused on a region closer to the S4 pocket, here we suggest that differences in AEP catalytic efficiency and substrate specificity will be defined by subtle amino acid differences in the β_{II}-β_{III} region, β_I sheet and β5-β_{IV} loop and the S1-S5 pocket (Yang et al., 2017). Moreover, this orientation of the tail away from the β_I sheet and β5-β_{IV} loop is hindered in other caspases which exhibit a relatively straight substrate channel (**Figure 5**) and thus prevent simultaneous attack of the thioacyl intermediate by the N-terminus loop upon scissile bond cleavage.

Following activation, AEP functionality has been illustrated to be dependent on a delicate balance between pH and stability, where endopeptidase activity is favored at a low pH (~pH 4) and ligase activity is favored at a higher pH (~pH 6) (Dall et al., 2015). A pH-dependent activity has been well documented for cysteine proteases and is a function of the formation of a thiolate and imidazolium ions on the catalytic dyad of Cys and His residues (Fuentes-Prior and Salvesen, 2004; Frankel et al., 2005). Given the hypothesized role of His in deprotonating the attacking N-terminal peptide chain in peptide macrocyclization it would therefore be expected that at a low pH the catalytic His would be more readily protonated and hence this reaction might be less frequent resulting in more hydrolysis at low pH. This hypothesis is supported by the low level of endopeptidase activity observed with Ala mutations of the catalytic His in

HaAEP1 and other proteases, illustrating that thiolates might form and that hydrolytic resolution of the thioacyl intermediate might occur in the absence of His (Frankel et al., 2007; Zhao et al., 2014). Previous reports have also suggested that a local Gly amide backbone might facilitate catalysis via a transfer of a proton to the leaving group of the tetrahedral intermediate from a hydrogen bonded water molecule; analysis of the HaAEP1 structure reveals Gly179 could potentially perform this role in the absence of His (Brady et al., 1999). Moreover, mutation of the acidic residue following the catalytic Cys to a basic residue in hAEP has previously been shown to enhance catalytic efficiency by decreasing the local pKa of the Cys residue (Dall and Brandstetter, 2013). Interestingly, we found an equivalent mutation in HaAEP1 to result in the loss of peptide macrocyclization indicating that E221 could also aid in deprotonation of the incoming N-terminus or that the residue's larger side chain and expected orientation in activating the catalytic Cys could impart steric hindrance upon the N-terminus attack of the thioacyl intermediate (**Figure 4**, **Figure 4—figure supplement 1**). However, given the finding that E221K mutation is unable to prevent re-ligation of the cap domain to the active core domain upon pH shift in HaAEP1, there is likely to be redundancy between the local residues in creating a nucleophilic amine group to complete a transpeptidation reaction.

The finding that AEPs from species that lack cyclic peptides may be coaxed into performing peptide macrocyclization of a linear peptide under favorable conditions significantly expands the potential use of AEPs for the production of cyclic peptides (**Figure 6**). Moreover, further investigation into the subtle nuances that define substrate specificity and catalytic activity between AEPs is warranted, with differences in the β_{II}-β_{III} region, β_{I} sheet, β5-β_{IV} loop and around the S1-5 pocket including the variable α5-β6 and β_{IV}-β_{V} loops likely to be key (James et

al., 2017). Indeed, in comparison to HaAEP1 and OaAEP1, the efficient peptide macrocyclizing AEP butelase 1 displays several different amino acids that could be responsible for this AEPs efficacy. These differences include shorter sidechains in the β_I sheet and $\beta5$-β_{IV} loops that may reduce steric hindrance on a peptides N-terminus during attack on a thioacyl intermediate (HaAEP1: P181, Q245, N247 OaAEP1: A178, T242, S244 Butelase-1: A168, G232, S234) and differences in the β_{IV}-β_V region that could generate substrate specificity (**Figure 5—figure supplement 1**).

Active HaAEP1 exhibits a succinimide at the same position as hAEP which has previously been postulated to perform a Cys-independent ligation reaction through cyclic rearrangement with the P1 Asx side chain (Dall et al., 2015). Functional analysis reveals that HaAEP1 is able to perform peptide macrocyclization despite D177G/A mutations that cannot form a succinimide group (**Figure 4B-C**). Moreover, these mutants continued to perform re-ligation of the cap domain when activated at pH >5.5 (**Figure 4—figure supplement 4**), a result that has previously been shown with mAEP (Zhao et al., 2014). In the absence of a critical requirement for succinimide formation in macrocyclization the question remains as to why this relatively unstable aspartimide appears stable in AEP crystal structures and is largely conserved in the C13 proteases. Succinimides have been shown to form more readily when adjacent to His residues as the His Nδ abstracts a proton from the His backbone NH allowing the deprotonated main chain amine to attack the Asx side chain (Brennan and Clarke, 1995; Takahashi et al., 2016). Once formed this succinimide could enhance the activity of the catalytic His by virtue of reducing stabilizing hydrogen bonding interactions with the carboxyl terminus and be involved in the proper positioning of the catalytic His. A further potential role for succinimides could

present upon their hydrolysis through racemization and favored formation of iso-Asp (Geiger and Clarke, 1987; Reissner and Aswad, 2003). This iso-Asp could potentially represent a subtle mode of AEP regulation as the orientation of this side chain towards the S1 pocket is likely to disrupt substrate binding. Such iso-Asp residues have previously been proposed to regulate protein activity by a time-dependent molecular switch (Geiger and Clarke, 1987; Ritz-Timme and Collins, 2002).

In the absence of caspases, plants have evolved a wide range of cysteine proteases to ensure functional redundancy in a myriad of highly regulated programmed cell death pathways in response to environmental and developmental cues (Fagundes et al., 2015). Furthermore, plants have developed a variety of strategies to control the destructive prowess of these proteases including the expression of proteases as inactive zymogens with cofactor dependency, compartmentalization and the production of protease inhibitors (Martinez et al., 2012). Of these cysteine proteases the AEPs have recently attracted considerable interest due to their ability to carry out peptide macrocyclization and their potential application in the synthesis of pharmacoactive cyclic peptides. Herein, the structural and functional analysis of HaAEP1 has revealed residues that are able to favor the production of cyclic or acyclic products from SFTI(D14N)-GLDN. Moreover, we have modelled a binding mode for productive macrocyclization of the HaAEP1 natural ligand SFTI-1, based on a comparison with related cysteine proteases that is likely to be conserved between AEPs, where substrate specificity is defined by the amino acids around the binding site for respective AEPs. Furthermore, we have shown that AEPs from diverse species lacking cyclic peptides are able to perform macrocyclization under favorable pH conditions. These findings provide the foundation for

448 further optimization of AEPs, potentially widening the array of peptide substrates that could be
449 cyclized by AEPs.

450 Methods

451 Protein expression and purification

452 DNA sequence encoding residues 28-491 of HaAEP1 (accession code: KJ147147), *Ricinus*
453 *communis* AEP (RcAEP1) residues 58-497 (accession code: D17401) and *Arabidopsis thaliana*
454 AEP (AtAEP2) residues 47-486 (accession code: Q39044) were cloned into a pQE30 (QIAGEN)
455 expression vector with an N-terminal six-histidine tag and expressed in the SHuffle strain *E. coli*
456 (New England Biolabs) transformed with pREP4 (QIAGEN). Briefly, cultures were grown at 30°C
457 to an OD_{600} of 0.8-1.0 in Luria Broth medium containing 100 μg/mL ampicillin and 35 μg/mL
458 kanamycin with expression induced at 16°C with 0.1 mM isopropyl β-D-1-thiogalactopyranoside
459 then cells cultured overnight. Cell pellets were collected by centrifugation and lysed by
460 ultrasonication in 50 mM Tris (pH 8.0), 100 mM sodium chloride, 0.1% Triton X-100. The soluble
461 fraction was then harvested by centrifugation and the supernatant was incubated (batch wise)
462 with Ni-NTA resin overnight at 4°C. The resin was then washed with 20 mL of 50 mM Tris (pH
463 8.0) and 20 mL of 50 mM Tris (pH 8.0) 20 mM imidazole and protein was eluted stepwise with
464 20 mL of 50 mM Tris (pH 8.0) 300 mM imidazole. For purification for crystal screens, six-
465 histidine tagged inactive protein was purified, using an ÄKTA FPLC platform, by anion-exchange
466 chromatography (HiTrap Q HP 5 mL) with a gradient of 0 to 500 mM sodium chloride over 60
467 min and then either activated by dialysis in 100 mM citric acid - sodium citrate buffer (pH 4.0)
468 containing 50 mM sodium chloride and 5 mM DTT, overnight at 16°C, or directly concentrated

and further purified by size-exclusion chromatography (HiLoad 16/600 Superdex 200) in 50 mM Tris (pH 8.0), 50 mM sodium chloride. Following dialysis active AEP was separated from insoluble material by centrifugation and purified by size-exclusion chromatography (HiLoad 16/600 Superdex 200) in 100 mM citric acid – sodium citrate buffer (pH 4.0) containing 50 mM sodium chloride. Protein was assessed for purity by SDS-PAGE.

HaAEP1 site-directed mutations were made following the Stratagene QuickChange protocol with the following primers: The N73A mutation was made with forward primer 5′-GTA GCA AAG GTT ATG GTG CTT ATC GTC ATC AGG CC-3′ and reverse primer 5′-GGC CTG ATG ACG ATA AGC ACC ATA ACC TTT GCT AC-3′; the N73D mutation with 5′-GTA GCA AAG GTT ATG GTG CTT ATC GTC ATC AGG CC-3′ and 5′-GCC TGA TGA CGA TAA TCA CCA TAA CCT TTG CTA C-3′; The D177A mutation with 5′-CTG TTT TAT AGC GCT CAT GGT GGT CCG G-3′ and 5′-CCG GAC CAC CAT GAG CGC TAT AAA ACA G-3′; the D177G mutation with 5′-CTG TTT TAT AGC GGC CAT GGT GGT CCG GG-3′ and 5′-CCC GGA CCA CCA TGG CCG CTA TAA AAC AG-3′; the H178A mutation with 5′-CTG TTT TAT AGC GAT GCT GGT GGT CCG GGT G-3′ and 5′-CAC CCG GAC CAC CAG CAT CGC TAT AAA ACA G-3′; the G185S mutation with 5′-GTC CGG GTG TTC TGA GTA TGC CGA ATG AAC-3′ and 5′-GTT CAT TCG GCA TAC TCA GAA CAC CCG GAC-3′; the C220S mutation with 5′-TGA TTT ATC TGG AAG CAT CTG AAA GCG GCA GCA T-3′ and 5′-ATG CTG CCG CTT TCA GAT GCT TCC AGA TAA ATC A-3′; and the E221K mutation with 5′-GAT TTA TCT GGA AGC ATG TAA GAG CGG CAG CAT TTT TGA AGG-3′ and 5′-CCT TCA AAA ATG CTG CCG CTC TTA CAT GCT TCC AGA TAA ATC-3′. Mutations were verified by sequencing and expressed as described for WT.

Crystallization and data collection

Protein was concentrated to 10-15 mg/mL and crystal screening performed using the sitting-drop-vapor diffusion method with 80 µL of reservoir solution in 96-well Intelli-Plates at 16°C. Protein to mother-liquor ratios for the sitting drops were varied in each condition: 0.1:0.1, 0.1:0.2, and 0.2:0.1 µL. Crystals of active HaAEP1 were obtained in 0.1 M HEPES (pH 7.5), 1.4 M sodium citrate tribasic dihydrate. Single crystals were soaked in mother-liquor supplemented with 20% glycerol as a cryoprotectant prior to being flash-frozen and stored in liquid nitrogen. Data collection was performed at 100 K on the Australian MX2 (micro-focus) beamline using a wavelength of 0.9537 Å and diffraction data for crystals were collected to a resolution of 1.8 Å.

Structural determination, refinement and model building.

Diffraction data were processed using iMOSFLM and scaled with AIMLESS from the CCP4 program suite (Battye et al., 2011; Winn et al., 2011) in space group $P3_121$ with unit cell dimensions $a = b = 77.03$, $c = 108.17$. A sequence alignment of HaAEP1 and human AEP1 was generated using ClustalO and used to create a search model of HaAEP1 based on the last common atom of human AEP (4FGU) using CHAINSAW. The structure of HaAEP1 was solved by molecular replacement with PHASER using this search model, followed by automatic building with ARP/WARP. Manual building and refinement was performed in iterative cycles with COOT and REFMAC5 using the CCP4 program suite. Structural analysis and validation were carried out with COOT and MolProbity (Emsley and Cowtan, 2004; Chen et al., 2010). Crystallographic data and refinement statistics are summarized in **Table 1** with Ramachandran plot values calculated from COOT. The peptide AAN modelled into the HaAEP1 active site was oriented into the active site based on the similar mode of cystatin binding to human AEP (4N6O)(Dall et al., 2015). The

tetrahedral intermediate was evidenced from initial visualization in Fo-Fc electron density difference maps using a polder OMIT map as implemented in phenix.polder (Liebschner et al., 2017). Coordinates and structure factors were deposited into the Protein Data Bank (PDB) under accession code 6AZT. Tetrahedral complex of a three residue peptide ligand (AAN) bound to the active site Cys220 defined as CX9 in PDB file. Figures illustrating structures were generated using PyMol, electrostatic surface potentials were contoured at ±10 kT/e using an APBS PyMol plugin (Dolinsky et al., 2007; Schrodinger, 2010). Models of *C. ensiformis* and *C. ternatea* AEPs *w*ere generated using the I-TASSER server (Roy et al., 2010).

Circular Dichroism

Proteins purified in 50 mM Tris (pH 8.0) 300 mM imidazole were concentrated in an 30 kDa Amicon centrifugal filter and buffer exchanged with an excess of 10 mM sodium phosphate buffer (pH 8). Concentrations were checked by absorbance at 280 nm with a NanoDrop using the extinction coefficient of Pro-AEP and samples diluted to 0.1 mg/ml or 0.2 mg/ml. CD measurements were made in triplicate using a Jasco J-810 CD spectrometer with a 0.1-cm quartz cuvette using a 1 nm bandwidth on a 0.1 mg/ml sample. CD wavelength spectra were collected from between 200-260 nm a rate of 1 nm/sec at 20°C. Melt curves were collected using the same bandwidth at 222 nm with temperature increasing at a rate of 1°C/min from 20 - 95°C on a 0.2 mg/ml sample. WT HaAEP1 melting temperature was interpolated from melt curve using a sigmoidal four parameter logistic regression fit (GraphPad Prism).

Protein activity analysis

AEPs were activated by dialysis for four hours at room temperature in activation buffer (20 mM sodium acetate pH 4.0, 5 mM DTT, 100 mM sodium chloride, 1 mM EDTA) followed by a second dialysis into ligation buffer (20 mM MES pH 6.5, 0.5 mM DTT, 100 mM sodium chloride, 1 mM EDTA acid). Protein concentrations were determined by measuring absorbance at A280 using a NanoDrop (1 Abs = 1.0 mg/mL). For mass spectrometry analysis of the processing of SFTI(D14N)-GLDN or native SFTI-GLDN by WT HaAEP1, RcAEP, AtAEP2 and mutant HaAEPs, AEPs at a concentration of 40 μg/mL were incubated with 0.25 mM peptide with a diselenide bond and 25 mM native (i.e. disulfide) SFTI-1 as an internal standard in activity buffer (20 mM MES pH 6.5, 5 mM DTT, 1 mM EDTA). Reactions were carried out at 37°C for 16 hours. Three independent reactions were performed for each protein tested. The reactions were stopped by dilution 100-fold in 50% acetonitrile, 0.1% formic acid and spotted with an α-cyano-4-hydroxycinnamic acid matrix onto a plate for analysis by MALDI-MS. Quantification of peak area by MALDI-MS was calculated using the internal standard to normalize for ionization efficiency as described previously (Bernath-Levin et al., 2015). Briefly, for activity probe analysis 50 μL of AEP at 10 μg/mL was incubated with 1 μM of the BODIPY probe JOPD1 (Lu et al., 2015) at room temperature overnight and protected from light. The labelling reaction was stopped by the addition of 10 μL of 6x SDS-PAGE loading buffer containing β-mercaptoethanol and proteins separated using 4-12% Bis-Tris SDS-PAGE gels as previously described (Lu et al., 2015). Labeled proteins were visualized in gel with excitation and emission wavelengths of 532 and 580 nm using a Typhoon 9500 (GE Healthcare).

Analysis of reversal of AEP activation

WT, D177G, N73A and E221K proteins were purified by affinity chromatography as described above and activated by overnight (4°C) dialysis at pH 4.0/5.5/6.5 in 100 mM citric acid - sodium citrate buffer containing 50 mM sodium chloride and 5 mM DTT, with pH adjusted by addition of 1.0 M Tris-HCl pH 8. Prior to pH-shift, sample aliquots were flash frozen and stored at -80°C for subsequent activity probe analysis and SDS-PAGE analysis. Remaining activated protein was then returned to the previous buffer but with pH adjusted to pH 8.0 by adding 1.0 M Tris-HCl pH 8.0, as previously described, and incubated overnight at 4°C (Hara-Nishimura et al., 1998). Following dialysis at pH 8.0 samples were flash frozen as described above. Proteins were then analyzed for activity, as described above, using the BODIPY activity based probe. Following in gel visualization of active protein gels were immediately fixed in Coomassie Blue stain for comparison of active and inactive protein.

Competing interests

The authors declare no financial and non-financial competing interests.

Acknowledgements

The authors thank Dan Tawfik for critical insights during project design and Renier van der Hoorn for the BODIPY probe JOPD1. J.S.M. was supported by an Australian Research Council (ARC) Future Fellowship (FT120100013). S.G.N. and K.V.S. were supported by the Australian Government's Research Training Program. J.H. and this work were supported by ARC grant DP160100107 to J.S.M. and Dan Tawfik. HaAEP1 is deposited in PDB under accession 6AZT.

Figure Captions

Figure 1. Examples of enzyme catalyzed formation of cyclic RiPPs. Cyclic RiPPs that are enzyme catalyzed from linear peptide precursors are commonly flanked by an N-terminus leader sequence and a C-terminus follower sequence prior to cyclization. These flanking sequences commonly aid in substrate recognition and catalysis. Well characterized examples of enzymes that are able to perform peptide macrocyclization currently include cysteine proteases (*Helianthus annuus AEP* – HaAEP, and *Oldenlandia affinis* AEP – OaAEP; following removal of the leader peptide by an as yet undefined enzyme), serine proteases (patellamide G – PatG, and peptide cyclase 1 – PCY1; following removal of the leader peptide by oligopeptidase 1 – OLP1) and the ATP dependent ATP-grasp ligases (microviridin C and B – MdnC/B, and microcin J25 C - McjC; following removal of the leader peptide by microcin J25 B – McjB). These examples also illustrate a range of cyclic peptide scaffolds that may have potential therapeutic applications.

Figure 2. Architecture of HaAEP1 versus other AEP structures. (**A**) Cartoon representation of the overall topology of HaAEP1 with major α-helices and β-sheets annotated. Residues 304-309 that exhibited weak electron density and are missing from the model are shown as red dotted loop. (**B**) Comparison of HaAEP1 core domain (green) with OaAEP1 (magenta) and hAEP (grey) with bound chloromethylketone inhibitor (black) illustrating high overall structural similarity. Expanded surface and cartoon representations of highlighted regions of β_{IV}-β_V substrate binding region (blue backgrounds) and catalytic region (β_I sheet, β5-β_{IV} loop and β_I-β_{III} region) orientated over the catalytic His residue (yellow backgrounds) are shown below illustrating the

residue differences that could alter substrate specificity (shown in stick format). Also see Figure 2–figure supplement 1-3.

Figure 3. Outstanding features of the HaAEP1 active site. (**A**) Catalytic residues with dual conformations illustrated in the HaAEP1 active site with simulated annealing omit electron density maps (2 F_{obs} - F_{calc}) contoured at 1σ level. (**B**) Cross-eyed stereo view of polder OMIT map (F_{obs} - F_{calc}) calculated in the absence of the shown overlaid AAN tetrahedral intermediate, contoured at 3σ level. (**C**) Schematic representation of interactions between AAN tetrahedral intermediate (purple sticks) and labelled active site residues generated using LigPlot^{+} (Laskowski and Swindells, 2011). Residues forming hydrogen bonds (green) are illustrated with black sticks with distances shown (Å). Residues and atoms that provide hydrophobic interactions are highlighted with orange eyelash symbols. Also see Figure 3–figure supplement 1.

Figure 4. Mutagenesis studies of HaAEP1. (**A**) Activity based probe analysis of HaAEP1 WT and C220S mutant illustrates that WT is active at pH 6.5 after activation at pH 4 whereas C220S mutant remains inactive. In-gel fluorescence of activity based probe (right) and post-fluorescence Coomassie stain of SDS-PAGE gel (left). (**B**) MALDI-TOF spectra of SFTI(D14N)-GLDN processing by WT and a range of HaAEP1 mutants directed at altering the ability of HaAEP1 to cleave and macrocyclize the substrate SFTI(D14N)-GLDN. NEC – no enzyme control. (**C**) Quantitation of peak areas from **B**. Peak areas of mass 1608 - cyclic SFTI(D14N), mass 1626 - acyclic-SFTI(D14N) and mass 2025 - unprocessed seleno-Cys modified SFTI(D14N)-GLDN substrate, were normalized for ionization efficiency using an internal standard mass 1515 - native SFTI-1. Black - cyclic SFTI-1(D14N), grey - acyclic-SFTI(D14N) and white - unprocessed

613 seleno-Cys modified SFTI(D14N)-GLDN substrate. Peak areas with acyclic and cyclic forms have
614 previously been shown to exhibit similar ionization efficiencies (Bernath-Levin et al., 2015).
615 Error bars illustrate standard deviation n=3 (D177A n=2) technical replicates. Also see Figure 4–
616 figure supplement 1-4.

617 **Figure 5. Comparison of Cys protease active site topologies.** Substrates are oriented towards
618 the catalytic dyad of Cys and His, highlighted in yellow and blue sticks, respectively. Regions and
619 residues likely to impart steric hindrance on the proposed mode of macrocyclization by AEPs in
620 other Cys proteases are labelled. (**A**) Proposed binding mode for SFTI-1 macrocyclization (cyan
621 sticks and dots) based on the position of the tetrahedral intermediate in the HaAEP1 structure
622 with the N-terminus attacking the intermediate and the GLDN-tail (red-dots) oriented over the
623 catalytic His towards the β_{II}-β_{III} region. (**B**) Structure of sortase A (PDB 2KID) covalently bound to
624 an analog of the sorting signal (purple sticks). Residues implicated in transpeptidation are
625 highlighted with green sticks, in a region analogous to the β5-β_{IV} loop and β_{I}-β_{II} loop of HaAEP1.
626 (**C**) Structure of papain (PDB 3IMA) bound to residues 2-7 and 49-53 of tarocystatin (purple
627 sticks) illustrating restricted access to the catalytic dyad. (**D**) Structure of the metacaspase
628 $MCA2_{C213A}$ (PDB 4AFV) with residues 30-33 of the N-terminal domain (purple sticks) bound in
629 the predicted direction for substrate binding. Flexible regions of low electron density are shown
630 with grey dots. Also see Figure 5–figure supplement 1.

631 **Figure 6. The ability to macrocyclize is inherent to AEPs.** MALDI-TOF spectra of SFTI-GLDN (**A**)
632 and SFTI(D14N)-GLDN (**B**) with AEPs from species thought not to contain cyclic peptides, namely
633 *Arabidopsis thaliana* (AtAEP2) and *Ricinus communis* (RcAEP1). (**C**) Quantitation of peak areas

from seleno-Cys modified substrates in **A** (mass 1609 - cyclic SFTI-1, mass 1627 - acyclic-SFTI and mass 2026 - unprocessed SFTI-GLDN substrate) and **B** (mass 1608 - cyclic SFTI(D14N), mass 1626 - acyclic-SFTI(D14N) and mass 2025 - unprocessed seleno-Cys modified SFTI(D14N)-GLDN substrate) normalized to mass 1515 - native SFTI-1. Black – cyclic peptide, grey – acyclic peptide and white – unprocessed substrate. Error bars illustrate standard deviation n=3 technical replicates. (**D**) SDS-PAGE analysis of inactive (pH 8) and active (pH 6.5) AtAEP2 and RcAEP1 proteins.

Table

Table 1. Crystallography data collection and refinement statistics. Numbers in parenthesis refer to the highest resolution bin.

Data collection	
Space group	$P3_1\ 2\ 1$
Unit cell dimensions	
a, b, c (Å)	77.03, 77.03, 108.17
α, β, γ (°)	90.00, 90.00, 120.00
Wavelength	0.9537
Resolution (Å)	1.8
R_{merge} (%)	6.0 (43.3)
$I/\sigma I$	14.7 (2.2)
Completeness (%)	93.8 (65.1)
Redundancy	4.1 (1.8)
$CC_{1/2}$	0.997 (0.727)
Refinement	
Resolution (Å)	66.71-1.80
No. reflections	31205
R_{work}/R_{free}	15.15/18.88
No. Atoms	2415
Protein	2168
Water	229
Ligand	18
Wilson B ($Å^2$)	15.7
Average refined B-factor ($Å^2$)	
Protein only ($Å^2$)	22.2
Water ($Å^2$)	35.0
Ligand ($Å^2$)	41.1
r.m.s deviations:	
Bond lengths (Å)	0.018
Bond angles (°)	1.84
Ramachandran analysis	
Favored (%)	98.84
Allowed (%)	1.16
Outliers (%)	0

Figure Supplements

Figure 2—figure supplement 1 | HaAEP1 auto-catalytically activates upon pH shift to pH 4. HiLoad 16/600 Superdex 200 size exclusion chromatograph of HaAEP1 purified at pH 8 (black) and HaAEP1 purified at pH 8 then dialyzed at pH 4 overnight at 16°C (purple). Right insert - SDS-PAGE analysis of fractions collected from labelled size exclusion peaks illustrating ~52 kDa inactive HaAEP1 (Lane 1), heterogeneous self-cleavage products of HaAEP1 activation (Lane 2), alternatively cleaved and removed C-terminal cap domains (Lane 3 and 4).

Figure 2—figure supplement 2 | Sequence alignment of C13 family of cysteine proteases. Proteins labelled with UniProt codes, AOAOG2RI59 – HaAEP1. Plant AEPs grouped in green box. Mammalian AEPs grouped in dark blue box. GPI8 proteins grouped in grey box. Identical residues shown with white text and red box, similar residues shown with red text. Potential cleavage sites of HaAEP1 shown with green stars. Catalytic triad highlighted with black stars. Succinimide forming aspartic acid, not conserved in GPI8 proteins highlighted with gold stars. Predicted N-linked glycosylation sites in HaAEP1 and human AEP shown with black and blue inverted triangles, respectively. Variable insertion sites between β_{IV}:β_V and α5:β6 strands in HaAEP1 highlighted with cyan brackets.

Figure 2—figure supplement 3. Topology diagram of active HaAEP1. Topology diagram based on previously published diagram of highly similar structure of hAEP (Dall and Brandstetter, 2013). Polyproline insertion site found in HaAEP1 between β_{IV}:β_V shown in red. Residues 304-309 which exhibit low electron density are illustrated with red dotted loop.

Figure 3—figure supplement 1 | HaAEP1 active site geometry. (**A**) Alternate conformations of catalytic residues H178 and C220, distances illustrated with text (Å) and dots. Distance of His178 to other local residues indicates that, with the exception of N73 and Snn177, H178 is relatively free from steric hindrance. (**B**) Tetrahedral peptide conformation bound to C220 (purple sticks) suggests that H178 and the backbone amines of G179 and C220 contribute to stabilizing this intermediate by forming an oxyanion hole, hydrogen bonds illustrated with text (Å) and dots. Furthermore, H178 is ideally situated to deprotonate an attacking nucleophilic group leading to formation of a tetrahedral intermediate and protonate the leaving group upon resolution of the intermediate.

Figure 3—figure supplement 2 | Polder OMIT maps define the presence of a tetrahedral intermediate in the HaAEP1 active site. Polder OMIT maps (F_{obs} - F_{calc}) calculated in the absence of an AAN peptide ligand, contoured at 3σ level, indicate that models of alternative intermediate AAN peptide structures (thioacyl intermediate – purple) and cleavage product (hydrolysis product – gray) appear not to fit the observed electron density as well as the modelled tetrahedral intermediate (second tetrahedral intermediate – green). Model peptides are shown from above (left) and side on (right) to aid visualization of electron density maps with incongruous regions of models circled (red dashed line).

Figure 4—figure supplement 1 | Analysis of WT HaAEP1 and mutants activity. (**A**) Expanded view of MALDI-TOF spectra of seleno-Cys SFTI(D14N)-GLDN processing by WT and a range of HaAEP1 mutants. Peak of mass 1608 - cyclic SFTI-1(D14N) and mass 1626 - acyclic-SFTI(D14N) indicated with blue and red dashed lines, respectively. (**B**) Structure of activity based probe with

fluorescent BODIPY structure (yellow), P1 Asp residue (blue) and acyloxymethyl ketone (AOMK) highlighted. (**C**) SDS-PAGE analysis of Pro-AEP (Mw ~52 kDa) HaAEP1 WT and mutants at 0.1 mg/ml (pH 8). (**D**) Activity based probe analysis of HaAEP1 WT and mutants at pH 6.5 following activation at pH 4 illustrates heterogeneity of active HaAEP1 mutants sizes and drastically reduced levels of activity of G185S, H178A and C220S mutants.

Figure 4—figure supplement 2 | Model of SFTI-1 processing from PawS1. Model aligned on framework of tetrahedral intermediate in the active site and where the substrate should orient through the S1-S5 pocket, between the β5-β6 and β_{IV}:β_V loops, using a PawS1 NMR structure (PDB 5U87). HaAEP1 modelled PawS1 cleavage at Asn49 (**A**), Asn18 (**B**) and Asp14 (**C**) with expanded views of active site shown below. HaAEP1 green, SFTI-1 precursor light cyan, Preproalbumin cyan. Cleavage residues highlighted with dark blue sticks. N-terminal Gly, shown as magenta sticks, acts as nucleophile to attack C-terminus of SFTI-GLDN (**C**).

Figure 4—figure supplement 3. Circular dichroism analysis of HaAEP1 secondary structure. (**A**) Melt curve of pro-AEP (pH 8) HaAEP1 measured at 222 nm, [0.2 mg/mL], shows a melting temperature of ~52°C. n=3 technical replicates data fitted with sigmoidal four parameter logistic regression fit. (**B**) Circular dichroism spectroscopy of pro-AEP (pH 8) WT HaAEP1 and active site mutants, [0.1 mg/mL], average of n=2 technical replicates.

Figure 4—figure supplement 4 | Re-ligation of cap domain at pH 8 favored after activation at pH 6.5. In-gel fluorescence of activity based probe (right) and post-fluorescence Coomassie stain of SDS-PAGE gel (left) of WT HaAEP1, N73A, E221K and D177G mutants. Activity based probe illustrates presence of active HaAEPs after dialysis at pH 4/5.5/6.5. Coomassie staining of

gels reveals that upon pH shift to pH 8 re-ligation of the cap domain is favored post-dialysis at pH 6.5, with strong pro-AEP bands also seen post-dialysis at pH 5.5 with N73A and D177G mutants.

Figure 5—figure supplement 1 | Comparison of crystal structures of AEPs and a model of an efficient macrocyclizing AEP. Crystal structures of AEPs with APBS generated electrostatic potential maps contoured at ±10 kT/e and catalytic Cys and His shown in yellow and blue sticks, respectively. Overall topology of AEPs (left) appear similar with slight differences in surface electrostatic potential. Expanded views of catalytic region (β_I sheet, β5-β_{IV} loop and β_I-β_{III} region) orientated over the catalytic His residue (right) illustrates variation around the active site, with residues that are proposed to alter substrate specificity shown with sticks. A model of an efficient macrocyclizer, *C. ensiformis* AEP (accession code: AIB06797), built on the HaAEP1 structure which displays 76.3 % amino acid identity, is shown below the dotted line. *C. ensiformis* AEP appears to exhibit a more accessible active site, in comparison to other AEP structures, due to the presence of several short side-chain amino acids.

References

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., Cotter, P.D., Craik, D.J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P.C., Entian, K.D., Fischbach, M.A., Garavelli, J.S., Goransson, U., Gruber, C.W., Haft, D.H., Hemscheidt, T.K., Hertweck, C., Hill, C., Horswill, A.R., Jaspars, M., Kelly, W.L., Klinman, J.P., Kuipers, O.P., Link, A.J., Liu, W., Marahiel, M.A., Mitchell, D.A., Moll, G.N., Moore, B.S., Muller, R., Nair, S.K., Nes, I.F., Norris, G.E., Olivera, B.M., Onaka, H., Patchett, M.L., Piel, J., Reaney, M.J., Rebuffat, S., Ross, R.P., Sahl, H.G., Schmidt, E.W., Selsted, M.E., Severinov, K., Shen, B., Sivonen, K., Smith, L., Stein, T., Sussmuth, R.D., Tagg, J.R., Tang, G.L., Truman, A.W., Vederas, J.C., Walsh, C.T., Walton, J.D., Wenzel, S.C., Willey, J.M., and van der Donk, W.A. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat. Prod. Rep. **30,** 108-160.

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

Bergmann, M., and Fruton, J.S. (1938). Some synthetic and hydrolytic experiments with chymotrypsin. J. Biol. Chem. **124,** 321-329.

Bernath-Levin, K., Nelson, C., Elliott, A.G., Jayasena, A.S., Millar, A.H., Craik, D.J., and Mylne, J.S. (2015). Peptide macrocyclization by a bifunctional endoprotease. Chem. Biol. **22,** 571-582.

Bhardwaj, G., Mulligan, V.K., Bahl, C.D., Gilmore, J.M., Harvey, P.J., Cheneval, O., Buchko, G.W., Pulavarti, S.V., Kaas, Q., Eletsky, A., Huang, P.S., Johnsen, W.A., Greisen, P.J., Rocklin, G.J., Song, Y., Linsky, T.W., Watkins, A., Rettie, S.A., Xu, X., Carter, L.P., Bonneau, R., Olson, J.M., Coutsias, E., Correnti, C.E., Szyperski, T., Craik, D.J., and Baker, D. (2016). Accurate *de novo* design of hyperstable constrained peptides. Nature **538,** 329-335.

Bollini, R., and Chrispeels, M.J. (1979). The rough endoplasmic reticulum is the site of reserve-protein synthesis in developing *Phaseolus vulgaris* cotyledons. Planta **146,** 487-501.

Bowles, D.J., Marcus, S.E., Pappin, D.J., Findlay, J.B., Eliopoulos, E., Maycox, P.R., and Burgess, J. (1986). Posttranslational processing of concanavalin A precursors in jackbean cotyledons. J. Cell Biol. **102,** 1284-1297.

Brady, K.D., Giegel, D.A., Grinnell, C., Lunney, E., Talanian, R.V., Wong, W., and Walker, N. (1999). A catalytic mechanism for caspase-1 and for bimodal inhibition of caspase-1 by activated aspartic ketones. Bioorg. Med. Chem. **7,** 621-631.

Brennan, T.V., and Clarke, S. (1995). Effect of adjacent histidine and cysteine residues on the spontaneous degradation of asparaginyl- and aspartyl-containing peptides. Int. J. Pept. Protein Res. **45,** 547-553.

Buller, A.R., and Townsend, C.A. (2013). Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. Proc. Natl. Acad. Sci. U. S. A. **110,** E653-661.

Chekan, J.R., Estrada, P., Covello, P.S., and Nair, S.K. (2017). Characterization of the macrocyclase involved in the biosynthesis of RiPP cyclic peptides in plants. Proc. Natl. Acad. Sci. U. S. A. **114,** 6551-6556.

Chen, J.M., Dando, P.M., Rawlings, N.D., Brown, M.A., Young, N.E., Stevens, R.A., Hewitt, E., Watts, C., and Barrett, A.J. (1997). Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase. J. Biol. Chem. **272,** 8090-8098.

Chen, V.B., Arendall, W.B., III, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D Biol. Crystallogr. **66,** 12-21.

Chu, M.H., Liu, K.L., Wu, H.Y., Yeh, K.W., and Cheng, Y.S. (2011). Crystal structure of tarocystatin-papain complex: implications for the inhibition property of group-2 phytocystatins. Planta **234,** 243-254.

Clancy, K.W., Melvin, J.A., and McCafferty, D.G. (2010). Sortase transpeptidases: insights into mechanism, substrate specificity, and inhibition. Biopolymers **94,** 385-396.

Clark, A.C. (2016). Caspase Allostery and Conformational Selection. Chem. Rev. **116,** 6666-6706.

Clark, R.J., Jensen, J., Nevin, S.T., Callaghan, B.P., Adams, D.J., and Craik, D.J. (2010). The engineering of an orally active conotoxin for the treatment of neuropathic pain. Angew. Chem. Int. Ed. **49,** 6545-6548.

Clark, R.J., Fischer, H., Dempster, L., Daly, N.L., Rosengren, K.J., Nevin, S.T., Meunier, F.A., Adams, D.J., and Craik, D.J. (2005). Engineering stable peptide toxins by means of backbone cyclization: stabilization of the alpha-conotoxin MII. Proc. Natl. Acad. Sci. U. S. A. **102,** 13767-13772.

Dall, E., and Brandstetter, H. (2013). Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. Proc. Natl. Acad. Sci. U. S. A. **110,** 10940-10945.

Dall, E., and Brandstetter, H. (2016). Structure and function of legumain in health and disease. Biochimie **122,** 126-150.

Dall, E., Fegg, J.C., Briza, P., and Brandstetter, H. (2015). Structure and mechanism of an aspartimide-dependent peptide ligase in human legumain. Angew. Chem. Int. Ed. **54,** 2917-2921.

Dodson, G., and Wlodawer, A. (1998). Catalytic triads and their relatives. Trends Biochem. Sci. **23,** 347-352.

Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G., and Baker, N.A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res. **35,** W522-525.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallographica Section D **60,** 2126-2132.

Fagundes, D., Bohn, B., Cabreira, C., Leipelt, F., Dias, N., Bodanese-Zanettini, M.H., and Cagliari, A. (2015). Caspases in plants: metacaspase gene family in plant stress responses. Funct. Integr. Genomics **15,** 639-649.

Franke, B., Colgrave, M.L., Mylne, J.S., and Rosengren, K.J. (2016). Mature forms of the major seed storage albumins in sunflower: A mass spectrometric approach. J. Proteomics **147,** 177-186.

Franke, B., James, A.M., Mobli, M., Colgrave, M.L., Mylne, J.S., and Rosengren, K.J. (2017). Two proteins for the price of one: structural studies of the dual destiny Preproalbumin with Sunflower Trypsin Inhibitor-1. J. Biol. Chem. **292,** 12398-12411.

Frankel, B.A., Kruger, R.G., Robinson, D.E., Kelleher, N.L., and McCafferty, D.G. (2005). *Staphylococcus aureus* sortase transpeptidase SrtA: insight into the kinetic mechanism and evidence for a reverse protonation catalytic mechanism. Biochemistry **44,** 11188-11200.

Frankel, B.A., Tong, Y., Bentley, M.L., Fitzgerald, M.C., and McCafferty, D.G. (2007). Mutational analysis of active site residues in the *Staphylococcus aureus* transpeptidase SrtA. Biochemistry **46,** 7269-7278.

Fuentes-Prior, P., and Salvesen, G. (2004). The protein structures that shape caspase activity, specificity, activation and inhibition. Biochem. J. **384,** 201-232.

Geiger, T., and Clarke, S. (1987). Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. J. Biol. Chem. **262,** 785-794.

Gould, A., Ji, Y., Aboye, T.L., and Camarero, J.A. (2011). Cyclotides, a novel ultrastable polypeptide scaffold for drug discovery. Curr. Pharm. Des. **17,** 4294-4307.

Hara-Nishimura, I., Takeuchi, Y., and Nishimura, M. (1993). Molecular characterization of a vacuolar processing enzyme related to a putative cysteine proteinase of *Schistosoma mansoni*. Plant Cell **5,** 1651-1659.

Hara-Nishimura, I., Kinoshita, T., Hiraiwa, N., and Nishimura, M. (1998). Vacuolar processing enzymes in protein-storage vacuoles and lytic vacuoles. J. Plant Physiol. **152,** 668-674.

Harris, K.S., Durek, T., Kaas, Q., Poth, A.G., Gilding, E.K., Conlan, B.F., Saska, I., Daly, N.L., van der Weerden, N.L., Craik, D.J., and Anderson, M.A. (2015). Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. Nat. Commun. **6,** 10199.

Hatsugai, N., Kuroyanagi, M., Yamada, K., Meshi, T., Tsuda, S., Kondo, M., Nishimura, M., and Hara-Nishimura, I. (2004). A plant vacuolar protease, VPE, mediates virus-induced hypersensitive cell death. Science **305,** 855-858.

Hetrick, K.J., and van der Donk, W.A. (2017). Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. Curr. Opin. Chem. Biol. **38,** 36-44.

Hiraiwa, N., Nishimura, M., and Hara-Nishimura, I. (1999). Vacuolar processing enzyme is self-catalytically activated by sequential removal of the C-terminal and N-terminal propeptides. FEBS Lett. **447,** 213-216.

Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. **38,** W545-W549.

James, A.M., Haywood, J., and Mylne, J.S. (2017). Macrocyclization by asparaginyl endopeptidases. New Phytol. **10.1111/nph.14511**.

Jayasena, A.S., Franke, B., Rosengren, K.J., and Mylne, J.S. (2016). A tripartite 'omic analysis identifies the major sunflower seed albumins. Theor. Appl. Genet. **129,** 613-629.

Ji, Y., Majumder, S., Millard, M., Borra, R., Bi, T., Elnagar, A.Y., Neamati, N., Shekhtman, A., and Camarero, J.A. (2013). *In vivo* activation of the p53 tumor suppressor pathway by an engineered cyclotide. J. Am. Chem. Soc. **135,** 11623-11633.

Kuroyanagi, M., Yamada, K., Hatsugai, N., Kondo, M., Nishimura, M., and Hara-Nishimura, I. (2005). Vacuolar processing enzyme is essential for mycotoxin-induced cell death in *Arabidopsis thaliana*. J. Biol. Chem. **280,** 32914-32920.

Laskowski, R.A., and Swindells, M.B. (2011). LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model **51,** 2778-2786.

Lee, T.-W., and James, M.N.G. (2008). 1.2Å-resolution crystal structures reveal the second tetrahedral intermediates of streptogrisin B (SGPB). Biochim Biophys Acta **1784,** 319-334.

Liebschner, D., Afonine, P.V., Moriarty, N.W., Poon, B.K., Sobolev, O.V., Terwilliger, T.C., and Adams, P.D. (2017). Polder maps: improving OMIT maps by excluding bulk solvent. Acta Crystallogr D Struct Biol **73,** 148-157.

Lu, H., Chandrasekar, B., Oeljeklaus, J., Misas-Villamil, J.C., Wang, Z., Shindo, T., Bogyo, M., Kaiser, M., and van der Hoorn, R.A. (2015). Subfamily-specific fluorescent probes for cysteine proteases display dynamic protease activities during seed germination. Plant Physiol. **168,** 1462-1475.

Luckett, S., Garcia, R.S., Barker, J.J., Konarev, A.V., Shewry, P.R., Clarke, A.R., and Brady, R.L. (1999). High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. J. Mol. Biol. **290,** 525-533.

Manoury, B., Hewitt, E.W., Morrice, N., Dando, P.M., Barrett, A.J., and Watts, C. (1998). An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation. Nature **396,** 695-699.

Martinez, M., Cambra, I., Gonzalez-Melendi, P., Santamaria, M.E., and Diaz, I. (2012). C1A cysteine-proteases and their inhibitors in plants. Physiol. Plant **145,** 85-94.

Mathieu, M.A., Bogyo, M., Caffrey, C.R., Choe, Y., Lee, J., Chapman, H., Sajid, M., Craik, C.S., and McKerrow, J.H. (2002). Substrate specificity of schistosome versus human legumain determined by P1-P3 peptide libraries. Mol. Biochem. Parasitol. **121,** 99-105.

Mazmanian, S.K., Liu, G., Ton-That, H., and Schneewind, O. (1999). *Staphylococcus aureus* sortase, an enzyme that anchors surface proteins to the cell wall. Science **285,** 760-763.

McLuskey, K., Rudolf, J., Proto, W.R., Isaacs, N.W., Coombs, G.H., Moss, C.X., and Mottram, J.C. (2012). Crystal structure of a *Trypanosoma brucei* metacaspase. Proc. Natl. Acad. Sci. U. S. A. **109,** 7469-7474.

Min, W., and Jones, D.H. (1994). *In vitro* splicing of concanavalin A is catalyzed by asparaginyl endopeptidase. Nat. Struct. Mol. Biol. **1,** 502-504.

Mylne, J.S., Colgrave, M.L., Daly, N.L., Chanson, A.H., Elliott, A.G., McCallum, E.J., Jones, A., and Craik, D.J. (2011). Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. Nat. Chem. Biol. **7,** 257-259.

Mylne, J.S., Chan, L.Y., Chanson, A.H., Daly, N.L., Schaefer, H., Bailey, T.L., Nguyencong, P., Cascales, L., and Craik, D.J. (2012). Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis. Plant Cell **24,** 2765-2778.

Nguyen, G.K.T., Wang, S., Qiu, Y., Hemu, X., Lian, Y., and Tam, J.P. (2014). Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. Nat. Chem. Biol. **10,** 732-738.

Otto, H.H., and Schirmeister, T. (1997). Cysteine proteases and their inhibitors. Chem. Rev. **97,** 133-172.

Pattabiraman, V.R., and Bode, J.W. (2011). Rethinking amide bond synthesis. Nature **480,** 471-479.
Poth, A.G., Chan, L.Y., and Craik, D.J. (2013). Cyclotides as grafting frameworks for protein engineering and drug design applications. Biopolymers **100,** 480-491.
Radisky, E.S., Lee, J.M., Lu, C.J.K., and Koshland, D.E. (2006). Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. Proc. Natl. Acad. Sci. U. S. A. **103,** 6835-6840.
Rawlings, N.D., Barrett, A.J., and Finn, R. (2016). Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. **44,** D343-D350.
Reissner, K.J., and Aswad, D.W. (2003). Deamidation and isoaspartate formation in proteins: unwanted alterations or surreptitious signals? Cell Mol Life Sci **60,** 1281-1295.
Ritz-Timme, S., and Collins, M.J. (2002). Racemization of aspartic acid in human proteins. Ageing Res. Rev. **1,** 43-59.
Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc **5,** 725-738.
Schechter, I., and Berger, A. (1967). On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. **27,** 157-162.
Schrodinger, L. (2010). The PyMOL Molecular Graphics System, Version 1.4.
Shafee, T., Harris, K., and Anderson, M. (2015). Biosynthesis of cyclotides. In Advances in Botanical Research (Academic Press), pp. 227-269.
Sheldon, P.S., Keen, J.N., and Bowles, D.J. (1996). Post-translational peptide bond formation during concanavalin A processing *in vitro*. Biochem. J. **320,** 865-870.
Shewry, P.R., and Halford, N.G. (2002). Cereal seed storage proteins: structures, properties and role in grain utilization. J. Exp. Bot. **53,** 947-958.
Sojka, D., Hajdusek, O., Dvorak, J., Sajid, M., Franta, Z., Schneider, E.L., Craik, C.S., Vancova, M., Buresova, V., Bogyo, M., Sexton, K.B., McKerrow, J.H., Caffrey, C.R., and Kopacek, P. (2007). IrAE: an asparaginyl endopeptidase (legumain) in the gut of the hard tick *Ixodes ricinus*. Int. J. Parasitol. **37,** 713-724.
Suree, N., Liew, C.K., Villareal, V.A., Thieu, W., Fadeev, E.A., Clemens, J.J., Jung, M.E., and Clubb, R.T. (2009). The structure of the *Staphylococcus aureus* sortase-substrate complex reveals how the universally conserved LPXTG sorting signal is recognized. J. Biol. Chem. **284,** 24465-24477.
Swedberg, J.E., Li, C.Y., de Veer, S.J., Wang, C.K., and Craik, D.J. (2017). Design of potent and selective cathepsin G inhibitors based on the sunflower trypsin inhibitor-1 scaffold. J. Med. Chem. **60,** 658-667.
Takahashi, O., Manabe, N., and Kirikoshi, R. (2016). A computational study of the mechanism of succinimide formation in the Asn-His sequence: intramolecular catalysis by the His side chain. Molecules **21,** 327.
Truman, A.W. (2016). Cyclisation mechanisms in the biosynthesis of ribosomally synthesised and post-translationally modified peptides. Beilstein Journal of Organic Chemistry **12,** 1250-1268.
Tsiatsiani, L., Van Breusegem, F., Gallois, P., Zavialov, A., Lam, E., and Bozhkov, P.V. (2011). Metacaspases. Cell Death Differ. **18,** 1279-1288.

Vercammen, D., van de Cotte, B., De Jaeger, G., Eeckhout, D., Casteels, P., Vandepoele, K., Vandenberghe, I., Van Beeumen, J., Inze, D., and Van Breusegem, F. (2004). Type II Metacaspases Atmc4 and Atmc9 of *Arabidopsis thaliana* cleave substrates after arginine and lysine. J. Biol. Chem. **279,** 45329-45336.

Vernet, T., Tessier, D.C., Chatellier, J., Plouffe, C., Lee, T.S., Thomas, D.Y., Storer, A.C., and Menard, R. (1995). Structural and functional roles of asparagine 175 in the cysteine protease papain. J. Biol. Chem. **270,** 16645-16652.

White, A.M., and Craik, D.J. (2016). Discovery and optimization of peptide macrocycles. Expert Opin. Drug Discov. **11,** 1151-1163.

Wilmouth, R.C., Edman, K., Neutze, R., Wright, P.A., Clifton, I.J., Schneider, T.R., Schofield, C.J., and Hajdu, J. (2001). X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate. Nat. Struct. Biol. **8,** 689-694.

Winn, M. D. , Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., S. Pannu, N., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., and Wilson, K.S. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr **67,** 235-242.

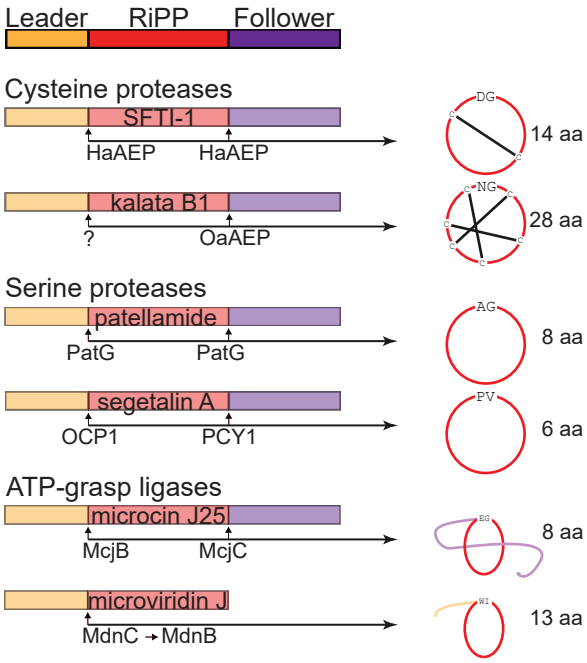
Yamada, K., Shimada, T., Nishimura, M., and Hara-Nishimura, I. (2005). A VPE family supporting various vacuolar functions in plants. Physiol. Plant **123,** 369-375.

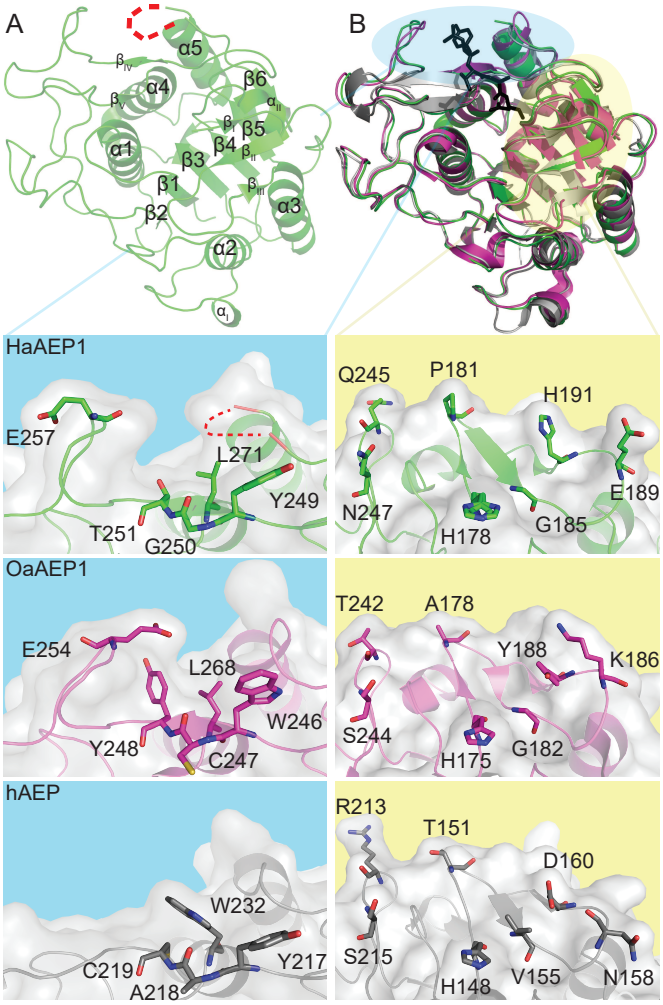
Yang, R., Wong, Y.H., Nguyen, G.K.T., Tam, J.P., Lescar, J., and Wu, B. (2017). Engineering a catalytically efficient recombinant protein ligase. J. Am. Chem. Soc. **139,** 5351-5358.

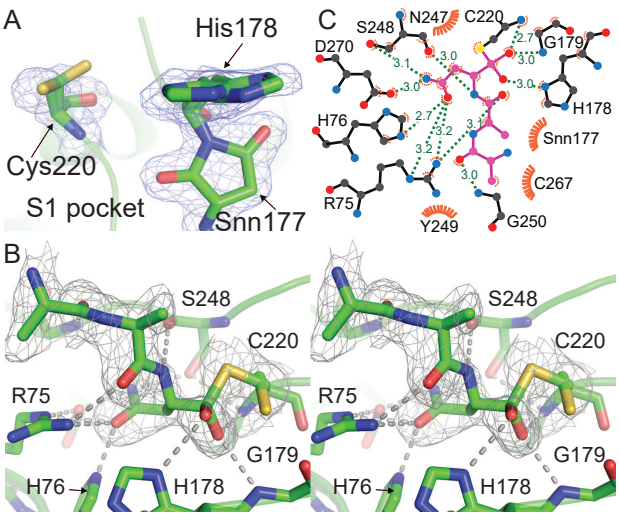
Youle, R.J., and Huang, A.H.C. (1978). Albumin storage proteins in the protein bodies of castor bean. Plant Physiol. **61,** 13.

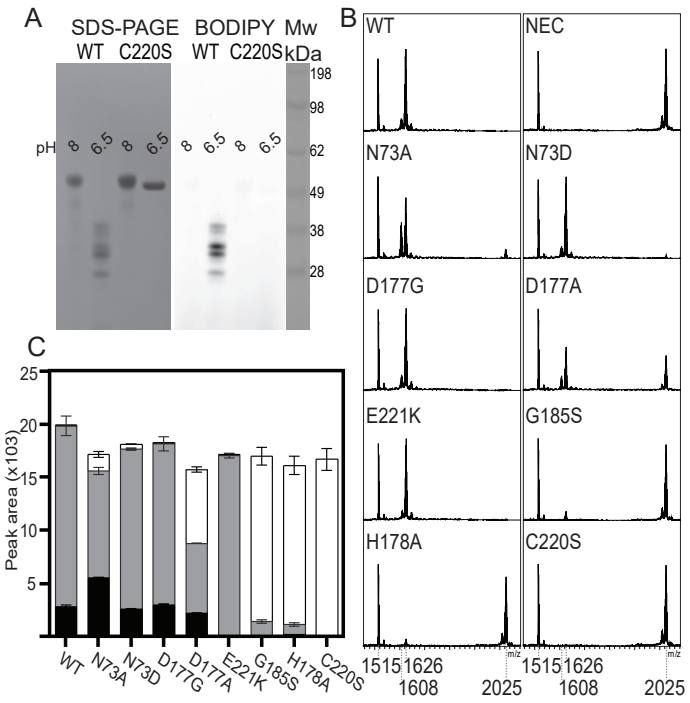
Zacks, M.A., and Garg, N. (2006). Recent developments in the molecular, biochemical and functional characterization of GPI8 and the GPI-anchoring mechanism Mol. Membr. Biol. **23,** 209-225.

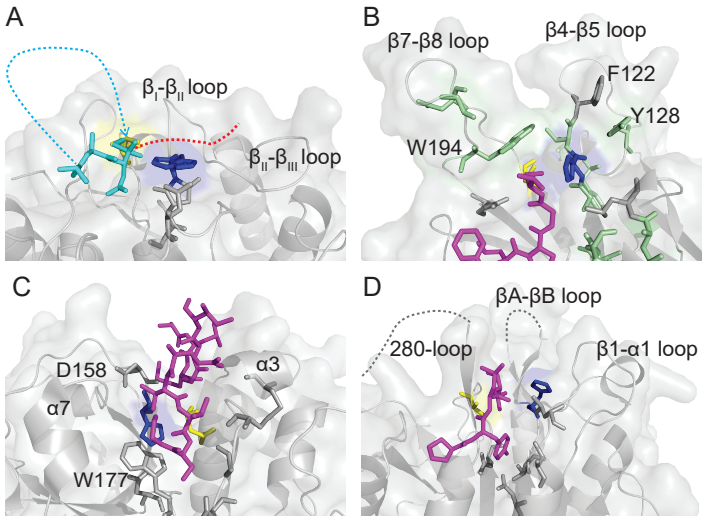
Zhao, L., Hua, T., Crowley, C., Ru, H., Ni, X., Shaw, N., Jiao, L., Ding, W., Qu, L., Hung, L.-W., Huang, W., Liu, L., Ye, K., Ouyang, S., Cheng, G., and Liu, Z.-J. (2014). Structural analysis of asparaginyl endopeptidase reveals the activation mechanism and a reversible intermediate maturation stage. Cell Res. **24,** 344-358.

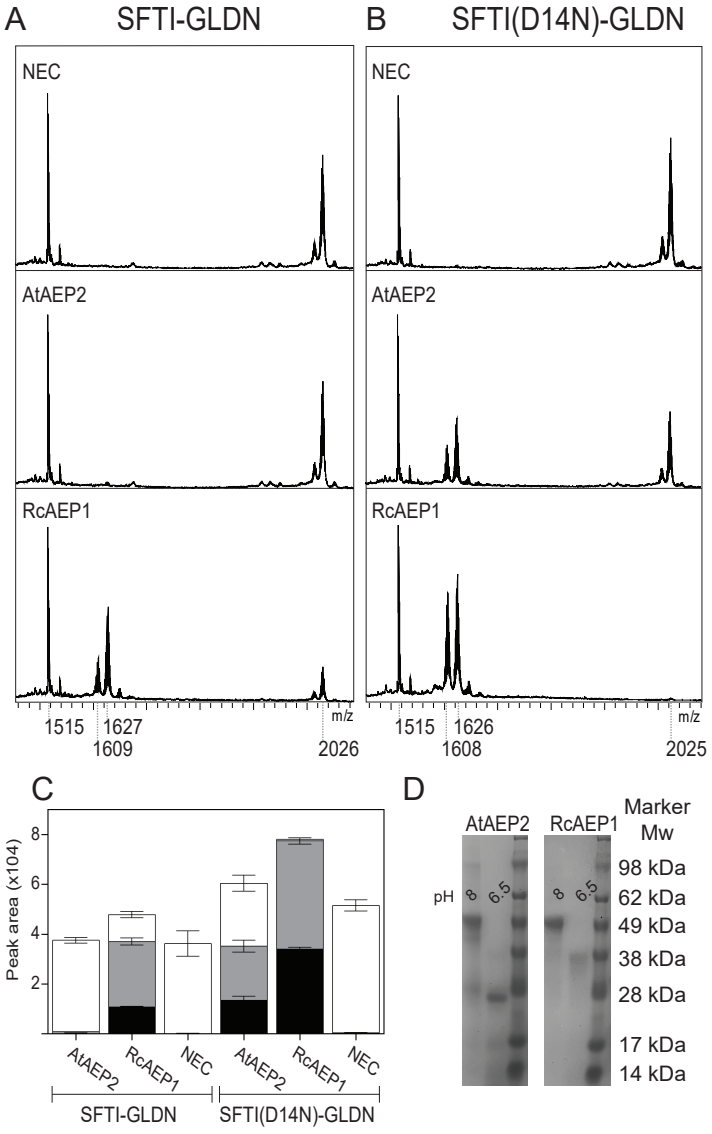

Leader
RiPP
Follower
Cysteine proteases
SFTI-1
HaAEP
HaAEP
DG
14 aa
kalata B1
?
OaAEP
NG
28 aa
Serine proteases
patellamide
PatG
PatG
AG
8 aa
segetalin A
OCP1
PCY1
PV
6 aa
ATP-grasp ligases
microcin J25
McjB
McjC
8 aa
microviridin J
MdnC → MdnB
13 aa

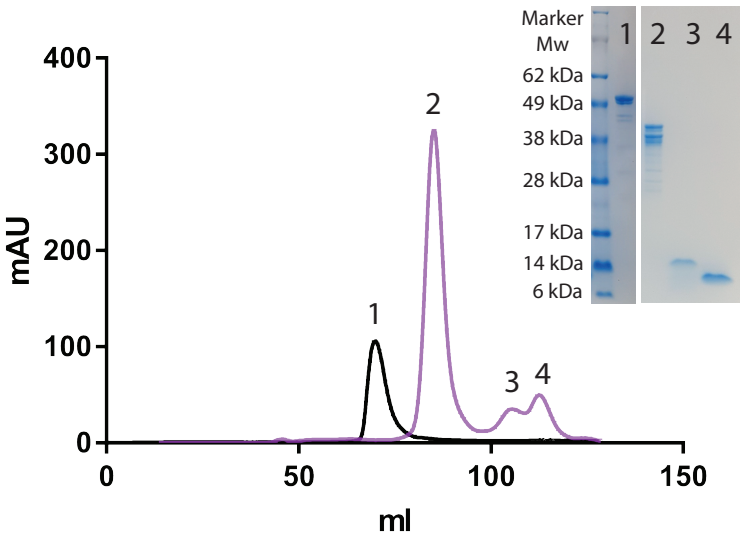

A
B
α5
β6
α4
β5
α1
β4
β3
β1
β2
α3
α2
HaAEP1
E257
L271
Y249
T251
G250
Q245
P181
H191
E189
N247
H178
G185
OaAEP1
E254
L268
W246
Y248
C247
T242
A178
Y188
K186
S244
H175
G182
hAEP
W232
C219
Y217
A218
R213
T151
D160
S215
H148
V155
N158

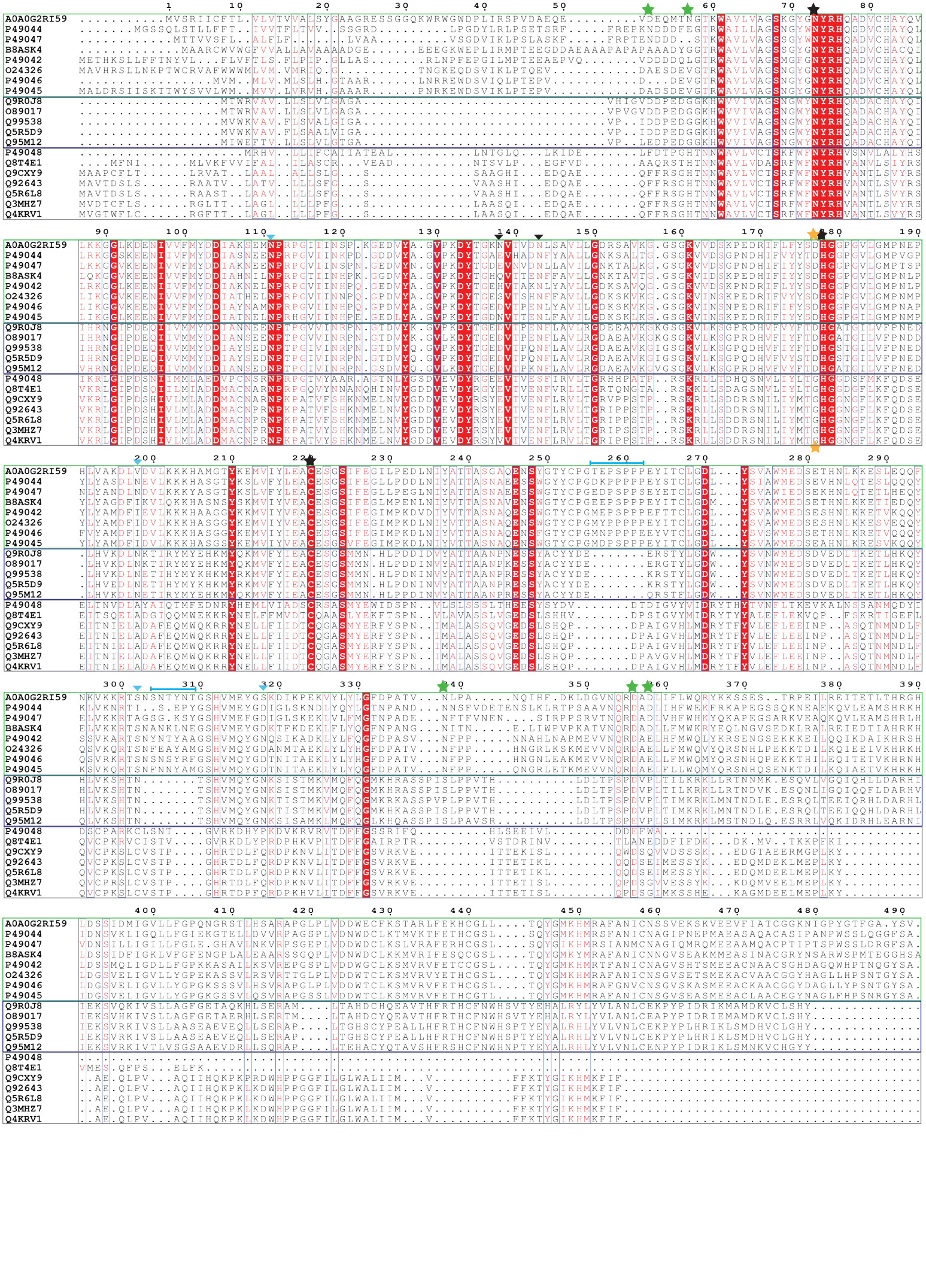

A
His178
Cys220
S1 pocket
Snn177
C
S248
N247
C220
D270
G179
H76
H178
Snn177
C267
R75
Y249
G250
B
S248
C220
R75
G179
H76
H178
S248
C220
R75
G179
H76
H178

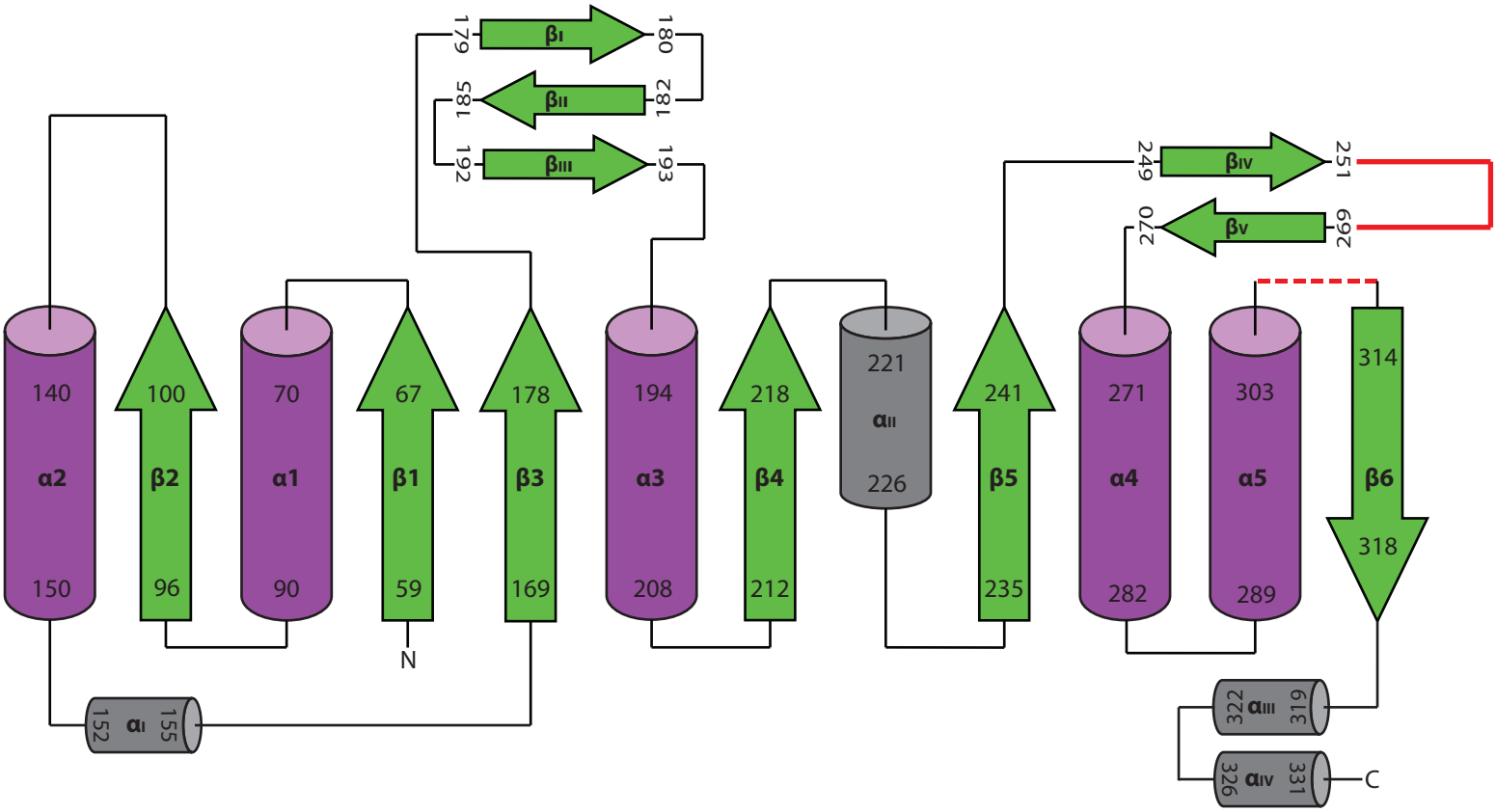

A
SDS-PAGE
WT C220S
BODIPY
WT C220S
Mw
kDa
198
98
62
49
38
28
pH 8 6.5 8 6.5 8 6.5 8 6.5
B
WT
NEC
N73A
N73D
D177G
D177A
E221K
G185S
H178A
C220S
1515 1626
1608
2025
m/z
C
Peak area (x103)
25
20
15
10
5
WT
N73A
N73D
D177G
D177A
E221K
G185S
H178A
C220S

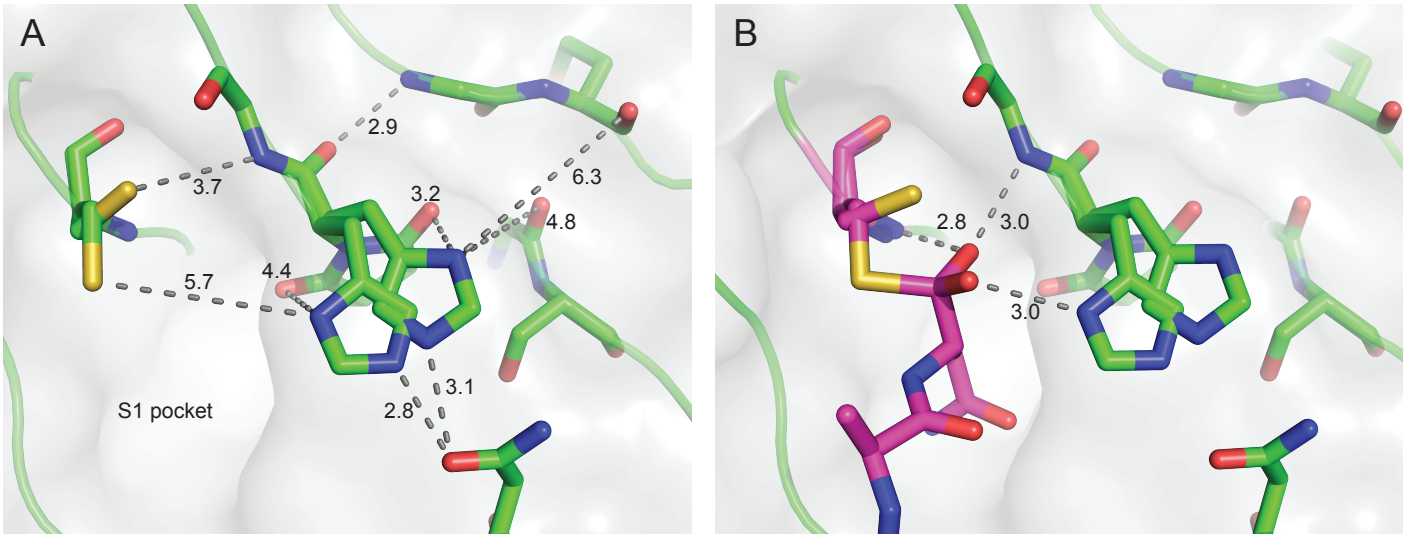

A
β_I-β_{II} loop
β_{II}-β_{III} loop
B
β7-β8 loop
β4-β5 loop
F122
Y128
W194
C
D158
α3
α7
W177
D
βA-βB loop
280-loop
β1-α1 loop


A
SFTI-GLDN
B
SFTI(D14N)-GLDN
NEC
AtAEP2
RcAEP1
m/z
1515
1627
1609
2026
1515
1626
1608
2025
C
Peak area (x104)
8
6
4
2
AtAEP2
RcAEP1
NEC
AtAEP2
RcAEP1
NEC
SFTI-GLDN
SFTI(D14N)-GLDN
D
AtAEP2
RcAEP1
Marker Mw
pH
8
6.5
8
6.5
98 kDa
62 kDa
49 kDa
38 kDa
28 kDa
17 kDa
14 kDa

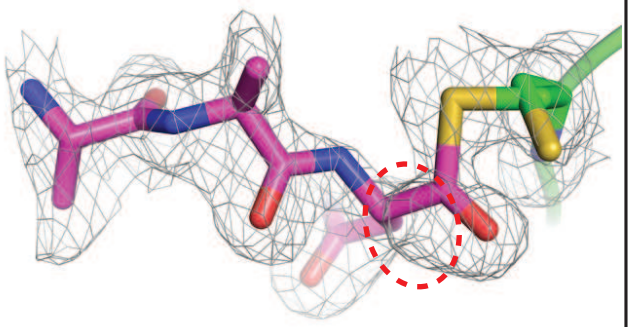

400
300
200
100
0
mAU
0
50
100
150
ml
1
2
3
4
Marker
Mw
62 kDa
49 kDa
38 kDa
28 kDa
17 kDa
14 kDa
6 kDa
1
2
3
4



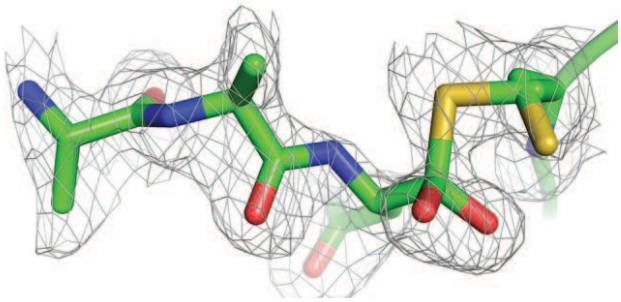

179
βI
180
181
βII
185
192
βIII
193
249
βIV
251
270
βV
269
140
α2
150
100
β2
96
70
α1
90
67
β1
59
N
178
β3
169
194
α3
208
218
β4
212
221
αII
226
241
β5
235
271
α4
282
303
α5
289
314
β6
318
152
αI
155
323
αIII
319
326
αIV
331
C


A
2.9
3.7
3.2
6.3
4.8
4.4
5.7
3.1
2.8
S1 pocket
B
2.8
3.0
3.0

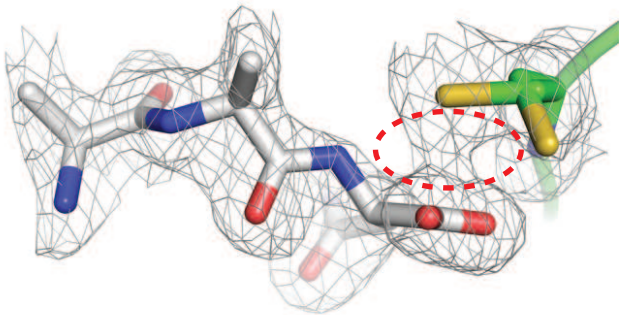
Thioacyl intermediate



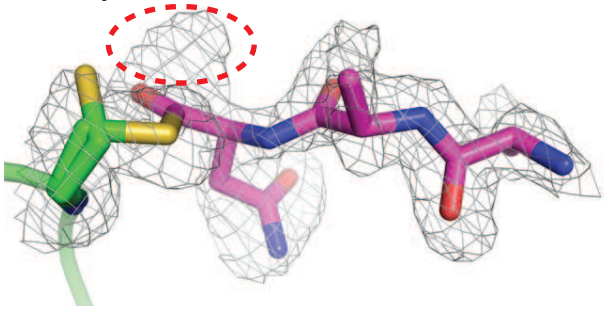
Second tetrahedral intermediate



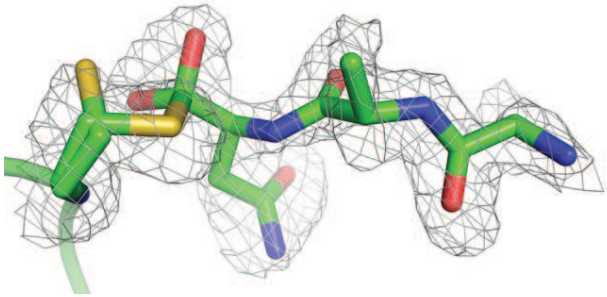
Hydrolysis product



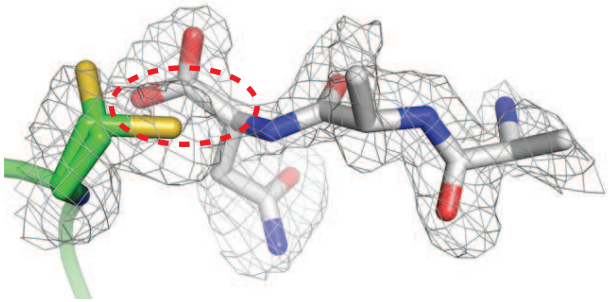
Thioacyl intermediate

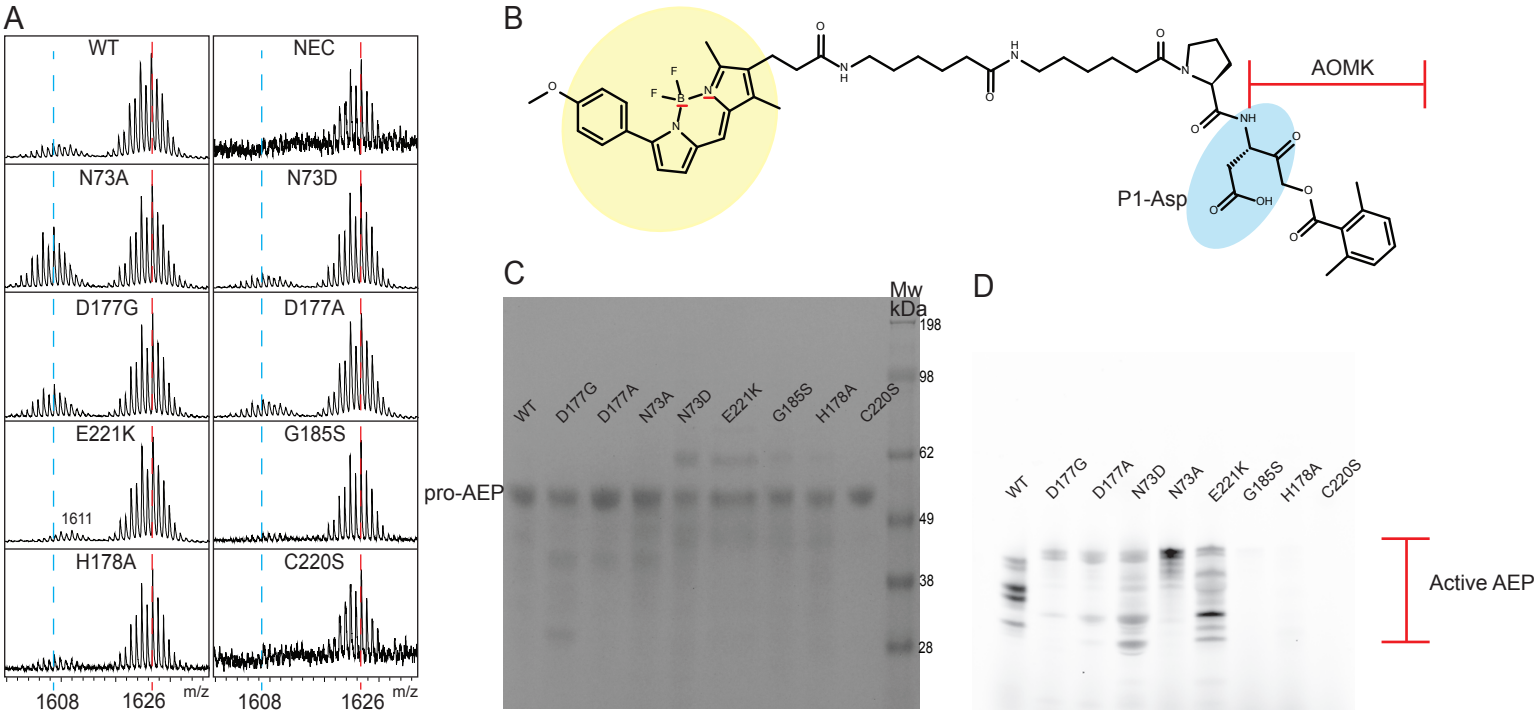


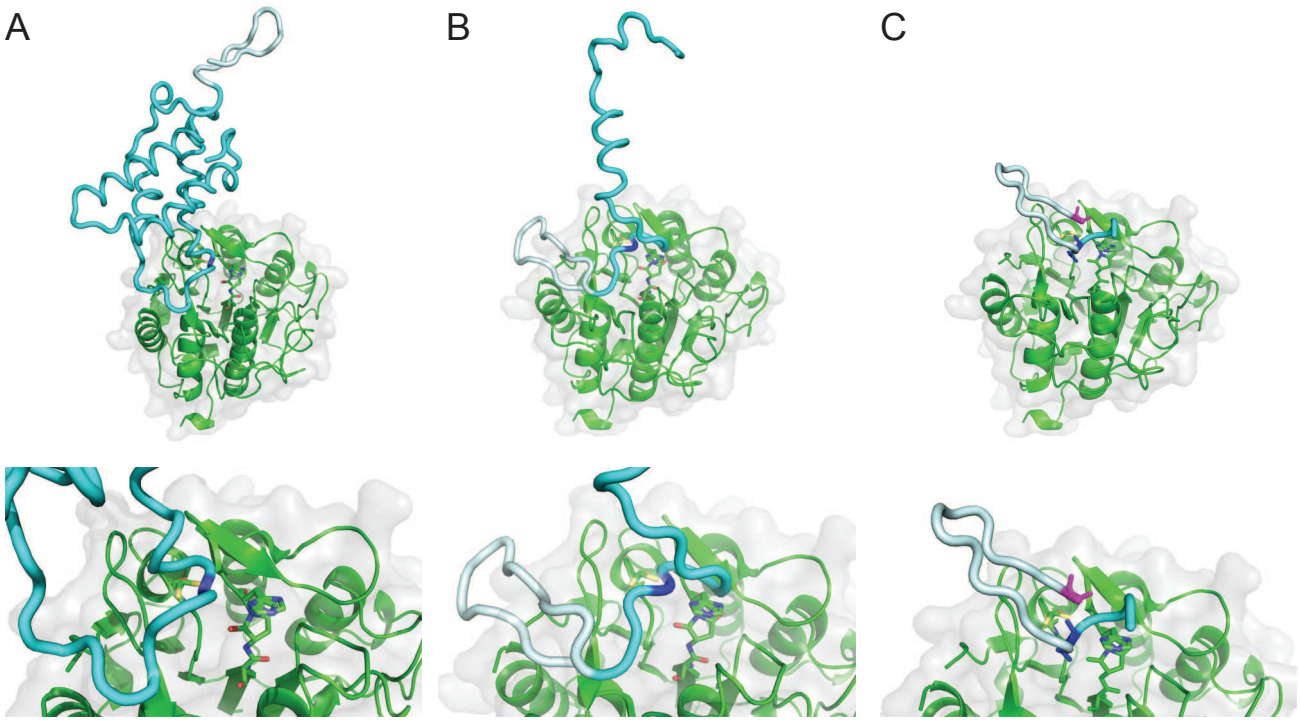
Second tetrahedral intermediate

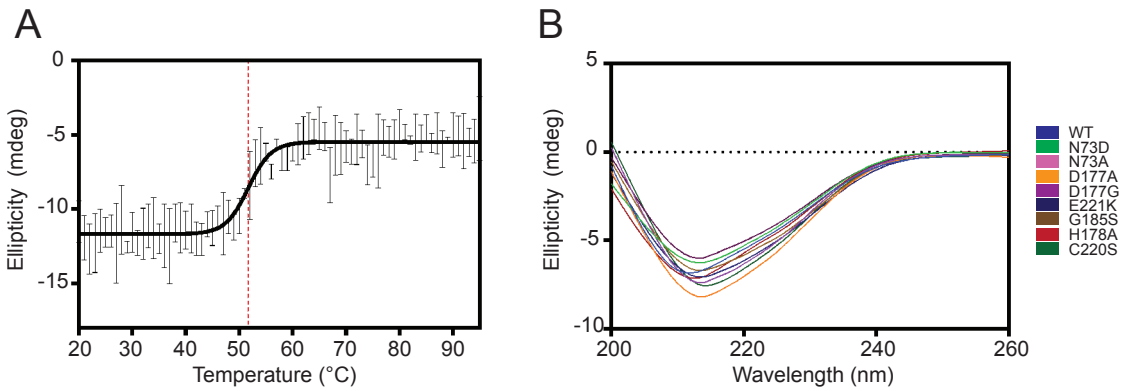


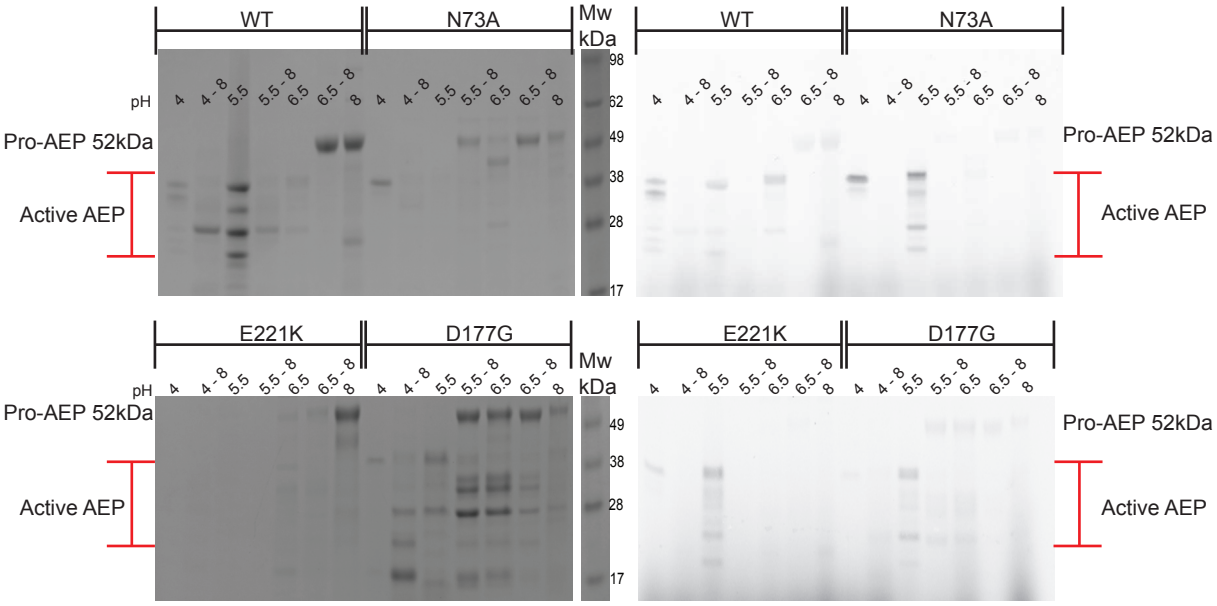
Hydrolysis product



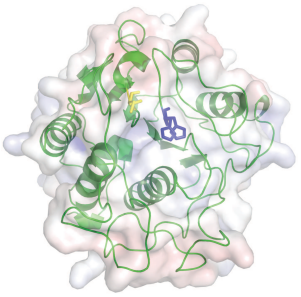

A
WT
NEC
N73A
N73D
D177G
D177A
E221K
G185S
1611
H178A
C220S
1608
1626
m/z
1608
1626
m/z
B
AOMK
P1-Asp
C
Mw
kDa
198
98
62
49
38
28
WT
D177G
D177A
N73A
N73D
E221K
G185S
H178A
C220S
pro-AEP
D
WT
D177G
D177A
N73D
N73A
E221K
G185S
H178A
C220S
Active AEP

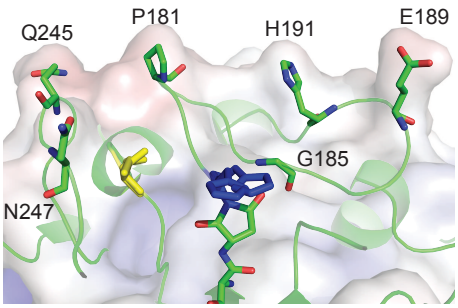

A
B
C


A
0
-5
-10
-15
Ellipticity (mdeg)
20
30
40
50
60
70
80
90
Temperature (°C)
B
5
0
-5
-10
Ellipticity (mdeg)
200
220
240
260
Wavelength (nm)
WT
N73D
N73A
D177A
D177G
E221K
G185S
H178A
C220S

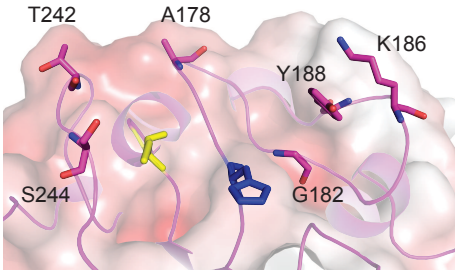

WT
N73A
Mw
kDa
WT
N73A
pH
4
4 - 8
5.5
5.5 - 8
6.5
6.5 - 8
8
98
62
49
38
28
17
Pro-AEP 52kDa
Active AEP
Pro-AEP 52kDa
Active AEP
E221K
D177G
Mw
kDa
E221K
D177G
pH
49
38
28
17
Pro-AEP 52kDa
Active AEP
Pro-AEP 52kDa
Active AEP

H. annuus

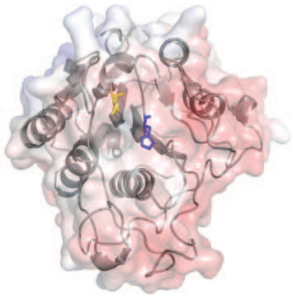


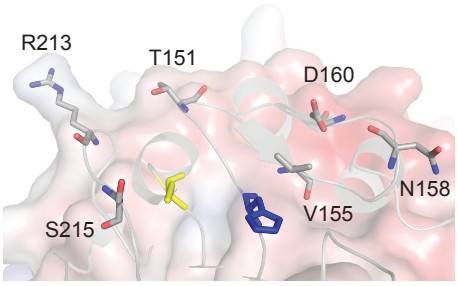


O. affinis

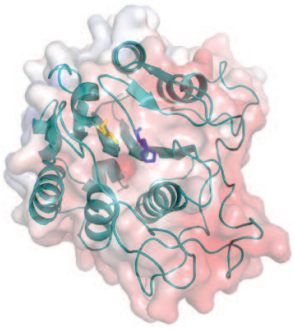


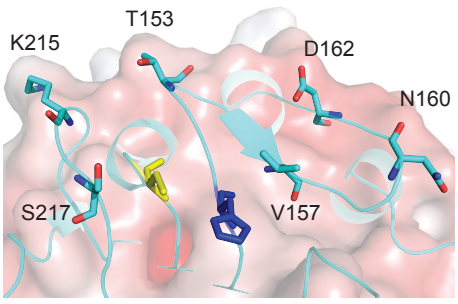
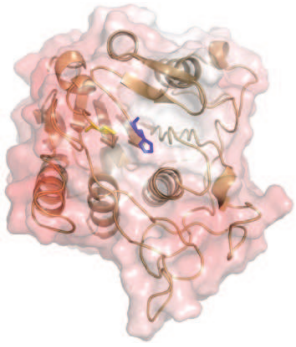
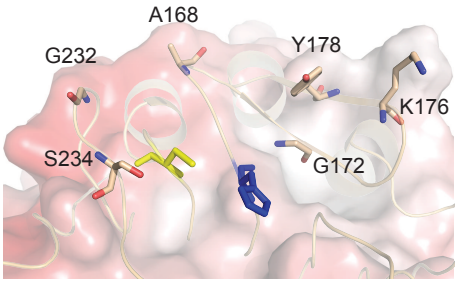
H. sapiens





M. musculus





C. ternatea