Structural and functional studies of *Arabidopsis thaliana* legumain beta reveal isoform specific mechanisms of activation and substrate recognition

Authors:
Elfriede Dall[†*#], Florian B. Zauner[†*], Wai Tuck Soh[†], Fatih Demir[§], Sven O. Dahms[†], Chiara Cabrele[†], Pitter F. Huesgen[§‖⊥], Hans Brandstetter[†#].

Affiliation:
[†]Department of Biosciences, University of Salzburg, 5020 Salzburg, Austria.
[§]Central Institute for Engineering, Electronics and Analytics, ZEA-3, Forschungszentrum Jülich, 52428 Jülich, Germany.
[‖]CECAD, Medical Faculty and University Hospital, University of Cologne, 50931 Cologne, Germany
[⊥]Institute for Biochemistry, Faculty of Mathematics and Natural Sciences, University of Cologne, 50674 Cologne, Germany
*These authors contributed equally to this work.

[#]Corresponding authors:
Elfriede Dall: elfriede.dall@sbg.ac.at
Hans Brandstetter: hans.brandstetter@sbg.ac.at

Running title: Structure and Function of AtLEGβ

Keywords: cysteine protease, crystal structure, pH regulation, transpeptidation, structural biology, plant biochemistry, protein stability

Abstract

The vacuolar cysteine protease legumain plays important functions in seed maturation and plant programmed cell death. Because of their dual protease and ligase activity, plant legumains have become of particular biotechnological interest e.g. for the synthesis of cyclic peptides for drug design or for protein engineering. However, the molecular mechanisms behind their dual protease and ligase activities are still poorly understood, limiting their applications. Here we present the crystal structure of *Arabidopsis thaliana* legumain isoform β (AtLEGβ) in its zymogen state. Combining structural and biochemical experiments, we show for the first time that plant legumains encode distinct, isoform-specific activation mechanisms. While the autocatalytic activation of isoform γ (AtLEGγ) is controlled by the latency-conferring dimer state, the activation of the monomeric AtLEGβ is concentration independent. Additionally, in AtLEGβ the plant-characteristic two-chain intermediate state is stabilized by hydrophobic rather than ionic interactions as in AtLEGγ, resulting in significantly different pH-stability profiles. The crystal structure of AtLEGβ reveiled unrestricted non-prime substrate binding pockets, consistent with the broad substrate specificity as determined by degradomic assays. Further to its protease activity, we show that AtLEGβ exhibits a true peptide ligase activity. While cleavage-dependent transpeptidase activity has been reported for other plant legumains, AtLEGβ is the first example of a plant legumain capable of linking free termini. The discovery of these isoform specific differences will allow to identify and rationally design efficient ligases with application in biotechnology and drug development.

Introduction

The plant cysteine proteases of the legumain family (C13 family, EC 3.4.22.34) have an important role in processing and maturation of seed storage proteins within the vacuole and are therefore also referred to as vacuolar processing enzymes (VPEs) (1). Plant legumains are structurally related to the mammalian caspases and exhibit a strong substrate sequence preference for cleavage after asparagine and, to a lesser extent, aspartate residues (2,3). Therefore, they are also synonymously referred to as the asparaginyl endopeptidases (AEP). In contrast to mammals, where only one functional legumain isoform is expressed, *Arabidopsis thaliana* contains four genes coding for legumains (α, β, γ, δ-VPE) and other plants even up to eight functional variants (4). Plant legumains are expressed primarily in seeds and vegetative organs, consistent with their phylogenetic grouping into two angiosperm clades, the seed type (β-VPE) and non-seed or vegetative type VPEs (α-, γ- and δ-VPE) (5-8). Vegetative legumains are found in lytic vacuoles and have been suggested to play critical roles in plant programmed cell death (PCD) and may functionally substitute the caspases, which are absent in plants (9). Seed type legumains like *Arabidopsis thaliana* legumain isoform β (AtLEGβ) play important functions in the processing and maturation of seed storage proteins within storage vacuoles (10,11). The importance of legumains is especially illustrated in Arabidopsis mutant strains missing all four legumain genes (α, β, γ, δ), which were shown to accumulate aberrantly processed seed storage proteins (12). Importantly, AtLEGβ can compensate missing vegetative α and γ proteins, further confirming that AtLEGβ is the main player in precursor protein processing in seeds (10). Known substrates of AtLEGβ include the pro12S globulin and pro2S albumin proteins (5,10,12,13).

On top of that, several plant legumains possess peptide ligase and cyclase activity (14-20). Recently we could show that the vegetative type AtLEGγ harbors ligase activity (21). However, it is still unknown whether this is also true for the other three *A. thaliana* legumain isoforms, especially the phylogenetically more distant seed type AtLEGβ.

Cyclic peptides are important for plants' defense against pathogens (16,17,22,23). Well characterized examples include the kalata B1 peptide found in *Oldenlandia affinis* which has proven antimicrobial and insecticidal activities and the Sunflower trypsin inhibitor 1 (SFTI) (22,24). Cyclic peptides are very resistant to extremes in pH and temperature, making them ideal scaffolds for biotechnological applications and drug design (25-27). Peptide cyclisation in plants is typically catalyzed by legumains. Consequently, there is a high interest in understanding the ligation mechanism, specificity and efficacy of different plant legumain isoforms. Recent studies led to the discovery of a marker of ligase activity (MLA) and

a gatekeeper residue (Cys247, *Oldenlandia affinis* numbering) that allow to predict ligase activity based on sequence information (20,28). However, in order to validate these marker regions, experimental data on ligase activity of different legumain isoforms is indispensable.

Structural analysis of plant legumains showed that they are synthesized as inactive zymogens composed of a caspase-like catalytic domain with the AEP activity (AEP domain) and a C-terminal death domain like prodomain (LSAM domain, Legumain Stabilization and Activity Modulation domain) that are connected by an activation peptide (AP) harboring the α6-helix (20,21,29,30). Although this tripartite domain architecture (AEP-AP-LSAM) is conserved in mammalian and plant legumains, the activation process of vegetative-type proAtLEGγ (*Arabidopsis thaliana* prolegumain isoform γ) significantly differs from that of human legumain (31,32). Importantly, proAtLEGγ is present in an enzymatically latent dimer state that is mediated by AP-LSAM – AP'-LSAM' interactions and is dependent on pH and protein concentration (21). Furthermore, we have previously shown that conversion to the active, monomeric AEP form, i.e. release of the prodomain, proceeds via a previously unknown two-chain intermediate state. Two-chain AtLEGγ results from cleavage at the N-terminal side of the α6-helix within the AP and is suppressed by high protein concentration where AtLEGγ dimerization is favored. Even after an initial cleavage within the AP, an enzymatically latent, dimeric two-chain AtLEGγ intermediate form remains stable at neutral pH-environment. Only at acidic pH the dimer dissociates to monomeric two-chain legumain, which may further release the LSAM domain and thereby convert to the mature AEP form. The identification of the dimer and two-chain states allowed developing a pH-dependent four step activation-model mechanism of plant legumains, i.e. single chain – two chain conversion; α6-helix destabilization; dimer – monomer dissociation; and AEP – LSAM release. However, given the subtle regulation of these conversions, isoform specific differences in activation are to be expected, with experimental data still lacking.

Here we present the crystal structure of zymogenic proAtLEGβ which led to the discovery of a distinct activation mechanism, contrasting that of AtLEGγ. Combining structural and biochemical information, we show for the first time that plant legumains follow isoform specific autocatalytic activation mechanisms and differential strategies of activity regulation and stability. Furthermore, we provide evidence that seed type AtLEGβ is an active ligase capable of peptide cyclisation. AtLEGβ ligase activity is not strictly linked to peptide bond cleavage but enables also the efficient joining of free N- and C-termini. To our knowledge, AtLEGβ is the first example of a plant legumain for which we could demonstrate the ligation of free peptide termini.

This study broadens our understanding of isoform-specific differences in plant legumains and their relevance in plant physiology. Furthermore, the study discloses new avenues to rationally design peptide ligases with applications in biotechnology and drug development.

Results

Crystal structure of proAtLEGβ

To understand isoform specific differences between different AtLEGs we determined the crystal structure of seed-type proAtLEGβ to a resolution of 2.0 Å (Table S1). The asymmetric unit of the tetragonal space group contained 12 independent molecules. Like isoform γ, proAtLEGβ comprises an N-terminal caspase-like catalytic domain and a C-terminal Legumain Stabilization and Activity Modulation (LSAM) domain with death domain-like topology (Fig. 1 and S1). The AEP and LSAM domain are connected by an activation peptide that harbors the α6-helix. Overall, the structure of proAtLEGβ closely resembles the structure of the homologous two-chain AtLEGγ indicated by a Cα-rmsd of 0.49 Å. However, inspecting the individual sub-domains unraveled specific differences. While the catalytic AEP domains of AtLEGβ and γ superimpose very well with an overall Cα-rmsd of 0.39 Å, we observed bigger differences in the LSAM domains with a Cα-rmsd of 0.78 Å (determined with Pymol). This observation is also in agreement with a higher sequence identity of the β and γ catalytic domains (67% identity) compared to the LSAM domains (56% identity). Furthermore, we observed an isoform specific glycosylation at Asn309, located at the bottom of the enzyme, which is also conserved in human legumain (Fig. 1A and 2A).

proAtLEGβ forms atypical dimers in the crystal and is monomeric in solution

An important feature of proAtLEGγ is that it exists in a latent dimer state in solution, which is mediated by AP-LSAM – AP'-LSAM' interactions. This dimer controls both the activation and activity of AtLEGγ (21). Similarly, in the crystal structure of proAtLEGβ we found all twelve independent protomers in the crystallographic asymmetric unit to engage in symmetric dimer contacts which were mediated by LSAM – LSAM' interactions (Fig. 2A). However, these interactions were mediated by different amino acids and led to an approximately 90° tilted orientation of the monomers within the β- and γ-dimer, respectively (Fig. 2B, C). Indeed, detailed analyses of the β and γ dimer interfaces revealed significant, isoform specific differences. The proAtLEGγ dimer is mediated primarily by three symmetric anchoring sites, α6 and α7 helices and a conserved cyclic protein recognition motif (cPRM) on the c341-loop. The α6 and α7 helices form a 4-helix bundle that is stabilized around a symmetric hydrophobic core formed by $W363^{\gamma}$ and $Val383^{\gamma}$, $L384^{\gamma}$, respectively (AtLEGγ numbering; Fig. 2C, E). This hydrophobic core is further stabilized by a network of salt bridges on the N- ($R355^{\gamma}$ – $E371^{\gamma}$' and $D356^{\gamma}$ – $K376^{\gamma}$') and C-terminal ($K376^{\gamma}$ – $D356^{\gamma}$' and $E371^{\gamma}$ – $R355^{\gamma}$') ends of the α6-helices. By contrast, the proAtLEGβ dimers in the crystal structure were predominantly mediated by the α7 helix. This interaction was formed around the symmetric $H384^{\beta}$ ($H392^{\gamma}$) and further stabilized by one symmetric salt bridge ($E390^{\beta}$ – $K383^{\beta}$') as well as by a hydrophobic contact of the α7 C-terminal LFG motif (396^{β}-398^{β}) with $W355^{\beta}$' centered in the α6 helix (Fig. 2D, E). The hydrophobic core of the α6-α7, α6'-α7' four helix bundle was missing as was any stabilization by the conserved cPRM, despite key residues important for proAtLEGγ-like dimer formation being conserved in proAtLEGβ (Fig. S1). However, modeling a proAtLEGγ-like dimer uncovered repulsive charge densities of α7-α7' helix contact residues in AtLEGβ (R380 – R380', K373 – K383', D369 – D386') that will prohibit this γ-mode of dimerisation (Fig. 2F and S2). Together, these findings suggest that the observed β-dimer is weak and probably only transient in solution. To test this conclusion, we performed size exclusion chromatography (SEC) experiments. As expected, at pH 7.0 proAtLEGβ migrated at the expected size of a monomer, similar to human legumain (Fig. 2F). Accordingly, proAtLEGβ was a monomer in solution.

Conserved Gln346 keeps proenzyme in latent state

Comparing the crystal structure of proAtLEGβ with YVAD-cmk inhibited AtLEGγ we found that the AP binds to the non-prime substrate binding sites in a substrate-like orientation, similar as we previously observed in mammalian prolegumain (Fig. 1B and Fig. 3A,B). Thereby the AP is blocking substrate access, keeping the proenzyme in a latent, inactive state. Additionally, we observed a conserved Gln346 (AtLEGβ numbering) on the N-terminal end of the α6 helix. Gln346 is binding into the S1 pocket in an unproductive orientation and thereby preventing cleavage of the AP and further blocking substrate access to the active site (Figs. 1B and 3B). This interaction was similarly observed in the crystal structure of *A.thaliana* legumain isoform γ; additionally, Gln346 is conserved throughout the plant VPE sequences, strongly suggesting that the Gln346-S1 binding forms a conserved mechanism in plant legumain activity regulation. Additionally, this interaction is further strengthened by the neighboring Arg347, which forms ionic interactions to Glu212, directly next to the catalytic Cys211 (Fig 3C).

Similar to two-chain AtLEGγ and mammalian prolegumains, the LSAM domain is further stabilized by two conserved disulfide bonds (Figs. 1B, and 3B). On the C-terminal end of the LSAM domain, AtLEGβ harbours a potential VSS (vacuolar sorting signal), which is however not structured and therefore not visible in the electron density (Fig. 1A).

Activation proceeds via two-chain intermediate state

In an effort to unravel the basic principles of proAtLEGβ activation, we analyzed the inter-domain interfaces of AEP and LSAM domains. Interestingly we found that the interface has a hydrophobic character with only two salt bridges identified by PDBe Pisa, R347-E212 and K422-D187, which are also conserved in proAtLEGγ ($R355^{\gamma}$-$E220^{\gamma}$ and $K432^{\gamma}$-$D195^{\gamma}$; Fig. 4A). This is in stark contrast to proAtLEGγ where the interdomain interface has a mixed charged-hydrophobic character, which is reflected by eight interdomain salt bridges and a hydrophobic cluster localized to the prime substrate binding sites (Fig. 4B). Interestingly, the conserved $D358^{\gamma}$-$R74^{\gamma}$, ($D348^{\beta}$ -$R66^{\beta}$) and $D358^{\gamma}$-$H177^{\gamma}$ ($D348^{\beta}$ –$H169^{\beta}$)

form salt bridges in proAtLEGγ, but not in proAtLEGβ due to a local reorientation of the α6 helix. The residues involved in other AtLEGγ-specific interdomain salt bridges are not conserved in AtLEGβ, i.e. $K365^{\gamma}$-$E109^{\gamma}$ ($M357^{\beta}$-$L101^{\beta}$), $R375^{\gamma}$-$E109^{\gamma}$ ($K367^{\beta}$-$L101^{\beta}$), $R375^{\gamma}$-$E264^{\gamma}$ ($K367^{\beta}$-$I256^{\beta}$) and $R490^{\gamma}$-$D136^{\gamma}$ ($L482^{\beta}$-$S129^{\beta}$). Combined with the differences in oligomerization state, these findings led to the hypothesis that there will be pronounced differences in the activation and pH-stability profiles of the two *A.thaliana* legumain isoforms.

Since the interaction between the catalytic domain and the LSAM domain in proAtLEGβ is primarily hydrophobic in nature, we expected that its activation would be rather independent of pH. Surprisingly, an SDS-PAGE based, pH-dependent activation assay uncovered that the activation profile of AtLEGβ closely resembles that of mammalian legumain, with complete activation only occurring at very acidic pH (4.0) (Fig. 5A). Consequently, we hypothesized that autocatalytic activation requires conditions that will destabilize the LSAM domain in order to gain accessibility to the active site. Indeed, we found complete degradation of the LSAM domain at pH ≤ 4.0, indirectly indicating that the LSAM domain is destabilized at acidic pH conditions (Fig. 5A). Interestingly, upon incubation at pH 5.0, proAtLEGβ was split into catalytic (AEP) and LSAM domain. However, the LSAM domain was not degraded but remained stable on SDS-PAGE. This suggested to us that AtLEGβ might form a two-chain state, where cleavage between LSAM and catalytic domain occurred but both domains remained bound to each other. To test this, we performed SEC experiments using proAtLEGβ activated at pH 5.0. Indeed, we found a mixture of two-chain state and isolated AEP domain (Fig. S3). Importantly, there was no dimeric two-chain intermediate state of AtLEGβ observed in SEC after activation.

Proteolytic activation is initiated by cleavages in the AP

Using mass spectrometry, we could identify two main autocatalytic cleavage sites, Asn333 and Asn345 on the AP (Fig. 1). These sites were similarly observed in proAtLEGγ and seem to be equally accessible to cleavage. Upon incubation at pH < 5.0, we observed additional cleavage sites on the LSAM domain, including Asp363, Asp416 and Asp417 (Fig. 1B). Due to the architecture of the S1-pocket, cleavage after Asp is restricted to low pH conditions (< 5.0), in line with the observed cleavage pattern. Interestingly, Asp363 is localized between the α6- and α7-helices and could, in combination with processing at Asn333/345, therefore allow the selective release of the α6-helix (fragment Gln346-Asp363) as observed in mammalian legumain (31). Asp416 and Asp417 are localized within the $V^{415}DDW^{418}$ motif, right before the α9-helix (Fig. 1B and S1). This motif is conserved within plant legumains and cleavage within this sequence was previously shown to be critical for the autocatalytic activation of castor bean legumain (33). Taken together, activation of AtLEGβ at pH < 5.0 goes along with cleavage at the aforementioned Asn and Asp sites, which finally results in the complete removal of the AP (including the α6-helix) and the LSAM domain. Thereby rendering the active site accessible for substrates.

In addition to cleavage on the AP and LSAM domain, we observed another processing at the N-terminal end of the protein. Here it is important to note that our proAtLEG expression constructs typically carry an N-terminal His_6-tag followed by a TEV-recognition site (ENLYFQG; TEV: tobacco etch virus protease). We found that AtLEGβ was capable of cleaving after the Asn residue within the TEV-recognition site and thereby removing the His_6-tag, as evidenced by a Western-blot using an Anti-His-Antibody (Fig. 5B,C). Based on SDS-PAGE experiments we propose that the primary cleavage at the Asn333/345 cleavage site can be catalyzed by the two-chain form. However, since N-terminal processing within the TEV-recognition motive was only observed at very acidic pH conditions we suggest that the latter cleavage is performed by the fully activated AtLEGβ. Importantly, N-terminal cleavage is not a physiological event as the relevant sequence is not present in native proAtLEGβ (Fig. S1).

(pro)AtLEGβ is stable at intermediate pH

Based on the remarkable variances we observed at the AEP – LSAM interfaces of proAtLEGβ and γ we hypothesized that they would translate into differences of their pH-stability profiles. Indeed, when we measured the thermal stability of proAtLEGβ using differential scanning fluorimetry we found a stability optimum at pH 5.0 (Fig. 4C). This is very different to proAtLEGγ and

mammalian legumain where the stability optimum of the proenzyme is at neutral pH (21,32). Even more interestingly, we found that AtLEGβ activated at pH 4.0 and thereby lacking the LSAM domain, similarly showed a maximum in pH-stability at pH 5.0 (Fig. 4D). This is in stark contrast to AtLEGγ and also mammalian legumain, where the AEP domain is most stable at pH ~4. However, this difference becomes clear considering the hydrophobic interaction between AEP and LSAM domain in proAtLEGβ. Mammalian legumain and AtLEGγ harbor a highly charged electrostatic stability switch (ESS) on the AEP surface, located at the area surrounding the active site (32). At neutral pH conditions, the ESS causes electrostatic destabilization of the isolated AEP domain because of the high negative charge density which is not compensated by the LSAM domain. In human legumain and AtLEGγ the isolated AEP can be stabilized by protonation of the excess acidic residues, hence the maximum stability at pH 4. The AEP in AtLEGβ lacks the pronounced ESS, explaining why a strong acidic pH is not necessary for charge neutralization, in agreement with the pH optimum at 5.0. The interaction of the AEP with LSAM generally stabilizes the protein. In proAtLEGβ, the AEP-LSAM interaction and stabilization do not depend on neutral pH, whereas the tight electrostatic clamping of these domains in proAtLEGγ and human prolegumain depend on neutral pH. Consequently, proAtLEGβ is most stable at the pH which is also favorable for the isolated AEPβ.

Overall topology of AEP domain is highly conserved

Previous studies showed that the AEP domain in prolegumain is present already in an active conformation (21,34). Zymogenicity resulted solely from the steric blockage of the active site by the AP and LSAM domain. Therefore, we can use the crystal structure of proAtLEGβ to also analyze the active AtLEGβ state. When we superimposed the AEP domains of AtLEGβ and γ we found that their fold is highly conserved (Fig. 6A). AtLEGβ exhibits a caspase-like topology, i.e. a 6-stranded central β-sheet that is surrounded by 5 major α-helices (Figs. S1, S4) (35). Furthermore, AtLEGβ harbours the c341- and c381-loops, which form the non-prime substrate binding sites. The c341-loop encodes a plant VPE specific disulfide bond that is stabilizing the proline-rich insertion that is extending the c341-loop compared to mammalian legumain (Fig. 6B). Mutation of Cys244 or Cys258 resulted in a complete loss of protein expression, confirming that the disulfide is also critical for folding. Furthermore, we observed 2 cis-imide peptide bonds (Thr180-Pro181 and Asn248-Pro249) with relevance for stabile bend and turn formation (Fig. 3B) (36). Interestingly, both turns are located to the substrate binding sites. The Asn248-Pro249 cis-peptide bond is on the c341-loop (non-prime side) and presents the Asn248 carbonyl oxygen as main chain recognition site for the P4 amide. Thr180-Pro181 is part of the eastern rim of the S2' pocket.

AtLEGβ has a wide S3-S4 pocket

When looking into the active site, we found that also the active site residues Cys211, His168 and Asn64 superimpose very well with the related AtLEGγ (Fig. 6B). Furthermore, the residues forming the S1-specificity pocket, Arg66, His67, E209 and D261 adopt identical conformations as observed in AtLEGγ. The highly conserved architecture of the active site suggested similar substrate specificity and catalytic activity of AtLEGβ and γ. However, when we compared the catalytic activity towards the fluorogenic Ala-Ala-Asn-AMC substrate, we observed a surprisingly low catalytic activity for AtLEGβ as compared to γ (Fig. 6C). Since the positioning of the active site residues were basically identical in β and γ, we did not expect this difference in activity to originate from a k_{cat} effect, but rather from differences in substrate affinity (K_M). Beyond the highly similar S1-pocket, we indeed identified major differences on the c341- and c381-loops on the non-prime side (Fig. 6B and S5). Variations in sequence and conformation resulted in a narrow S3 – S4 pocket in AtLEGγ but a rather wide pocket in AtLEGβ (Fig. 4A,B, Fig. 6 and S5). To test whether these differences were a result of induced fit of the YVAD-cmk inhibitor, we superposed the crystal structures of proAtLEGβ, two-chain (pro)AtLEGγ and active YVAD-AtLEGγ and compared their active sites. Interestingly, we found that the conformations of the substrate specificity loops c341 and c381 of proAtLEGβ most closely resembled the active state of AtLEGγ. Thereby we could exclude that induced fit was a main regulator of substrate affinity (Fig. S5). However, the situation might be different in AtLEGγ, where we observed pronounced conformational changes of

the c381-loop between the proenzyme and the YVAD-cmk inhibited form. Modeling a peptidic substrate, based on the YVAD-cmk-AtLEGγ crystal structure, we found tight interactions in AtLEGγ but less interactions to AtLEGβ. While AtLEGβ offered an open, broad surface to accommodate the YVAD substrate, AtLEGγ was tightly embracing the peptidic substrate as visible in Fig. 4A,B and 6B. We could assign $Tyr240^{\beta}$/$Trp248^{\gamma}$ on the c341-loop and $Gly300^{\beta}$/$Tyr307^{\gamma}$ on the c381-loop as the main determinants for this difference. Together, this suggested to us, that small peptidic substrates would bind with lower affinity to AtLEGβ as compared to gamma, because of missing enzyme-substrate interactions. Indeed, when we determined K_M values for AtLEGβ and γ towards the AAN-AMC substrate we found high affinity binding (K_M = 57 ± 3 µM) to AtLEGγ but low affinity for AtLEGβ (K_M = 337 ± 3 µM) (Fig. 6D). Importantly, we found similar k_{cat} (AtLEGβ: 4.5 x 10^{-3} min^{-1} and AtLEGγ: 6.3 x 10^{-3} min^{-1}) and V_{max} values (AtLEGβ: 0.9 x 10^{-3} µmol/min and AtLEGγ: 1.1 x 10^{-3} µmol/min) for both enzymes. These findings confirmed that the difference in catalytic activity between AtLEGβ and γ was explained by differences in substrate affinity. Interestingly, when we used a VAN-AMC substrate instead of AAN-AMC we observed a reduction in enzymatic activity both for AtLEGβ and γ (Fig. 6C). Accordingly, the smaller alanine is preferred over the branched valine at the P3 position in both AtLEG isoforms. Furthermore, we found an activity optimum for AAN-AMC turnover at pH 5.5 which is also in agreement with the pH-stability requirements of the AEP-domain (Fig. S6).

c381-loop is variable in length and sequence

Together, these observations made us hypothesize that the c341- and c381-loop might serve as K_M-switch. To analyze this further, we superposed all plant legumain structures available in the PDB. While the main structural elements superimposed very well in all available structures, we observed big differences on the c381-loops. It is variable in length, sequence and may even contain a glycosylation site (Fig. 6E and S4). Together these findings suggested that the c381-loop is a main determinant of the proteolytic activity of legumains, similar to caspases. The relevance of the c381-loop for legumain activity is further supported by a previous analysis suggesting it as a marker of ligase activity (MLA) (28).

AtLEGβ substrate specificity is pH-dependent

To further analyse the substrate specificity of AtLEGβ, we carried out PICS experiments, which use proteome-derived peptides as substrate libraries (37,38). Here we used a peptide library that was generated from an *E.coli* proteome by digestion with trypsin for AtLEG specificity profiling at three different pH conditions. As expected, we observed a strong preference for Asn in P1 position at all investigated pH values (Fig. 7A). Interestingly, we also observed an increasing frequency of cleavage at Asp residues upon prolonged incubation times (18h). This time-dependence illustrates that substrates with Asn in P1 are kinetically favored over Asp. The substrate preference was also pH-dependent, i.e. the turnover rate of P1-Asp substrates increased with lower pH values, which nicely agrees with the bi-polar architecture of the S1-specificity pocket and with previously published data for human legumain (Fig. 6B) (32).

AtLEGβ has a strong preference for hydrophobic residues in P2'

Furthermore, we observed a slight preference for small, polar residues in P1' position which was especially visible at the shorter incubation times (Fig. 7A), suggesting that P1'-Gly is kinetically preferred. Additionally, we found a pronounced preference for Leu in P2' position. Leucine has previously been proven to be beneficial at P2' position in legumain ligase substrates (16). Together, these results are in nice agreement with the architecture of the S1' and S2' binding sites: While the S1' binding site is flat and not allowing much interaction with the enzyme, the S2'-binding site forms a pronounced pocket (Fig. 4A) (29). Small residues in P1' position will facilitate the simultaneous binding of the P1 and P2' residues into the respective S1 and S2' binding pockets while still maintaining enough flexibility to allow efficient cleavage of the scissile peptide bond. The bottom of the S2' pocket is formed by Gly176 and the eastern wall by His182 (Fig. 7B). Gly176 is conserved in all plant legumains that have been structurally characterized so far (Fig. 7C). The eastern wall is mostly histidine and tyrosine, with some exceptions. Interestingly, mammalian legumain harbors a valine at position 176, making the S2'-pocket shallower and thereby less specific

at this position (39). Furthermore, AtLEGδ has the glycine replaced by alanine (Fig. 7C), suggesting that it will also have a less pronounced specificity at P2' position. To test the relevance of His182 for prime side substrate specificity, we repeated the PICS experiments using AtLEGγ, which has a tyrosine at the equivalent position (Fig. S7). Interestingly, we found highly similar preferences on the non-prime and prime substrate binding sites, further confirming that Gly176 is the main determinant at the S2' site.

AtLEGβ has a strong preference for small residues in P1' position in protein substrates

In a next step we analyzed the substrate specificity of AtLEGβ towards protein substrates, using proteome extracts isolated under non-denaturing conditions from leaves of the *A. thaliana vpe0* mutant lacking expression of all four VPE isoforms as a substrate library. After incubation with recombinant AtLEGβ, recombinant AtLEGγ or buffer control, free N-terminal α-amines where labeled with three different formaldehyde isotopologues, and cleavage sites determined using the HUNTER N-termini enrichment and mass spectrometry (40). Based on three biological replicates we identified 381 N-terminal peptides significantly accumulating after incubation with AtLEGβ at pH 6.0 (Figure 8A,B and Supplementary Table 1), matching to 363 unique cleavage sites (Fig. 8C) in 289 proteins (Fig. 8D). As expected, we found a pronounced preference for Asn at P1 position (Fig. 8B). Furthermore, we observed a stronger preference for small and polar residues in P1' position, suggesting that the accessibility of the scissile peptide bond is enhanced when it is flanked by a small residue. Additionally, we also noticed a slightly increased preference for the more bulky and charged Asp and Glu amino acids. As in the peptide-based PICS experiment, we again observed a preference for hydrophobic amino acids in P2' position. For AtLEGγ, we identified 412 significantly accumulating N-terminal peptides (Fig. 8E, Supplementary Table 1). These matched 390 unique cleavage sites (Fig. 8C) in 304 proteins (Fig. 8D) that exhibited a very similar cleavage profile in line with our observations using peptide substrates (Fig. 8F). Notably, a vast majority of 313 of the cleavage sites in 257 proteins were cut by both enzymes, while only 50 cleavages in 32 proteins were strongly preferred substrates of AtLEGβ and 77 cleavages sites of 47 proteins were selectively cut by AtLEGγ (Fig. 8C,D and G).

AtLEGβ is a broad spectrum transpeptidase

To characterize the cyclase activity of AtLEGβ we co-incubated it with different SFTI-derived linear peptides and measured the formation of the cyclic product using mass spectrometry. Indeed, we found that AtLEGβ could cleave the SFTI-GL precursor peptide to the linear L-SFTI (lacking GL) version and further cyclize it to cyclic SFTI (c-SFTI) (Fig. 9A). Cyclisation worked most efficient at pH 6.0, which is in agreement with the previously reported pH-requirements of legumain ligase activity (29,41). Using the SFTI-GL precursor peptide, which harbors an Asp at P1 position, we observed a product formation rate of about 60%. This is less than compared to AtLEGγ, which resulted in approx. 80% product formation (29). Interestingly, when the P1 residue was replaced by Asn, as is the case in SFTI(N14)-GL, AtLEGβ was still able to catalyze peptide cyclisation, contrasting the situation of AtLEGγ. When we replaced the P1'-P2' Gly-Leu by His-Val residues, which is the preferred sequence found for butelase-1 (*C. ternatea* legumain), we observed a similar cyclisation efficiency (Fig. S8) (16). Showing us that albeit optimized for butelase-1 the HV-dipeptide is not facilitating peptide ligation in AtLEGβ.

AtLEGβ is a broad spectrum ligase

Along that line we also tested whether AtLEGβ would be able to cyclize linear L-SFTI and L-SFTI(N14) peptides, which are lacking amino acids on P1' and P2' positions of the protease substrate. Surprisingly, AtLEGβ was indeed able to join the free termini and form the cyclic product, suggesting that AtLEGβ is not only a transpeptidase but a real ligase (Fig. 9A and B). Using the SFTI peptides carrying Asn at P1 position (N14), cyclisation worked equally well with or without preceding cleavage of prime side residues. In case of Asp at P1 position, transpeptidation (cleavage-linked ligation) was preferred as compared to joining free ends. Again, product formation was pH dependent, working best at near neutral pH conditions. So far, there was not a single report of a (plant) legumain capable of efficiently linking free peptide termini.

Discussion

Dimerization is a critical regulatory event for caspase-like proteins. In case of the apoptotic caspases, dimerization is mediated primarily by the β6 strand on the catalytic domain and is associated with structural rearrangements that render the caspase active (Fig. S1). Similarly, dimerization was also observed in plant legumains. The crystal structures of OaAEP1 (pdb entry 5hoi) and AtLEGγ (5nij) both showed a dimer state that was mediated by the α6 and α7 helices on the LSAM domain. However, in these cases dimerization was not associated with activation but rather with inactivation. Under conditions where dimerization is maintained, such as high protein concentration, the proenzyme will not auto-process to the active AEP form. Additionally, there is a two-chain intermediate state, which is active to some extent. In this study we show for the first time that there are isoform specific differences in the activation and activity regulation of *A.thaliana* legumains. Firstly, we observed that proAtLEGβ is monomeric in solution. In this respect, autocatalytic activation of proAtLEGβ resembles more the mechanism known from mammalian legumain, which also lacks a stable, latency-conferring dimer state (Fig. 10). We should point out, however, that in the crystal we found six equivalent proAtLEGβ dimers per asymmetric unit. Nonetheless, this atypical dimer interaction is transient and short-lived, hence could not be observed in solution experiments. Secondly, we found that the AEP – LSAM interface is rather hydrophobic and not charged in nature. Consequently, the stability profile of AtLEGβ differs from AtLEGγ and mammalian legumains (Fig. 10). Thirdly, AtLEGβ encodes autocatalytic cleavage sites on both ends of the α6-helix (Asn345 and Asp363), which in principle allows the selective removal of the AP, like in mammalian legumain (32). While N-terminal cleavage was observed at pH < 6.0, cleavage on the C-terminal end of the α6-helix is restricted to pH < 5.0, which is in agreement with the charge requirements of the S1 pocket (Fig. 5A and 6B). Additionally, at acidic pH the ionic clamp that is linking the N-terminal end of the α6-helix (Arg347) to the active site (Glu212), will loosen (Fig. 3C), which will further facilitate the release of the AP (α6-helix). Therefore, an AEP–LSAM complex might represent a critical intermediate state, which initiates the complete removal of the LSAM domain by proteolytic degradation and/or conformational destabilization. However, as we did not observe a stable AEP–LSAM complex in our experiments, it will only be short lived (Fig. 5 and 10).

These unique characteristics provide another new regulatory mechanism distinct from that of AtLEGγ. Different oligomerisation states will cause AtLEGβ to favor activation at high local concentrations but will favor the latent two-chain state in AtLEGγ. On the other hand, the transient dimers observed in the AtLEGβ crystal might possibly play a role in cooperative substrate processing. Together these findings suggest that AtLEGβ and γ represent examples of two distinct classes of plant legumains, not only concerning their physiological function but also with regard to completely different mechanisms of zymogenicity, activation and stability.

All plant legumains are specific for cleaving after P1-Asn. However, we could show that subtle differences in the non-prime substrate binding sites translate into pronounced kinetic differences. Consequently, different legumain isoforms will feature kinetically-driven substrate preferences, which may be modulated by the amount and time of substrate availability (Fig. 7A). We provide evidence that the c381-loop can encode such kinetic differences. The corresponding sequences and conformations differ significantly in plant legumains, making it the single most variable region within the plant legumain catalytic domain. Differences in substrate affinity (K_M) can be kinetically assayed using specific substrates. PICS assays with proteome-derived peptide libraries are typically insensitive to such differences due to the mixed and unknown concentration of individual peptide substrates. However, if the substrate affinity is extremely different, such preferences can become apparent. Indeed, using time-series experiments we show that P1-Asp is a low affinity legumain substrate at increasing pH values. Presenting a P1-Asp may consequently serve as a strategy to kinetically regulate substrate turnover, i.e. to release a certain cleavage product in a slow, and pH-controlled manner. An example includes the autocatalytic activation of proAtLEGβ which critically depends on cleavage after Asp residues on the LSAM domain and which is thereby restricted to low pH. Together this indicates that the differences in the c381-loop among the plant legumains will have an impact on cleavage kinetics

rather than on sequence specificity. In line with these observations, we found mostly overlapping AtLEGβ and γ cleavage sites in protein substrates *in vitro*.

Previously the c381-loop was described as marker of ligase activity (28). More precisely, a deletion in that region was associated with an increase in ligase activity. However, both AtLEGβ and γ encode relatively long c381-loops, yet both are active ligases. Furthermore, we could show that not only the sequence but also the conformation of this loop can be quite different, although it might be similar in length (Fig. S4). Therefore, we suggest that the c381-loop is primarily a determinant of protease activity. Since protease and ligase activities are inversely coupled, the c381-loop may be an indirect marker of ligase activity: If the affinity of the non-prime (protease) substrate is low, the affinity of the prime-side ligase substrate might be relatively high in comparison. Such a situation favors transpeptidation over substrate hydrolysis. Furthermore, low affinity of non-prime substrates may also result in less re-cleavage of cyclic products, and thereby again indirectly favor ligation. This hypothesis also fits to our observation that AtLEGβ, which has a non-prime binding site optimized for low affinity binding, is a ligase with broad substrate specificity.

In general, we found that SFTI-derived peptides harboring Asp at P1 position are better ligase substrates resulting in most efficient formation of cyclic product. This observation fits with the notion that poor (high K_M) non-prime substrates are more likely to find a prime ligase substrate at the active site, which in turn excludes the catalytic water molecule from the active site. In concert, the poor non-prime substrate affinity should favor aminolysis of prime substrate over hydrolysis – by the catalytic water that is excluded from the active site. For P1-Asp substrates this is particularly true at near neutral pH, where ligation is favored (Fig. 10). Additionally, the residence time of the ligation product is very short, making re-cleavage of the cyclic product unfavorable and consequently indirectly stabilizing the cyclic product. However, L-SFTI, which lacks P1' and P2' residues, resulted in less formation of cyclic product, indicating that P1-Asp will only be tolerated as a substrate at near neutral pH if coupled to prime side amino acids. P1-Asn as a free C-terminal end worked better, probably because Asn is in general a better K_M substrate at pH 6.0. Likely, the K_M will also be influenced by prime side residues. As a result, a substrate with P1-Asp linked to prime side amino acids will have a critically superior (lower) K_M as compared to C-terminally free Asp, giving the P1-Asp substrate the possibility for binding and transpeptidation to happen. Looking at the prime substrate binding sites, we found that AtLEGβ and γ encode nearly identical substrate binding sites. Taken together, differences in ligation efficiency between AtLEGβ and γ might therefore be explained by their different non-prime substrate binding sites optimized for low and high affinity binding respectively.

In addition to the marker of ligase activity, Cys247 (*O. affinis* numbering) was identified as a gate keeper residue for ligase activity (20). Mutation to Ala247 resulted in an enzyme with superior ligase activity. Since all AtLEG isoforms harbor a glycine at the equivalent position (Gly241, AtLEGβ numbering), this residue cannot explain the observed isoform specific differences in ligase activity. Similarly, the sequence motif Gly171-Pro172 (AtLEGβ numbering; Fig. S1) which is located close to the S1' pocket and was recently found to be beneficial for ligase activity is conserved in both *A.thaliana* legumains (30). However, directly next to Gly241 is Tyr240, which is a critical part of the non-prime binding site (S2 – S3) and which is different in AtLEGγ (Trp248). Based on this observation and the differences in the close-by c381-loop, we suggest that it is rather the overall architecture of the non-prime substrate binding sites that affect substrate affinity and might positively affect ligase activity.

The ability of AtLEGβ to join free ends is also interesting from a biotechnological point of view, as it will allow to link targets without the necessity of introducing artificial cleavage sites. However, still with the prerequisite of a P1-Asn or Asp. Furthermore, it also highlights that joining free termini is a general feature encoded in selected plant legumain isoforms. Given that all plants express a variety of different legumain isoforms, it is very likely that there is an AtLEGβ-like enzyme present in every plant.

Previously we could show that the two-chain state observed in AtLEGγ is especially interesting with regard to ligase activity, as it is stable at neutral pH-environments where ligase activity is favored. Since two-chain AtLEGβ has the same pH-stability profile as active AEP with a pH stability optimum at 5.0, two-chain AtLEGβ will most likely not be a

superior ligase. However, it may implement differences in substrate specificity and catalytic efficacy. Indeed, we could previously demonstrate that human two-chain legumain with the C-terminal LSAM domain still present, exhibits carboxy-peptidase activity rather than endopeptidase activity (32). The carboxypeptidase activity is structurally encoded by LSAM-derived arginine residues, which anchor the carboxy-terminus at the primed recognition site. Interestingly, we observed a slight preference for aspartate and glutamate residues in the P1' position of protein substrates (Fig. 8B) together with a relative depletion of basic residues (K, R), which could similarly indicate carboxypeptidase activity of two-chain AtLEGβ. However, this observation has to be taken with some caution, as the relative increase in specificity for Asp and Glu at the P1'-position was low. Finding out whether or not two-chain AtLEGβ indeed harbours carboxy-peptidase activity will require further experiments and may be the subject of future studies.

Experimental procedures

Protein preparation

The *Arabidopsis thaliana* vacuolar processing enzyme (VPE, legumain) isoform β (AtLEGβ) full-length clone U12200 (locus AT1G62710) was obtained from the Arabidopsis Biological Resource Center (ABRC). Using this as a template we subcloned an N-truncated variant missing the N-terminal signal sequences into the pLEXSY-sat2 (Jena Bioscience, Germany) vector using PCR amplification and XbaI and NotI restriction enzymes. The final expression construct carried an N-terminal signal sequence for secretory expression in the LEXSY supernatant and an N-terminal His_6-tag followed by a TEV recognition site. Furthermore, we prepared a C211A dead mutant using the round-the-horn site directed mutagenesis technique, which is based on the inverse PCR method (42). Primers were designed which allowed the amplification of the cyclic plasmid template, harbouring the proAtLEGβ wild-type insert, to a linear full-length PCR product carrying the desired mutation on one end of the PCR product. Following gel extraction of the PCR product and blunt end ligation, an intact plasmid carrying the desired mutation was generated and transformed into *E.coli* Xl2(blue) cells. The C211A-mutant was used for crystallization experiments. Correctness of all constructs was confirmed by DNA sequencing. The so generated expression constructs were stably transfected into the LEXSY P10 host strain and stable cell lines were grown as described previously (21). Protein expression and purification was performed as described elsewhere (21,31). The final proAtLEGβ protein was stored in a buffer composed of 20 mM Hepes pH 7.0 and 50 mM NaCl. ProAtLEGγ was prepared following the same protocol.

Crystallization, data collection and refinement

Initial screening was performed using the sitting-drop vapour-diffusion method utilizing a Hydra II Plus one liquid-handling system. Crystals of proAtLEGβ were obtained in a condition composed of 0.5 M ammonium sulfate, 1 M Lithium sulfate and 0.1 M trisodium citrate. Crystals grew within 2 weeks at a protein concentration of 10 mg/ml. To prevent autocatalytic activation we used a C211A dead mutant. Following pre-incubation in a cryo-protectant solution containing 0.8 M ammonium sulfate, 1.5 M lithium sulfate, 0.1 M trisodium citrate and 10% sucrose. Crystals were flash frozen in liquid nitrogen and subjected to Xray measurements. A high-resolution data set was collected at the ESRF on beamline ID30B. The beamline was equipped with a Pilatus 6M detector. Data collection was performed at a wavelength of 0.94 Å, 0.037 s exposure time and 15.3% transmission. 1,000 images were collected at an oscillation range of 0.1° and 100 K. Diffraction images were processed using xds and scala from the CCP4 program suite (43,44). An initial model could be obtained by molecular replacement using PHASER (45), using the crystal structure of two-chain AtLEGγ combined with the sequence of proAtLEGβ. Following iterative cycles of model building in coot (46) and refinement in phenix (47) a final model was obtained and coordinates and structure factors were deposited to the PDB under the accession code 6YSA.
Electrostatic surface potentials were created with APBS (48) after assigning charges at pH 7.0 using Pdb2pqr (49). Surface potentials were contoured at +/-5 kT/e.

Autoactivation

To test the pH-dependence of auto-activation of proAtLEGβ we incubated it at 0.4 mg/ml concentration in buffer composed of 100 mM buffer substance (pH 3.5 – 6.0: citric acid; pH 6.5: MES; pH 7.0: Hepes), 100 mM NaCl and 2 mM DTT for 1 h at 25 °C. Reactions were stopped by the addition of 10 mM MMTS (S-methyl methane thiosulfonate; Sigma-Aldrich) before subjecting them to SDS-PAGE.
To generate active AtLEGβ in large scale we incubated the proenzyme in a buffer composed of 100 mM citric acid pH 4.0, 100 mM NaCl and 2 mM DTT at 25 °C for 1 h. Completion of auto-activation was analyzed by SDS-PAGE. Active AtLEGβ was buffer exchanged using a NAP column (GE Healthcare) pre-equilibrated in a buffer composed of 20 mM citric acid pH 4.0 and 50 mM NaCl. Active AtLEGγ was prepared following the protocol described in (21).

Enzymatic activity assays

The enzymatic activity of active AtLEGβ was investigated using the peptidic Z-Ala-Ala-Asn-7-amino-4-methylcoumarin (Z-AAN-AMC; Bachem) and Z-Val-Ala-Asn-AMC (VAN-AMC) substrates. Activity was measured in assay buffer composed of 100 mM citric acid pH 5.5, 100 mM NaCl, 2 mM DTT and 100 µM of the respective substrate at 25 °C after adding the enzyme at 60 nM concentration. Assays were carried out in an infinite M200 plate reader (Tecan). Increase in fluorescence was measured at 460 nm upon excitation at 380 nm. K_M values were determined upon incubation of AtLEGβ or γ with serial dilutions of the AAN-AMC substrate in assay buffer. Kinetic data was processed using GraphPad and K_M values were calculated using implemented algorithms.

Characterization of oligomerization state

To test the oligomerization state of proAtLEGβ, 200 µl of sample were loaded on a S200 10/300 GL column (GE Healthcare) equilibrated in a buffer composed of 20 mM Hepes pH 7.5 and 100 mM NaCl. To test the oligomerization state of pH 5.0 activated AtLEGβ, we loaded it on a S200 column pre-equilibrated in buffer composed of 20 mM citric acid pH 5.0 and 100 mM NaCl. BSA served as a size standard.

Determination of melting temperatures

To access the thermal stability of proAtLEGβ and pH 4.0 activated AtLEGβ we used the Thermofluor method. Experiments were setup as described previously (50). The investigated assay buffers were composed of 100 mM buffer substance (pH 4.0, 5.0, 6.0: citric acid; pH 7.0: Hepes) and 100 mM NaCl. Fluorescence data was analyzed as described elsewhere (51).

Western blot

Protein samples to be analysed were separated on an SDS-PAGE gel. Subsequently, proteins were blotted onto a Amersham Protran 0.45 NC membrane (GE Healthcare) using a Trans-Blot SD semi-dry transfer cell (Bio-Rad). The membrane was blocked with 1 x TBST supplemented with 5% (w/v) nonfat dry milk. Subsequently, the membrane was incubated with 5% milk-TBST supplemented with 1:10,000 (v/v) Anti-His-HRP antibody (ROTH). Chemiluminescent detection of His-tagged protein was performed by using the Amersham ECL prime Western blotting detection reagent (GE Healthcare) together with an Odyssey Fc imaging system (Li-Cor).

Substrate specificity profiling

To test the substrate specificity of AtLEGβ and –γ we carried out Proteomic Identification of protease Cleavage Sites (PICS) assays using peptide libraries generated from *Escherichia coli* Bl21 cells (37,38). The peptide library was prepared as described previously (52). The proteome (2.2 mg/ml) was digested with trypsin at a ratio of 1:100 in 100 mM Hepes pH 7.5 overnight at 37 °C. The peptide library (2 mg/ml) was incubated with AtLEG proteases (10 µg/ml) in assay buffer composed of 50 mM buffer substance (pH 4.0 and 5.5: citric acid; pH 6.5: MES) and 100 mM NaCl at 25 °C. Samples were taken after 1 h and 18 h of incubation. Protease treated samples were stable isotope-labeled with 20 mM heavy formaldehyde ($^{13}CD_2O$) and 20 mM sodium cyanoborohydride and control reactions with 20 mM light formaldehyde (CH_2O) and 20 mM sodium cyanoborohydride for 2 h and quenched with 100 mM Tris pH 8.0 for 1h. Protease-treated and control samples were mixed and purified using C18 StageTips.

Substrate specificity profiling of AtLEGβ and γ using intact *A. thaliana* leaf proteome

A. thaliana VPE quadruple mutant (VPE0, (53)) was obtained from the Nottingham Arabidopsis Stock center (accession N67918). Leaves were harvested from 8 week old plants grown on soil under short day conditions (9 h/15 h photoperiod, 22ºC/ 18ºC, 120 µmol photons m^{-2} s^{-1}). The harvested leaves were homogenized with a Polytron PT-2500 homogenizer (Kinematica, Luzern, Switzerland) in extraction buffer containing 0.05 M MES pH 6.0, 0.15 M NaCl, 10% (w/v) sucrose, 0.01 M DTT and HALT protease inhibitor cocktail (ThermoFisher, Dreieich, Germany) on ice. The lysate was then filtered through Miracloth (Merck, Darmstadt, Germany), followed by centrifugation at 4000 x g at 4 ºC, for 5 min. The protein concentration was determined by the Bradford assay using BSA as a reference.
Recombinant AtLEGβ, recombinant AtLEGγ or buffer control were added to the isolated proteome at a protease to proteome (1 mg) ratio of 1:100 (w/w) in the extraction buffer and incubated in parallel at 25 ºC for 3 hours. The reactions were terminated by addition of 50 µM caspase-1 inhibitor (YVAD-cmk, Bachem, Switzerland). The reaction mixtures were purified by chloroform-methanol precipitation (54) and resuspended in 6 M GuaHCl, 0.1 M HEPES pH 7.5. The protein concentrations were determined using the BCA assay (ThermoFisher, Dreieich, Germany). The digested proteomes were reduced with 5 mM DTT at 56 ºC for 30 min followed by alkylation with 15 mM iodoacetamide for 30 min at 25ºC and quenched by addition of 15 mM DTT for 15 min. The three samples were differentially dimethyl labeled with 20 mM light formaldehyde ($^{12}CH_2O$) and 20 mM sodium cyanoborohydride (light label), 20 mM medium formaldehyde ($^{12}CD_2O$) and 20 mM sodium cyanoborohydride (medium label) or 20 mM heavy formaldehyde ($^{13}CD_2O$) and 20 mM sodium cyanoborodeuteride. After 16 hours incubation at 37 °C, the same amounts of fresh reagents were added and incubated for another 2 hours. The reactions were quenched with 0.1 M Tris (final concentration) at pH 7.4 and 37 ºC for 1 hour. Equal amounts of protein were pooled, purified by chloroform-methanol precipitation and resuspended in 0.1 M HEPES pH 7.4. The sample was then digested with trypsin in a 1:100 protease:protein ratio (SERVA Electrophoresis, Heidelberg, Germany) at 37 ºC for 16 hours. Enrichment of N-terminal peptides was performed according to the HUNTER method (40). In brief, trypsin-digested sample was tagged with undecanal at a ratio of 50:1 (w/w) in 40% ethanol supplemented with 20 mM sodium cyanoborohydride at 50 ºC for 45 min. Additional 20 mM sodium cyanoborohydride was added for another 45 min under the same condition. The reaction was then acidified with a final concentration of 1% TFA and centrifuged at 21,000 x g for 5 min. Next, the supernatant was injected through a pre-activated HR-X (M) cartridge (Macherey-Nagel, Düren, Germany). The flow-through containing N-terminal peptides was collected. Remaining N-terminal peptides on the HR-X (M) cartridge were eluted with 40% ethanol containing 0.1% TFA, pooled with the first eluate and subsequently evaporated in the SpeedVac to a small volume suitable for C18 StageTip purification prior to mass spectrometric analysis. The assays were performed in three biological triplicates.

Mass spectrometry data acquisition

Samples were analyzed on a two-column nano-HPLC setup (Ultimate 3000 nano-RSLC system with Acclaim PepMap 100 C18, ID 75 µm, particle size 3 µm columns: a trap column of 2 cm length and the analytical column of 50 cm length, ThermoFisher) with a binary gradient from 5-32.5% B for 80 min (A: H_2O + 0.1% FA, B: ACN + 0.1% FA) and a total runtime of 2 h per sample coupled to a high resolution Q-TOF mass spectrometer (Impact II, Bruker) as described (55). Data was acquired with the Bruker HyStar Software (v3.2, Bruker Daltonics,) in line-mode in a mass range from 200-1500 m/z at an acquisition rate of 4 Hz. The Top17 most intense ions were selected for fragmentation with dynamic exclusion of previously selected precursors for the next 30 sec unless intensity increased three-fold compared to the previous precursor spectrum. Intensity-dependent fragmentation spectra were acquired between 5 Hz for low intensity precursor ions (> 500 cts) and 20 Hz for high intensity (> 25k cts) spectra. Fragment spectra were averaged from t-stepped parameters, with 50% of the acquisition time manner with split parameters: 61 µs transfer time, 7 eV collision energy and a collision RF of

1500 Vpp followed by 100 µs transfer time, 9 eV collision energy and a collision RF of 1800 Vpp

Mass spectrometry data analysis

Acquired mass spectra were matched to peptide sequences at a FDR of 0.01 using MaxQuant (56) v.1.6.0.16 using standard Bruker QToF instrument settings. For PICS experiments, the UniProt *E.coli* K12 proteome database (downloaded Nov 2015, 4313 entries) with appended common contaminants was used. Search parameters considered semi-specific tryptic peptides, light (+28.031300) and heavy (+36.075670) dimethyl labeling at peptide N-terminal or Lys side chain amines, Cys carbamidomethylation as fixed and Met oxidation as variable modification. Identified peptides that showed at least a fourfold increase in intensity after protease treatment compared to the control treatment or were exclusively present in the protease-treated condition were considered as putative cleavage products. An in-house Perl script was used to remove putative library peptides (trypsin specificity on both sides of the identified peptide) and to reconstruct the full cleavage windows from the identified cleavage products as described (38) and visualized as IceLogos using software version 1.3.8 (57).
For HUNTER experiments, the *A. thaliana* UniProt proteome database (downloaded Dec 2018, 41592 entries) with appended list of common laboratory contaminants was used for searches that considered C-terminal cleavage by ArgC as digestion enzyme. Further search parameters included isotope labeling by light (+28.031300), medium (+32.056407) or heavy (+36.075670) dimethylation of peptide N termini or Lys residues, Cys carbamidomethylation as fixed and Met oxidation, N-terminal acetylation (+42.010565) or N-terminal pyroGlu formation from Glu (-18.010565) or Gln (-17.026549) as variable modifications. Further statistical data analysis, filtering and annotation was performed with the Perl script MANTI.pl 3.9.7 (https://manti.sourceforge.io).

Peptide cyclisation assay

SFTI-derived peptides were synthesized and analysed as described previously (29). Subsequently, cyclization experiments were carried out using 500 µM of the respective linear peptide and 0.5 µM AtLEGβ in a buffer composed of 100 mM NaCl and 50 mM Tris, Bis-Tris, citric acid pH 4.0 or pH 6.0. Reactions were incubated at 30 °C for 12 h. Subsequently the reactions were desalted using ZipTip C18 tips (Merck Millipore) and analyzed by MALDI-TOF-MS (Autoflex, Bruker Daltonics, matrix: α-cyano-4-hydroxycin-namic acid).

Data availability: The coordinates and structure factors presented in this paper have been deposited with the Protein Data Bank (PDB) with the accession code 6YSA. MS data have been deposited with the PRIDE (https://www.ebi.ac.uk/pride/archive/) (58) repository with the accession codes PXD019220 for the PICS dataset and PXD19276 for the HUNTER N-terminome dataset. All remaining data are contained within the article.

Acknowledgements
We thank Sabine Markovic-Ullrich for peptide synthesis.

Funding and additional information
This work was primarily supported by the Austrian Science Fund (FWF, project numbers W_01213 and P31867) with additional support by a starting grant of the European Research Council with funding from the European Union's Horizon 2020 program (grant 639905, to PFH.).

Conflict of Interest
The authors declare that they have no conflict of interest with the contents of this article.

References

1. 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. *The Plant cell* **5**, 1651-1659
2. Abe, Y., Shirane, K., Yokosawa, H., Matsushita, H., Mitta, M., Kato, I., and Ishii, S. (1993) Asparaginyl endopeptidase of jack bean seeds. Purification, characterization, and high utility in protein sequence analysis. *The Journal of biological chemistry* **268**, 3525-3529
3. Becker, C., Shutov, A. D., Nong, V. H., Senyuk, V. I., Jung, R., Horstmann, C., Fischer, J., Nielsen, N. C., and Muntz, K. (1995) Purification, cDNA cloning and characterization of proteinase B, an asparagine-specific endopeptidase from germinating vetch (Vicia sativa L.) seeds. *Eur J Biochem* **228**, 456-462
4. Julian, I., Gandullo, J., Santos-Silva, L. K., Diaz, I., and Martinez, M. (2013) Phylogenetically distant barley legumains have a role in both seed and vegetative tissues. *J Exp Bot* **64**, 2929-2941
5. Muntz, K., and Shutov, A. D. (2002) Legumains and their functions in plants. *Trends Plant Sci* **7**, 340-344
6. Nakaune, S., Yamada, K., Kondo, M., Kato, T., Tabata, S., Nishimura, M., and Hara-Nishimura, I. (2005) A vacuolar processing enzyme, deltaVPE, is involved in seed coat formation at the early stage of seed development. *The Plant cell* **17**, 876-887
7. Poncet, V., Scutt, C., Tournebize, R., Villegente, M., Cueff, G., Rajjou, L., Balliau, T., Zivy, M., Fogliani, B., Job, C., de Kochko, A., Sarramegna-Burtet, V., and Job, D. (2015) The Amborella vacuolar processing enzyme family. *Front Plant Sci* **6**, 618
8. Yamada, K., Basak, A. K., Goto-Yamada, S., Tarnawska-Glatt, K., and Hara-Nishimura, I. (2020) Vacuolar processing enzymes in the plant life cycle. *New Phytol* **226**, 21-31
9. Hatsugai, N., Yamada, K., Goto-Yamada, S., and Hara-Nishimura, I. (2015) Vacuolar processing enzyme in plant programmed cell death. *Front Plant Sci* **6**, 234
10. Shimada, T., Yamada, K., Kataoka, M., Nakaune, S., Koumoto, Y., Kuroyanagi, M., Tabata, S., Kato, T., Shinozaki, K., Seki, M., Kobayashi, M., Kondo, M., Nishimura, M., and Hara-Nishimura, I. (2003) Vacuolar processing enzymes are essential for proper processing of seed storage proteins in Arabidopsis thaliana. *The Journal of biological chemistry* **278**, 32292-32299
11. Shimada, T., Hiraiwa, N., Nishimura, M., and Hara-Nishimura, I. (1994) Vacuolar processing enzyme of soybean that converts proproteins to the corresponding mature forms. *Plant Cell Physiol* **35**, 713-718
12. Gruis, D., Schulze, J., and Jung, R. (2004) Storage protein accumulation in the absence of the vacuolar processing enzyme family of cysteine proteases. *The Plant cell* **16**, 270-290
13. Tiedemann, J., Schlereth, A., and Muntz, K. (2001) Differential tissue-specific expression of cysteine proteinases forms the basis for the fine-tuned mobilization of storage globulin during and after germination in legume seeds. *Planta* **212**, 728-738
14. Conlan, B. F., Gillon, A. D., Craik, D. J., and Anderson, M. A. (2010) Circular proteins and mechanisms of cyclization. *Biopolymers*
15. Craik, D. J., and Malik, U. (2013) Cyclotide biosynthesis. *Curr Opin Chem Biol* **17**, 546-554
16. Nguyen, G. K., 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
17. 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. *Chemistry & biology* **22**, 571-582
18. 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. *Nature communications* **6**, 10199
19. Saska, I., Gillon, A. D., Hatsugai, N., Dietzgen, R. G., Hara-Nishimura, I., Anderson, M. A., and Craik, D. J. (2007) An asparaginyl endopeptidase mediates in vivo protein backbone cyclization. *The Journal of biological chemistry* **282**, 29721-29728
20. 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

21. Zauner, F. B., Dall, E., Regl, C., Grassi, L., Huber, C. G., Cabrele, C., and Brandstetter, H. (2018) Crystal Structure of Plant Legumain Reveals a Unique Two-Chain State with pH-Dependent Activity Regulation. *The Plant cell* **30**, 686-699
22. 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
23. Craik, D. J. (2012) Host-defense activities of cyclotides. *Toxins (Basel)* **4**, 139-156
24. Gillon, A. D., Saska, I., Jennings, C. V., Guarino, R. F., Craik, D. J., and Anderson, M. A. (2008) Biosynthesis of circular proteins in plants. *The Plant journal : for cell and molecular biology* **53**, 505-515
25. 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
26. Lesner, A., Legowska, A., Wysocka, M., and Rolka, K. (2011) Sunflower trypsin inhibitor 1 as a molecular scaffold for drug discovery. *Curr Pharm Des* **17**, 4308-4317
27. Ireland, D. C., Wang, C. K., Wilson, J. A., Gustafson, K. R., and Craik, D. J. (2008) Cyclotides as natural anti-HIV agents. *Biopolymers* **90**, 51-60
28. Jackson, M. A., Gilding, E. K., Shafee, T., Harris, K. S., Kaas, Q., Poon, S., Yap, K., Jia, H., Guarino, R., Chan, L. Y., Durek, T., Anderson, M. A., and Craik, D. J. (2018) Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. *Nature communications* **9**, 2411
29. Zauner, F. B., Elsasser, B., Dall, E., Cabrele, C., and Brandstetter, H. (2018) Structural analyses of Arabidopsis thaliana legumain gamma reveal differential recognition and processing of proteolysis and ligation substrates. *The Journal of biological chemistry* **293**, 8934-8946
30. Hemu, X., El Sahili, A., Hu, S., Wong, K., Chen, Y., Wong, Y. H., Zhang, X., Serra, A., Goh, B. C., Darwis, D. A., Chen, M. W., Sze, S. K., Liu, C. F., Lescar, J., and Tam, J. P. (2019) Structural determinants for peptide-bond formation by asparaginyl ligases. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 11737-11746
31. Dall, E., and Brandstetter, H. (2012) Activation of legumain involves proteolytic and conformational events, resulting in a context- and substrate-dependent activity profile. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **68**, 24-31
32. Dall, E., and Brandstetter, H. (2013) Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10940-10945
33. 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
34. James, A. M., Haywood, J., Leroux, J., Ignasiak, K., Elliott, A. G., Schmidberger, J. W., Fisher, M. F., Nonis, S. G., Fenske, R., Bond, C. S., and Mylne, J. S. (2019) The macrocyclizing protease butelase 1 remains autocatalytic and reveals the structural basis for ligase activity. *The Plant journal : for cell and molecular biology* **98**, 988-999
35. Fuentes-Prior, P., and Salvesen, G. S. (2004) The protein structures that shape caspase activity, specificity, activation and inhibition. *Biochem J* **384**, 201-232
36. Stewart, D. E., Sarkar, A., and Wampler, J. E. (1990) Occurrence and role of cis peptide bonds in protein structures. *Journal of molecular biology* **214**, 253-260
37. Schilling, O., Huesgen, P. F., Barre, O., Auf dem Keller, U., and Overall, C. M. (2011) Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nature protocols* **6**, 111-120
38. Biniossek, M. L., Niemer, M., Maksimchuk, K., Mayer, B., Fuchs, J., Huesgen, P. F., McCafferty, D. G., Turk, B., Fritz, G., Mayer, J., Haecker, G., Mach, L., and Schilling, O. (2016) Identification of Protease Specificity by Combining Proteome-Derived Peptide Libraries and Quantitative Proteomics. *Molecular & cellular proteomics : MCP* **15**, 2515-2524

39. Vidmar, R., Vizovisek, M., Turk, D., Turk, B., and Fonovic, M. (2017) Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *EMBO J* **36**, 2455-2465
40. Weng, S. S. H., Demir, F., Ergin, E. K., Dirnberger, S., Uzozie, A., Tuscher, D., Nierves, L., Tsui, J., Huesgen, P. F., and Lange, P. F. (2019) Sensitive Determination of Proteolytic Proteoforms in Limited Microscale Proteome Samples. *Molecular & cellular proteomics : MCP* **18**, 2335-2347
41. 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 Engl* **54**, 2917-2921
42. Wang, J., and Wilkinson, M. F. (2001) Deletion mutagenesis of large (12-kb) plasmids by a one-step PCR protocol. *Biotechniques* **31**, 722-724
43. Kabsch, W. (2010) Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132
44. 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., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., 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
45. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674
46. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132
47. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., and Terwilliger, T. C. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58**, 1948-1954
48. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037-10041
49. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**, W665-667
50. Dall, E., Hollerweger, J. C., Dahms, S. O., Cui, H., Haussermann, K., and Brandstetter, H. (2018) Structural and functional analysis of cystatin E reveals enzymologically relevant dimer and amyloid fibril states. *The Journal of biological chemistry* **293**, 13151-13165
51. Niesen, F. (2010) Excel script for the analysis of protein unfolding data acquired by Differential Scanning Fluorimetry (DSF). (Niesen, F. ed., 3.0 Ed., Structural Genomics Consortium, Oxford
52. Dahms, S. O., Demir, F., Huesgen, P. F., Thorn, K., and Brandstetter, H. (2019) Sirtilins - the new old members of the vitamin K-dependent coagulation factor family. *J Thromb Haemost* **17**, 470-481
53. 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. *The Journal of biological chemistry* **280**, 32914-32920
54. Wessel, D., and Flugge, U. I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical biochemistry* **138**, 141-143
55. Rinschen, M. M., Hoppe, A. K., Grahammer, F., Kann, M., Volker, L. A., Schurek, E. M., Binz, J., Hohne, M., Demir, F., Malisic, M., Huber, T. B., Kurschat, C., Kizhakkedathu, J. N., Schermer, B., Huesgen, P. F., and Benzing, T. (2017) N-Degradomic Analysis Reveals a Proteolytic Network Processing the Podocyte Cytoskeleton. *J Am Soc Nephrol* **28**, 2867-2878
56. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319
57. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **6**, 786-787
58. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447-456

59. Karplus, P. A., and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* **336**, 1030-1033

FOOTNOTES

The abbreviations used are: proAtLEGβ, *Arabidopsis thaliana* prolegumain isoform beta; AtLEGβ, *Arabidopsis thaliana* legumain isoform beta; AEP, asparaginyl endopeptidase; LSAM, Legumain Stabilization and Activity Modulation domain; AP, activation peptide; VPE, vacuolar processing enzyme; SFTI, sunflower trypsin inhibitor; cPRM, cyclic protein recognition motif; rmsd, root mean square deviation

Table 1. Xray data collection and refinement statistics

	proAtLEGβ (6ysa)
Data collection	
Space group	$P4_1$
Cell dimensions	
a=*b*, *c* (Å)	170.4, 196.5
Resolution (Å)[a]	49.6–2.0 (2.04–2.01)
R_{merge}	0.12 (1.42)
R_{pim}	0.08 (0.99)
CC (1/2) (%)	0.99 (0.22)
$I/\sigma I$	6.8 (0.7)
Completeness (%)	90.2 (86.3)
Redundancy	2.8 (2.6)
Refinement	
Resolution (Å)	49.6–2.0
No. unique reflections	336594
R_{work} / R_{free}	20.8/21.8
No. atoms	
Protein	39124
Ligand/ion	763
Water	2254
Overall B-factor (Å^2)	36.0
R.m.s deviations	
Bond lengths (Å)	0.01
Bond angles (°)	1.15
Ramachandran plot	
No. outliers (%)	0.0
No. favored (%)	97.9

The structure was determined from a single crystal.

[a] Highest resolution shell is shown in parentheses.

The resolution cutoff was set by applying the CC1/2 criterion (59).

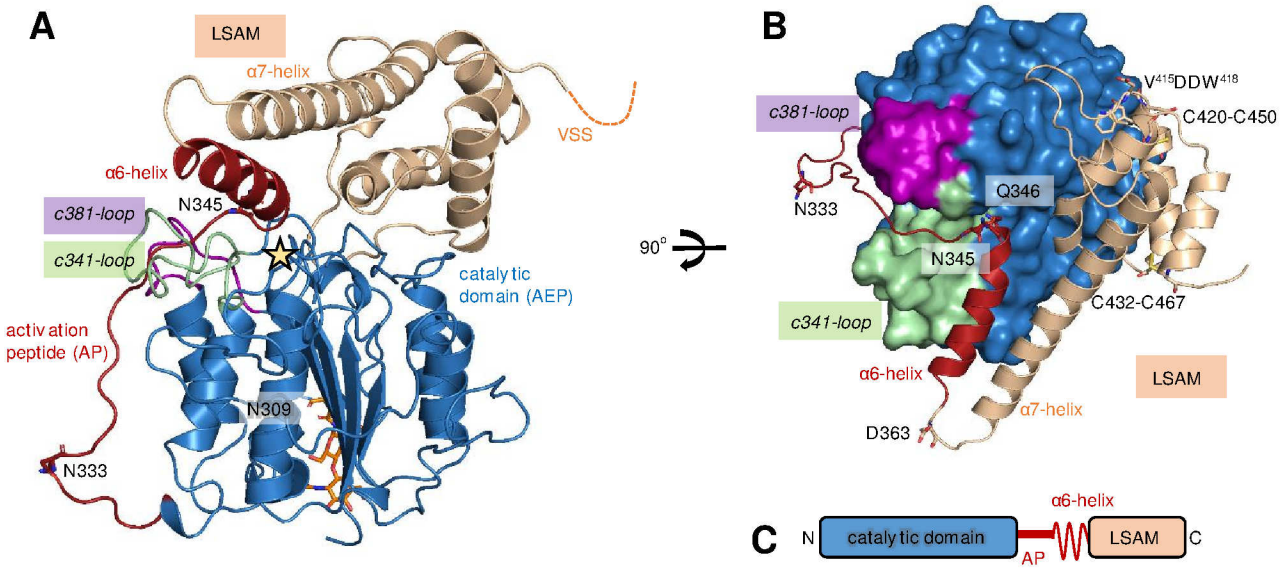

Figure 1. proAtLEGβ shares the typical prolegumain-like architecture. A. Cartoon representation of proAtLEGβ with the catalytic AEP domain shown in blue, the activation peptide harbouring the α6-helix in red and the LSAM domain in beige. Asn333 and 345 autocatalytic cleavage sites and the Asn309 glycosylation site are indicated as sticks, an asterisk is labeling the active site, the C-terminal vacuolar sorting signal (VSS) is indicated by a dashed line. C341- and c381-specificity loops are colored green and purple respectively. **B.** Top-view on the active site in standard orientation (substrate binding from left to right). Gln346 (red sticks) on the AP binds to the S1-pocket. Disulfide bonds on the LSAM domain are shown as sticks. The autocatalytic processing sites Asp363 and Asp416 (within the $V^{415}DDW^{418}$ motif) are indicated. **C.** Schematic representation of proAtLEGβ domain architecture.

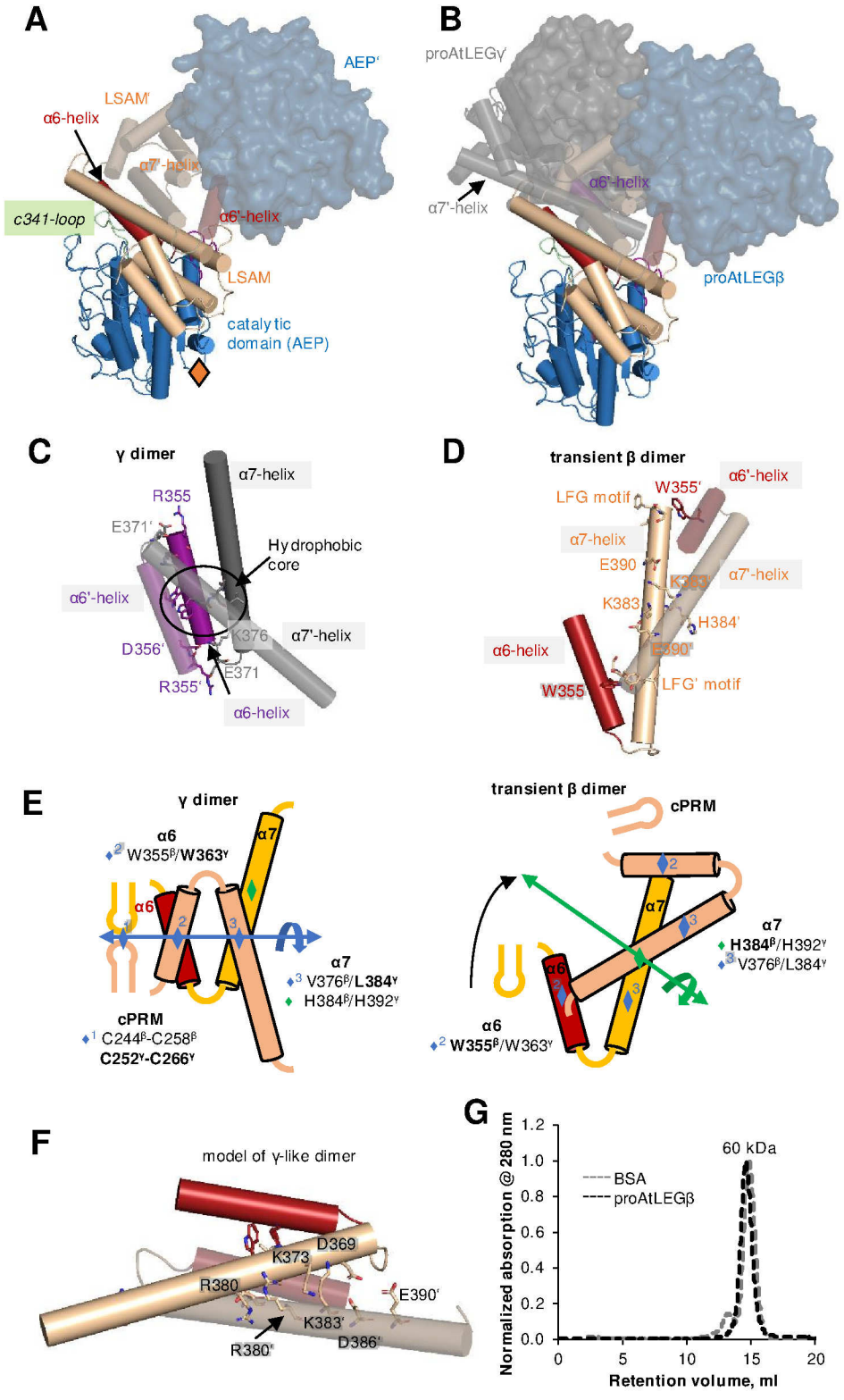

Figure 2. proAtLEGβ is monomeric in solution. A. Crystal packing induced proAtLEGβ dimerization. Monomer 1 is shown in cartoon representation, monomer 2 is labeled with a prime symbol (AEP' in surface representation). Interactions were mainly mediated by 2 symmetric salt bridges on the α7-helices (LSAM domain). The location of the Asn309 glycosylation site is indicated with an orange diamond. **B.**

Superposition of A) (proAtLEGβ dimer observed in the crystals) with dimeric two-chain AtLEGγ (pdb entry 5nij). Dimerization led to different spatial orientation of the AEP domains. **C.** Zoom-in view on the 4-helix bundle as observed in two-chain AtLEGγ. Interaction is mediated by a hydrophobic core that is surrounded by electrostatic interactions. **D.** Zoom-in view on the 4-helix bundle observed in proAtLEGβ. Interaction is mediated by a symmetric E390 – K383 salt bridge localized on the α7-helix and hydrophobic interactions between the LFG motif ($Leu396^{\beta}$-$Gly398^{\beta}$) on α7 helix and W355' on the α6' helix. Relative to B), the views in C) and D) are rotated by 90° along the y-axis. **E.** Schematic representation of the 4-helix bundle as observed in AtLEGγ and β. **F.** Model of an AtLEGγ-like dimerization mode in proAtLEGβ. AtLEGγ-like dimerization is not favored because of electrostatic repulsion of R380 – R380', K373 – K383' and D369 – D386' – E390' pairs. **G.** Size exclusion runs confirming monomeric state of proAtLEGβ. BSA served as a size standard.

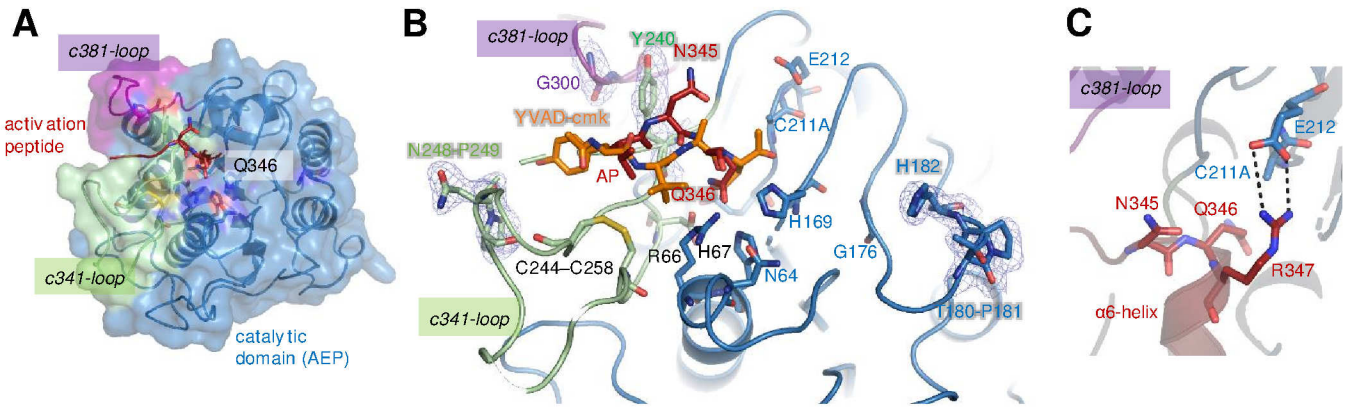

Figure 3. The activation peptide binds canonically to the active site. A. Top view on the active site of proAtLEGβ. The activation peptide (AP) harbouring the autocatalytic Asn345 cleavage site and Gln346 that is occupying the S1-pocket are shown in red. **B.** Zoom-in view on the non-primed and primed substrate binding sites with a YVAD-cmk peptide modeled based on the crystal structure of the YVAD – AtLEGγ complex (pdb entry 5obt). cis-imide peptide bonds (Thr180-Pro181 and Asn248-Pro249) are shown as sticks. For selected residues a 2Fo-Fc composite omit map is displayed at a contour level of 1 σ. **C.** Zoom-in view on the active site of proAtLEGβ. The ionic clamp (R347–E312) that is a linking the α6-helix to the active site is indicated.

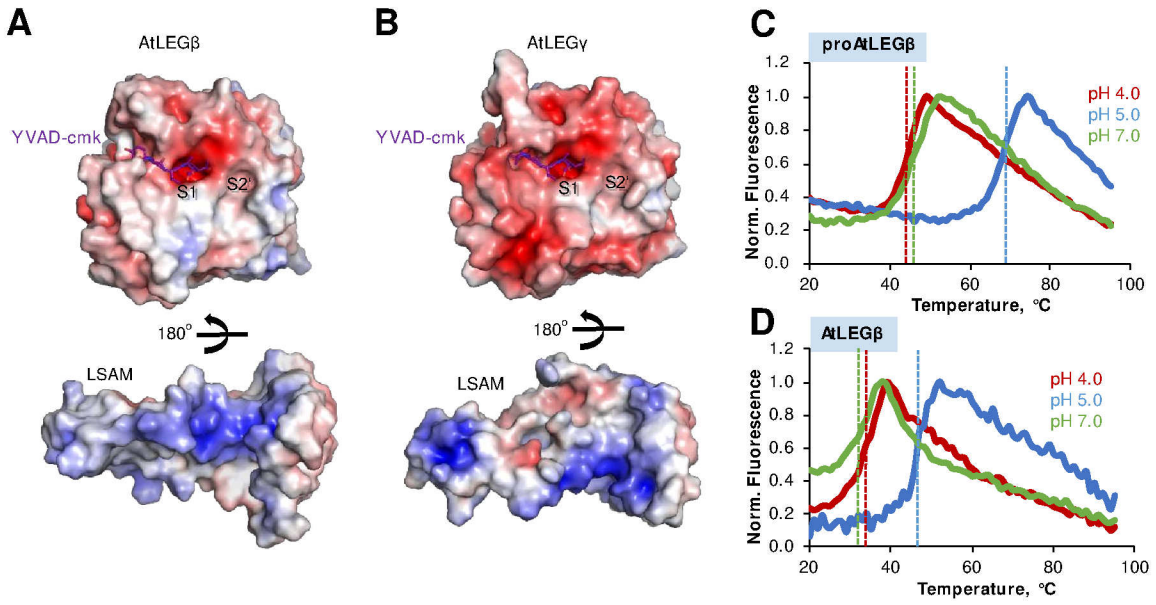

Figure 4. The AEP – LSAM interaction in proAtLEGβ is mostly hydrophobic. A. Color-coded electrostatic surface potential of AtLEGβ AEP and LSAM domains based on the crystal structure of proAtLEGβ (blue: positive charge, red: negative charge) calculated at pH 7.0 and contoured at +/- 5 kT/e. The LSAM domain has been rotated by 180° relative to the AEP domain. The YVAD-cmk inhibitor has been modeled based on the crystals structure of the AtLEGγ inhibitor complex (pdb entry 5obt). **B.** same as A., but calculated for AtLEGγ in complex with YVAD-cmk inhibitor. **C.** Melting curves of proAtLEGβ at indicated pH values show highest thermal stability at pH 5. Melting points are indicated by dashed lines. **D.** Melting curves of active AtLEGβ showing highest stability at pH 5.0.

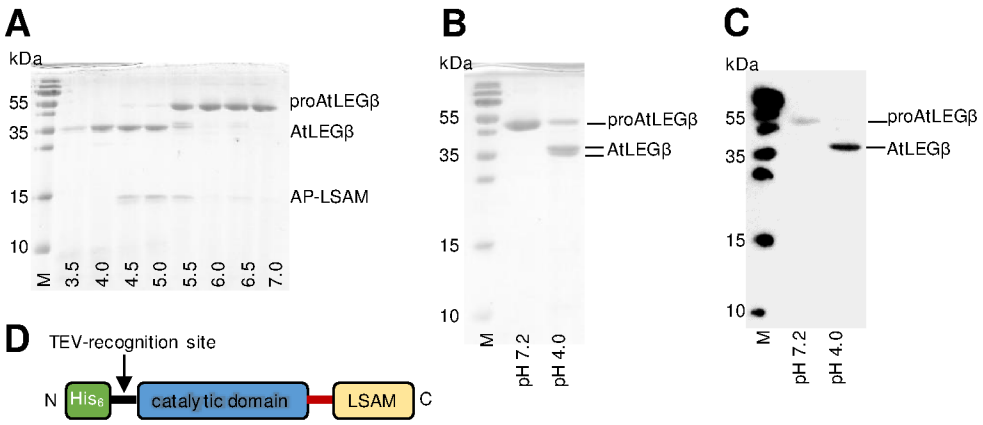

Figure 5. Auto-catalytic activation of AtLEGβ is pH-dependent and results in a two-chain intermediate state (pH 5.0) and active AEP state (pH 4.0). A. ProAtLEGβ after 1h incubation at indicated pH values. 'AtLEGβ' corresponds to the catalytic domain up to the autocatalytic cleavage site Asn333, and 'AP-LSAM' corresponds to the Gln346–Ala486 C-terminal fragment, that is generated by cleavage after Asn345). **B.** SDS-PAGE showing proAtLEGβ (pH 7.2) and AtLEGβ following activation at pH 4.0. Activation results in a double band at around 36 kDa. **C.** Western-blot using an Anti-His-HRP antibody, showing that only one AtLEGβ activation product harbours the N-terminal His_6-tag. **D.** Scheme illustrating the domain architecture of the recombinant expression construct.

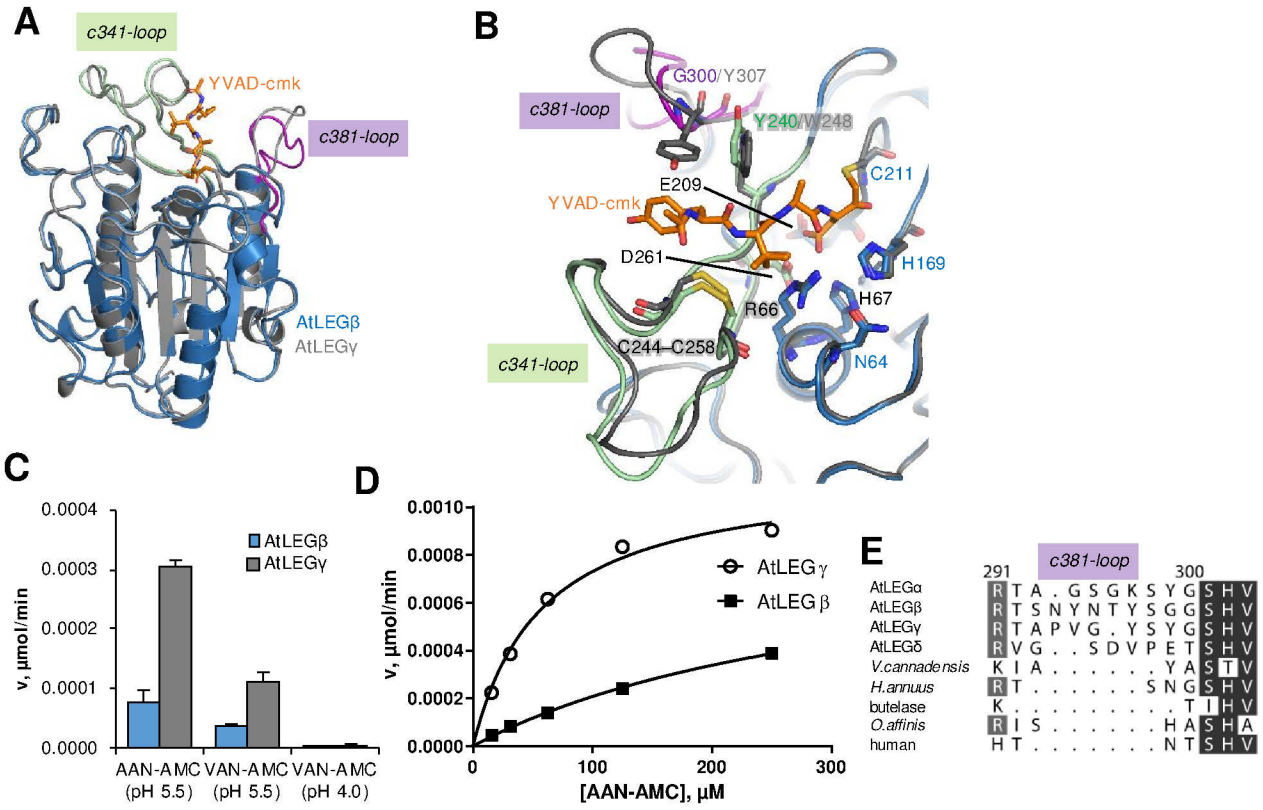

Figure 6. AtLEGs differ in their substrate specificity loops. A. Superposition of AtLEGβ (blue) and γ (grey) AEP domains. The YVAD-cmk inhibitor bound to AtLEGγ is shown in orange sticks, the c341-loop in green and the c381-loop in purple. **B.** Zoom in view on the active site. Catalytic residues are labeled in blue, residues forming the S1 specificity pocket in black. **C.** Catalytic activities of AtLEGβ and –γ towards peptidic AAN-AMC and VAN-AMC substrates at indicated pH values. **D.** K_M-determination for AtLEGβ and –γ towards the AAN-AMC substrate. **E.** Sequence alignment of the c381-loops of indicated (plant) species. Sequences were derived from structures deposited to the PDB, where applicable; AtLEGα (P49047), AtLEGβ (Q39044), AtLEGγ (5nij), AtLEGδ (Q9LJX8) *V.canadensis* (*Viola Canadensis*; 5zbi), *H.annuus* (*Helianthus annuus*; 6azt), butelase (*Clitoria ternatea*; 6dhi), *O.affinis* (*Oldenlandia affinis*; 5hoi).

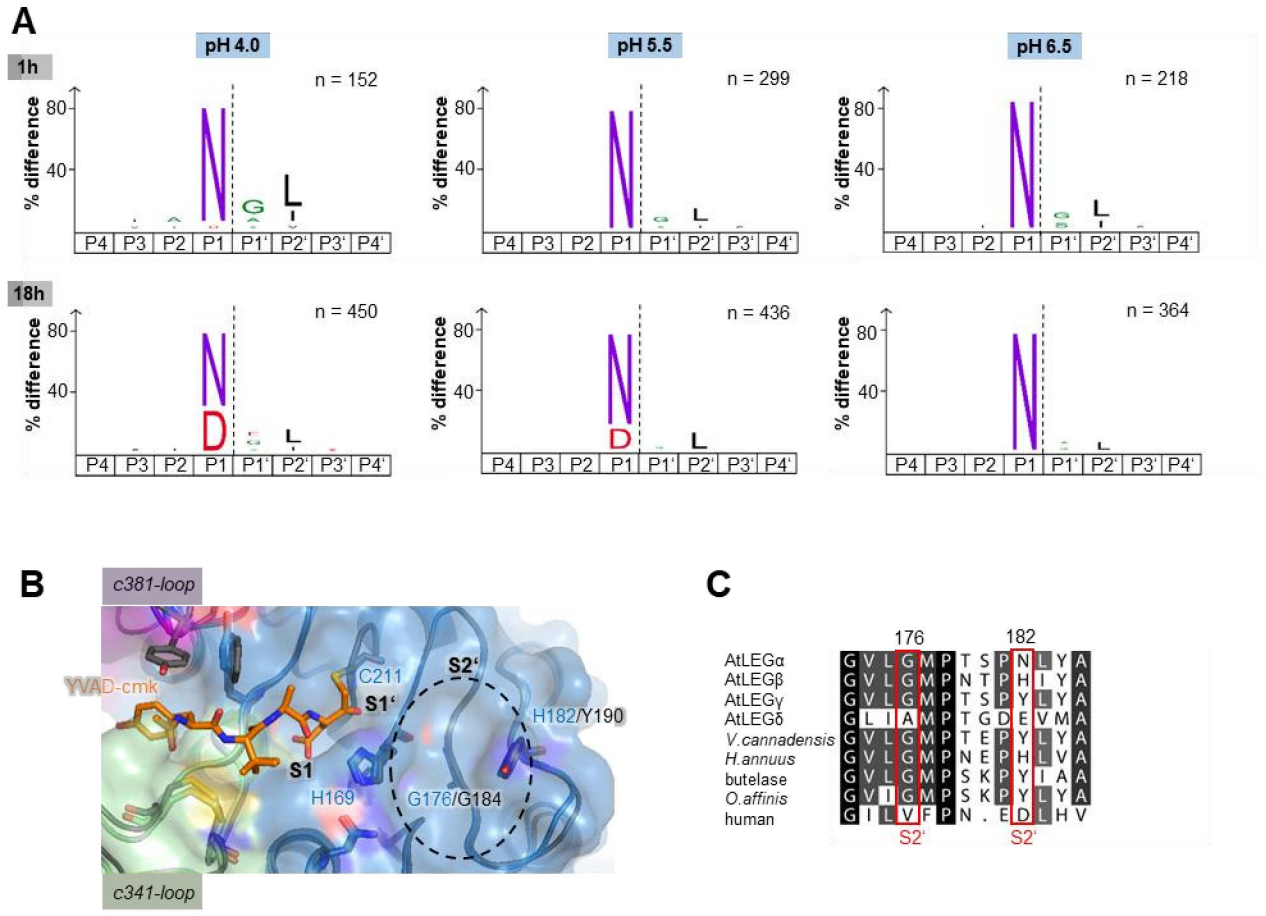

Figure 7. AtLEGβ has a pH-dependent substrate specificity. A. Cleavage site specificity determined by the PICS assay, using peptides generated by tryptic digest of an *E.coli* proteome as substrate library. iceLogos visualize the substrate preference surrounding the cleavage sites (p = 0.05) based on peptides cleaved by AtLEGβ after incubation at indicated pH values and times. Number of non-redundant cleavage sites used to generate the iceLogos are indicated. **B.** Top-view on the AtLEGβ substrate binding site. Binding of the YVAD-cmk inhibitor was modeled based on the crystal structure of the YVAD – AtLEGγ complex (pdb entry 5obt). **C.** Sequence alignment of the residues forming the prime-substrate binding site. Sequences used are the same as in Figure 5.

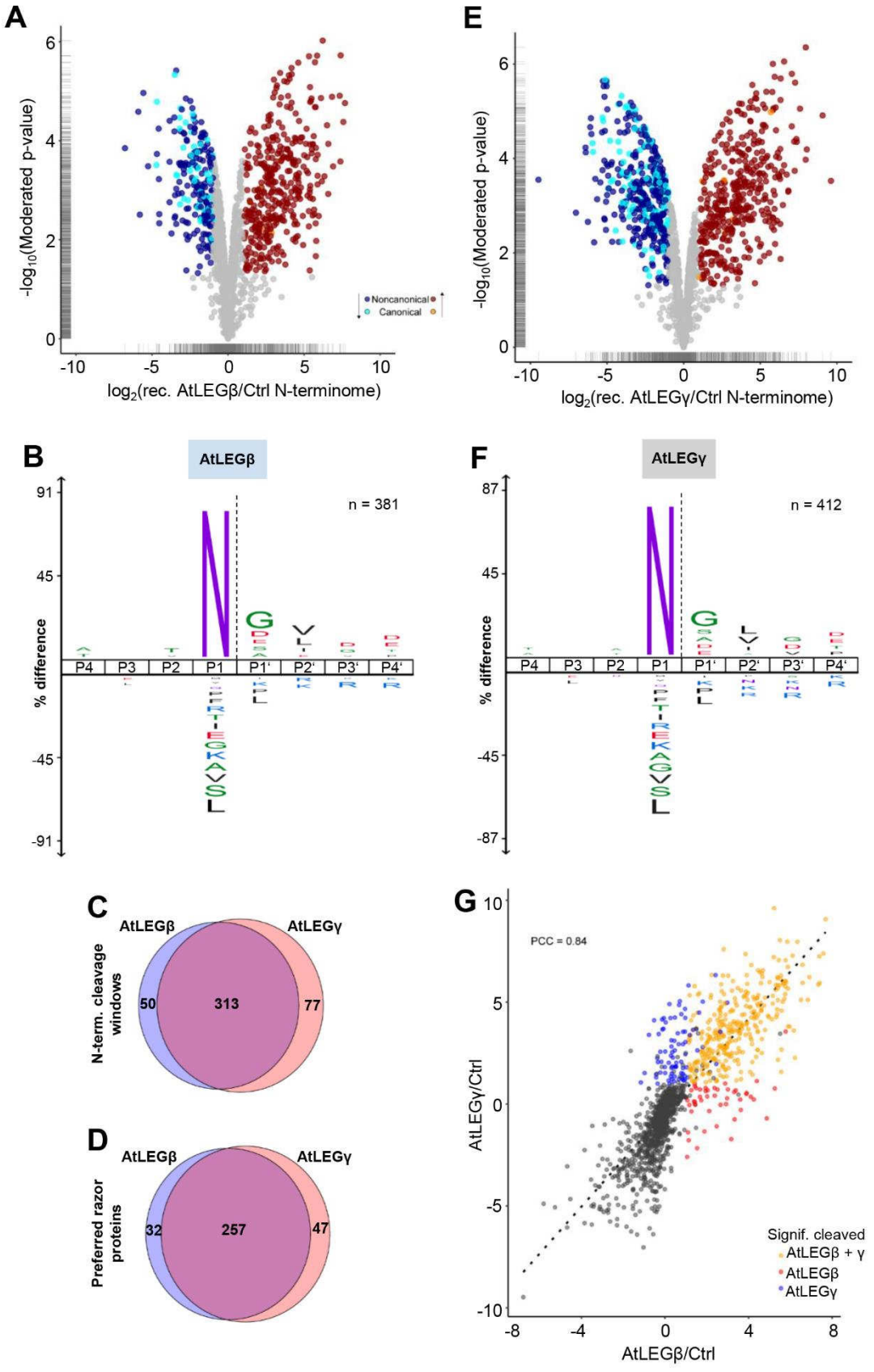
A
-log10(Moderated p-value)
log2(rec. AtLEGβ/Ctrl N-terminome)
Noncanonical
Canonical
E
-log10(Moderated p-value)
log2(rec. AtLEGγ/Ctrl N-terminome)
B
AtLEGβ
n = 381
% difference
P4 P3 P2 P1 P1' P2' P3' P4'
F
AtLEGγ
n = 412
% difference
P4 P3 P2 P1 P1' P2' P3' P4'
C
AtLEGβ
AtLEGγ
N-term. cleavage windows
50
313
77
D
AtLEGβ
AtLEGγ
Preferred razor proteins
32
257
47
G
PCC = 0.84
AtLEGγ/Ctrl
AtLEGβ/Ctrl
Signif. cleaved
AtLEGβ + γ
AtLEGβ
AtLEGγ

Figure 8. Substrate specificity of AtLEGβ and γ towards intact proteins extracted from *A. thaliana* leaves. Volcano plots identify protein N-terminal peptides significantly changing in abundance (greater 2-fold change in abundance supported by LIMMA-moderated t-test p-val <0.05) after *in vitro* incubation of *A. thaliana vpe0* proteome with recombinant **A**) AtLEGβ or **E**) AtLEGγ. Log_2 fold-change is the mean of 3 biological replicates. Accumulating N-terminal peptides indicative of AtLEGβ/γ cleavage are highlighted red, depleted peptides cleaved within their sequence in blue. iceLogos visualize the substrate preference surrounding the cleavage sites for **B**) AtLEGβ and **F**) AtLEGγ (p = 0.05). Number of non-redundant cleavage sites used to generate the iceLogos are indicated. Venn diagrams showing the overlap of **C**) cleavage sites and **D**) proteins cleaved in the *vpe0* proteome after incubation with AtLEGβ or γ. **G**) Correlation of N-terminal peptide abundance in both experiments (dimethylated N-terminal peptides quantified in at least 2 out of 3 replicates). Significantly accumulating dimethylated N-terminal peptides ($log_2FC > 1$, LIMMA-moderated t-test p-value < 0.05) indicate cleavage by AtLEGβ (red), AtLEGγ (blue), or both (orange). The linear fit confers a Pearson correlation coefficient (PCC) of 0.84 indicating a very high degree of overlap among the putative substrates.

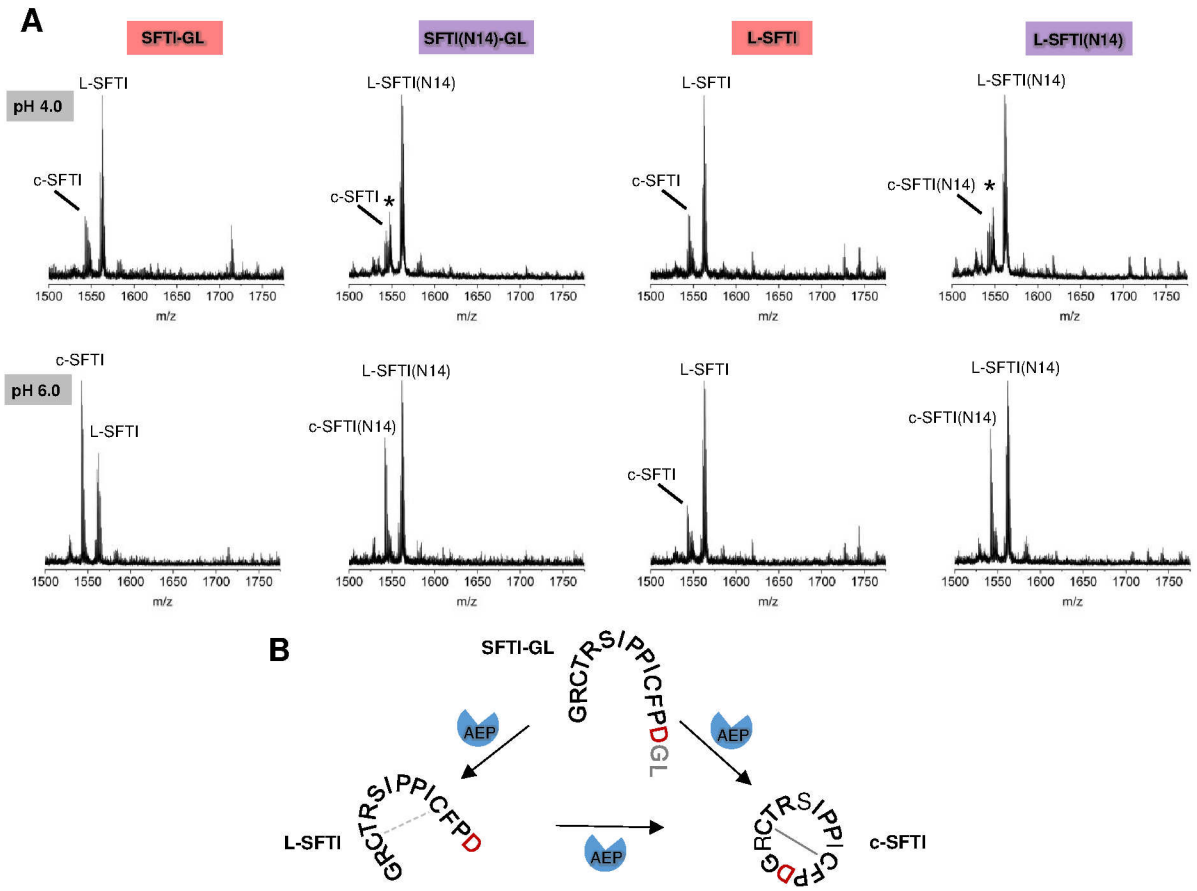

Figure 9. Cyclisation of SFTI-derived peptides by AtLEGβ is pH-dependent. A. Reactions were carried out at indicated pH values. An unidentified species is labeled with an asterisk. **B.** Reaction scheme of AtLEGβ catalyzed cyclisation of SFTI-GL peptides. The pre-cursor peptides SFTI-GL and SFTI(N14)-GL were synthesized in the reduced form, and were also observed mostly reduced in the assays. The linear L-SFTI and L-SFTI(N14) cleavage products were observed both in the reduced and oxidized forms, with the Cys3–Cys11 disulfide bond formed. c-SFTI and c-SFTI(N14) were mostly oxidized.

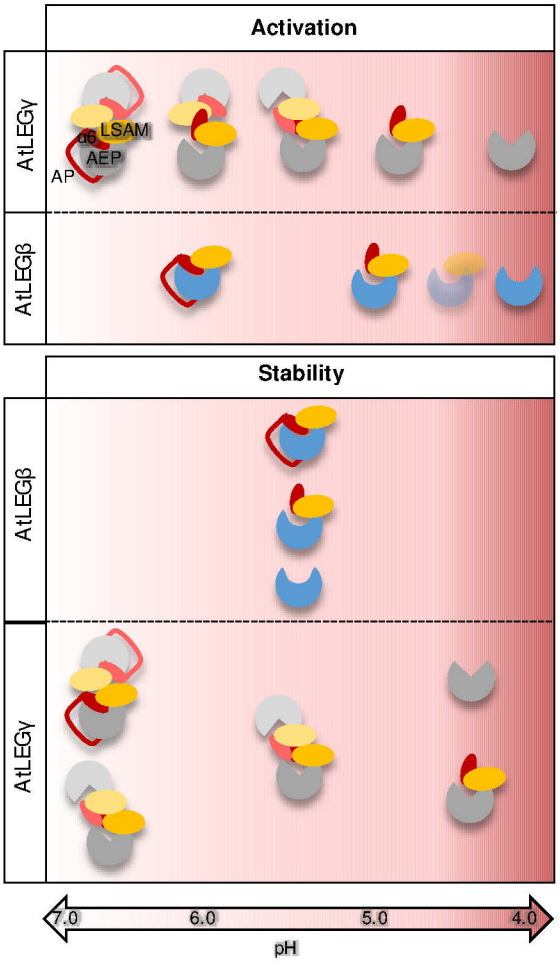

Figure 10. Activation and stability of AtLEGβ and AtLEGγ are pH dependent. In the zymogen forms of proAtLEGβ and γ the LSAM domain (orange) and activation peptide (AP, red line) that harbors the α6-helix (red ellipsoid) sit on top of the active site and thereby block access to the substrate binding sites. ProAtLEGγ forms a dimer at neutral to intermediate pH conditions, and is mostly present in its two-chain state, which is generated upon cleavage at the N-terminal end of the α6-helix. By lowering pH, the interaction of the α6-helix with the AEP domain gets weaker, as it is mainly mediated by electrostatic interactions. At pH < 4.5, the two-chain state will disassemble and thereby allow degradation of the α6-helix and the LSAM domain. By contrast, proAtLEGβ is a monomer in solution. Activation proceeds via cleavage after (1) Asn333/345 on the N-terminal end of the α6-helix at intermediate pH, followed by (2) multiple cleavages after aspartic acid residues at pH < 4.5, which finally result in AP-LSAM degradation. Activation likely proceeds via a short-lived intermediate state, that has the α6-helix selectively removed, but the LSAM domain still bound to the AEP domain (indicated by transparent coloring). While all AtLEGβ activation states show highest conformational stability at intermediate pH, two-chain (pro)AtLEGγ is stable at neutral to slightly acidic pH and monomeric two-chain AtLEGγ as well as the AEP domain are most stable at acidic pH.