



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

Received for publication, May 20, 2020, and in revised form, July 14, 2020. Published, Papers in Press, July 21, 2020. DOI 10.1074/jbc.RA120.014478

Elfriede Dall<sup>1,‡,\*</sup>, Florian B. Zauner<sup>1,‡</sup>, Wai Tuck Soh<sup>1</sup>, Fatih Demir<sup>2</sup>, Sven O. Dahms<sup>1</sup>, Chiara Cabrele<sup>1</sup>, Pitter F. Huesgen<sup>2,3,4</sup>, and Hans Brandstetter<sup>1,\*</sup>

From the <sup>1</sup>Department of Biosciences, University of Salzburg, Salzburg, Austria, <sup>2</sup>Central Institute for Engineering, Electronics and Analytics, ZEA-3, Forschungszentrum Jülich, Germany, <sup>3</sup>CECAD, Medical Faculty and University Hospital, University of Cologne, Cologne, Germany, and <sup>4</sup>Institute for Biochemistry, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany

Edited by Joseph M. Jez

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  $\beta$  (AtLEG $\beta$ ) in its zymogen state. Combining structural and biochemical experiments, we show for the first time that plant legumains encode distinct, isoform-specific activation mechanisms. Whereas the autocatalytic activation of isoform  $\gamma$  (AtLEG $\gamma$ ) is controlled by the latency-conferring dimer state, the activation of the monomeric AtLEG $\beta$  is concentration independent. Additionally, in AtLEG $\beta$  the plant-characteristic two-chain intermediate state is stabilized by hydrophobic rather than ionic interactions, as in AtLEG $\gamma$ , resulting in significantly different pH stability profiles. The crystal structure of AtLEG $\beta$  revealed unrestricted nonprime substrate binding pockets, consistent with the broad substrate specificity, as determined by degradomic assays. Further to its protease activity, we show that AtLEG $\beta$  exhibits a true peptide ligase activity. Whereas cleavage-dependent transpeptidase activity has been reported for other plant legumains, AtLEG $\beta$  is the first example of a plant legumain capable of linking free termini. The discovery of these isoform-specific differences will allow us to identify and rationally design efficient ligases with application in biotechnology and drug development.

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, therefore, are also referred to as vacuolar processing enzymes (VPEs) (1). Plant legumains are structurally related to the mam-

malian 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 ( $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ -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 ( $\beta$ -VPE) and nonseed or vegetative type VPEs ( $\alpha$ -,  $\gamma$ - and  $\delta$ -VPE) (5–8). Vegetative legumains are found in lytic vacuoles and have been suggested to play critical roles in plant programmed cell death and may functionally substitute for the caspases, which are absent in plants (9). Seed type legumains like *Arabidopsis thaliana* legumain isoform  $\beta$  (AtLEG $\beta$ ) 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 ( $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ ), which were shown to accumulate aberrantly processed seed storage proteins (12). Importantly, AtLEG $\beta$  can compensate for missing vegetative  $\alpha$  and  $\gamma$  proteins, further confirming that AtLEG $\beta$  is the main player in precursor protein processing in seeds (10). Known substrates of AtLEG $\beta$  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 showed that the vegetative type AtLEG $\gamma$  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 AtLEG $\beta$ .

Cyclic peptides are important for plant 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 cyclization in plants is typically catalyzed by legumains. Consequently, there is a high interest in understanding

This article contains [supporting information](#).

✂ Author's Choice—Final version open access under the terms of the Creative Commons CC-BY license.

‡ These authors contributed equally to this work.

\* For correspondence: Elfriede Dall, [elfriede.dall@sbg.ac.at](mailto:elfriede.dall@sbg.ac.at); Hans Brandstetter, [hans.brandstetter@sbg.ac.at](mailto:hans.brandstetter@sbg.ac.at).

## Structure and function of AtLEG $\beta$

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 us to predict ligase activity based on sequence information (20, 28). However, 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 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  $\alpha$ 6-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 proAtLEG $\gamma$  (*Arabidopsis thaliana* prolegumain isoform  $\gamma$ ) significantly differs from that of human legumain (31, 32). Importantly, proAtLEG $\gamma$  is present in an enzymatically latent dimer state that is mediated by AP-LSAM-AP'-LSAM' interactions and depends 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 AtLEG $\gamma$  results from cleavage at the N-terminal side of the  $\alpha$ 6-helix within the AP and is suppressed by high protein concentration, where AtLEG $\gamma$  dimerization is favored. Even after an initial cleavage within the AP, an enzymatically latent, dimeric two-chain AtLEG $\gamma$  intermediate form remains stable at neutral pH. Only at acidic pH does the dimer dissociate 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 the development of a pH-dependent four-step activation model of plant legumains, *i.e.* single chain–two chain conversion,  $\alpha$ 6-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 proAtLEG $\beta$ , which led to the discovery of a distinct activation mechanism, in contrast to AtLEG $\gamma$ . 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 AtLEG $\beta$  is an active ligase capable of peptide cyclization. AtLEG $\beta$  ligase activity is not strictly linked to peptide bond cleavage but enables the efficient joining of free N and C termini. To our knowledge, AtLEG $\beta$  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.

**Table 1**  
X-ray data collection and refinement statistics<sup>b</sup>

Parameter	Value(s) for proAtLEG $\beta$ (GySa)
Data collection	
Space group	<i>P</i> 4 <sub>1</sub>
Cell dimensions	
$a = b, c$ (Å)	170.4, 196.5
Resolution (Å) <sup>a</sup>	49.6–2.0 (2.04–2.01)
$R_{\text{merge}}$	0.12 (1.42)
$R_{\text{pim}}$	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. of unique reflections	336,594
$R_{\text{work}}/R_{\text{free}}$	20.8/21.8
No. atoms	
Protein	39,124
Ligand/ion	763
Water	2254
Overall B-factor (Å <sup>2</sup> )	36.0
RMSD	
Bond length (Å)	0.01
Bond angle (°)	1.15
Ramachandran plot	
No. of outliers (%)	0.0
No. favored (%)	97.9

<sup>a</sup>Highest-resolution shell is shown in parentheses.

<sup>b</sup>The structure was determined from a single crystal. The resolution cutoff was set by applying the CC1/2 criterion (59).

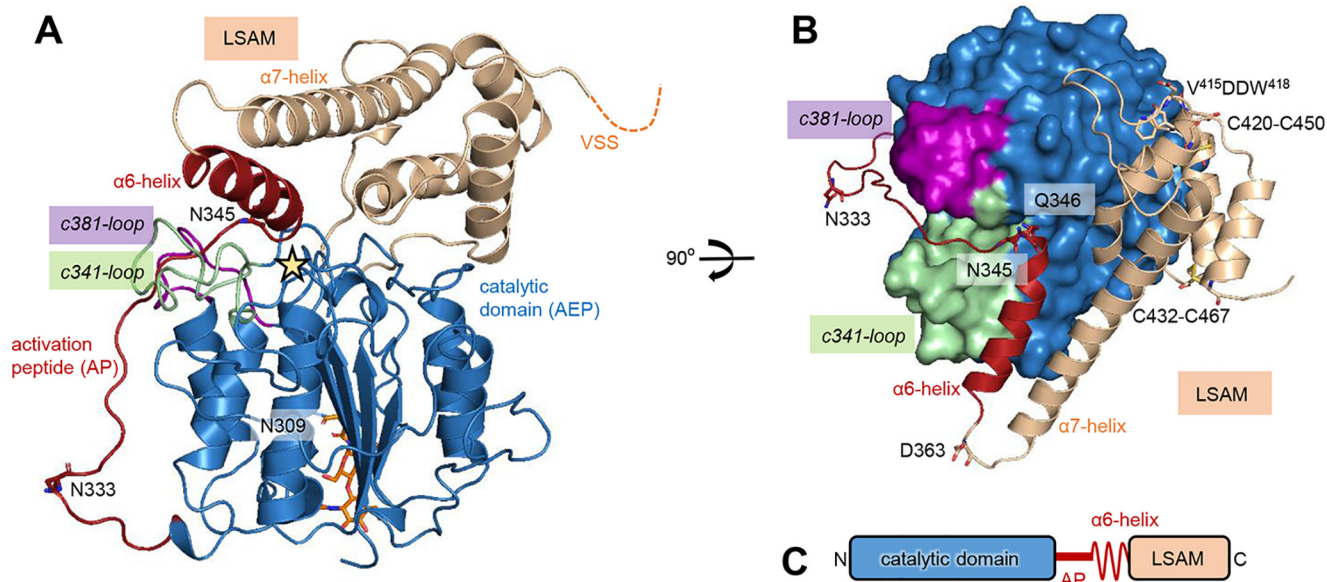
## Results

### Crystal structure of proAtLEG $\beta$

To understand isoform-specific differences between different AtLEGs, we determined the crystal structure of seed-type proAtLEG $\beta$  to a resolution of 2.0 Å (Table 1). The asymmetric unit of the tetragonal space group contained 12 independent molecules. Like isoform  $\gamma$ , proAtLEG $\beta$  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 Fig. S1). The AEP and LSAM domain are connected by an activation peptide that harbors the  $\alpha$ 6-helix. Overall, the structure of proAtLEG $\beta$  closely resembles the structure of the homologous two-chain AtLEG $\gamma$  indicated by a  $C\alpha$  root mean square deviation (RMSD) of 0.49 Å. However, inspecting the individual subdomains unraveled specific differences. Whereas the catalytic AEP domains of AtLEG $\beta$  and  $\gamma$  superimpose very well with an overall  $C\alpha$  RMSD of 0.39 Å, we observed bigger differences in the LSAM domains with a  $C\alpha$  RMSD of 0.78 Å (determined with Pymol). This observation is also in agreement with a higher sequence identity of the  $\beta$  and  $\gamma$  catalytic domains (67% identity) compared with 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).

### proAtLEG $\beta$ forms atypical dimers in the crystal and is monomeric in solution

An important feature of proAtLEG $\gamma$  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



**Figure 1. proAtLEG $\beta$  shares the typical prolegumain-like architecture.** *A*, cartoon representation of proAtLEG $\beta$  with the catalytic AEP domain shown in blue, the activation peptide harboring the  $\alpha 6$ -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, and 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 of 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<sup>413</sup>DDW<sup>418</sup> motif) are indicated. *C*, schematic representation of proAtLEG $\beta$  domain architecture.

activation and activity of AtLEG $\gamma$  (21). Similarly, in the crystal structure of proAtLEG $\beta$ , 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. 2*A*). However, these interactions were mediated by different amino acids and led to an  $\sim 90^\circ$  tilted orientation of the monomers within the  $\beta$ - and  $\gamma$ -dimer, respectively (Fig. 2, *B* and *C*). Indeed, detailed analyses of the  $\beta$  and  $\gamma$  dimer interfaces revealed significant, isoform-specific differences. The proAtLEG $\gamma$  dimer is mediated primarily by three symmetric anchoring sites,  $\alpha 6$  and  $\alpha 7$  helices, and a conserved cyclic protein recognition motif (cPRM) on the c341-loop. The  $\alpha 6$  and  $\alpha 7$  helices form a 4-helix bundle that is stabilized around a symmetric hydrophobic core formed by W363 $^\gamma$  as well as Val383 $^\gamma$  and L384 $^\gamma$ , respectively (AtLEG $\gamma$  numbering; Fig. 2, *C* and *E*). This hydrophobic core is further stabilized by a network of salt bridges on the N-terminal (R355 $^\gamma$ –E371 $^\gamma$ ' and D356 $^\gamma$ –K376 $^\gamma$ ') and C-terminal (K376 $^\gamma$ –D356 $^\gamma$ ' and E371 $^\gamma$ –R355 $^\gamma$ ') ends of the  $\alpha 6$ -helices. In contrast, the proAtLEG $\beta$  dimers in the crystal structure were predominantly mediated by the  $\alpha 7$  helix. This interaction was formed around the symmetric H384 $^\beta$  (H392 $^\beta$ ) and further stabilized by one symmetric salt bridge (E390 $^\beta$ –K383 $^\beta$ ') as well as by a hydrophobic contact of the  $\alpha 7$  C-terminal LFG motif (396 $^\beta$ –398 $^\beta$ ) with W355 $^\beta$ ' centered in the  $\alpha 6$  helix (Fig. 2, *D* and *E*). The hydrophobic core of the  $\alpha 6$ - $\alpha 7$ ,  $\alpha 6'$ - $\alpha 7'$  four-helix bundle was missing, as was any stabilization by the conserved cPRM, despite key residues important for proAtLEG $\gamma$ -like dimer formation being conserved in proAtLEG $\beta$  (Fig. S1). However, modeling a proAtLEG $\gamma$ -like dimer uncovered repulsive charge densities of  $\alpha 7$ - $\alpha 7'$  helix contact residues in AtLEG $\beta$  (R380–R380', K373–K383', and D369–D386') that will prohibit this  $\gamma$ -mode of dimerization

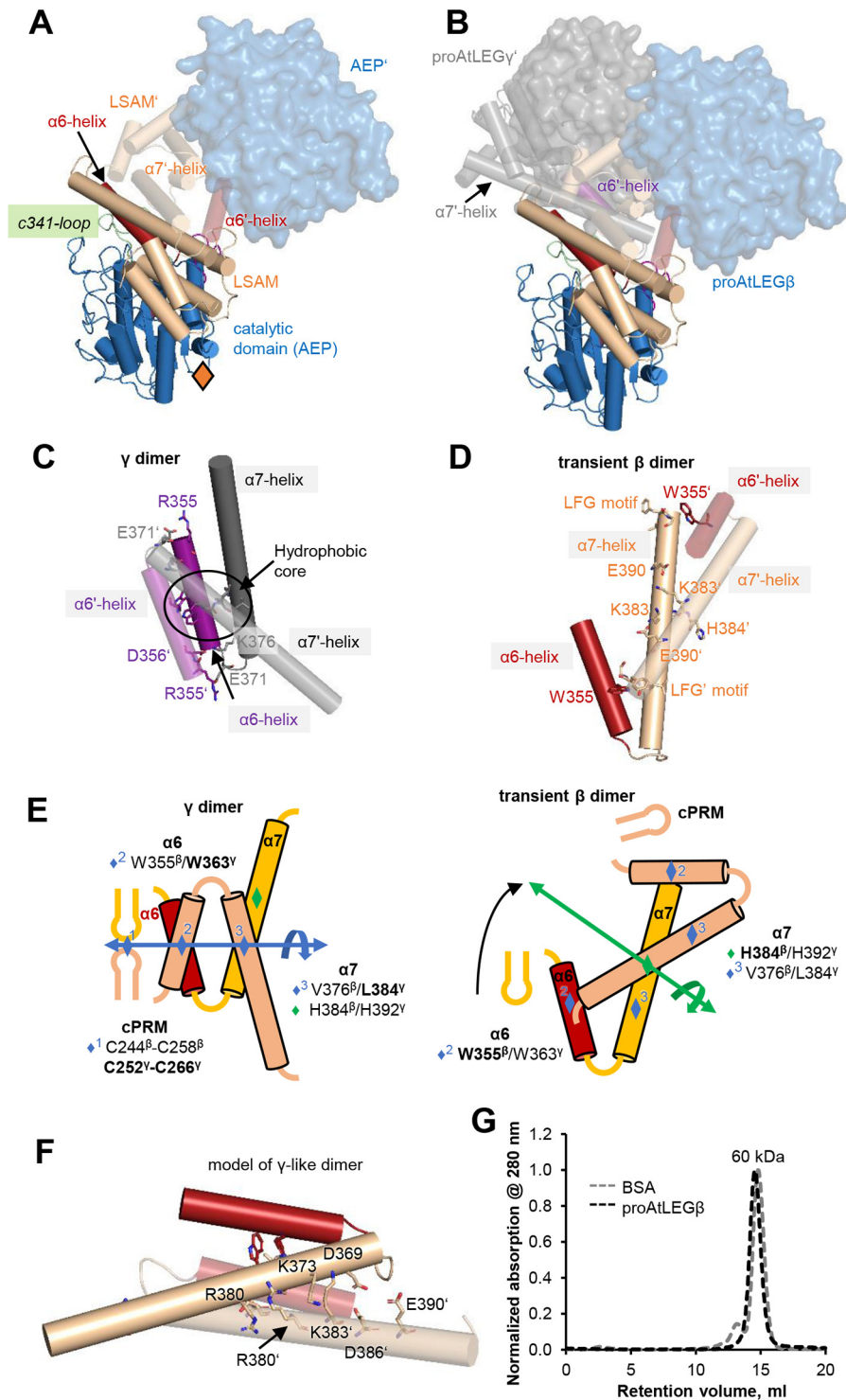
(Fig. 2*F* and Fig. S2). Together, these findings suggest that the observed  $\beta$ -dimer is weak and probably only transient in solution. To test this conclusion, we performed size exclusion chromatography (SEC) experiments. As expected, at pH 7.0, proAtLEG $\beta$  migrated at the expected size of a monomer, similar to human legumain (Fig. 2*F*). Accordingly, proAtLEG $\beta$  was a monomer in solution.

### Conserved Gln346 keeps proenzyme in latent state

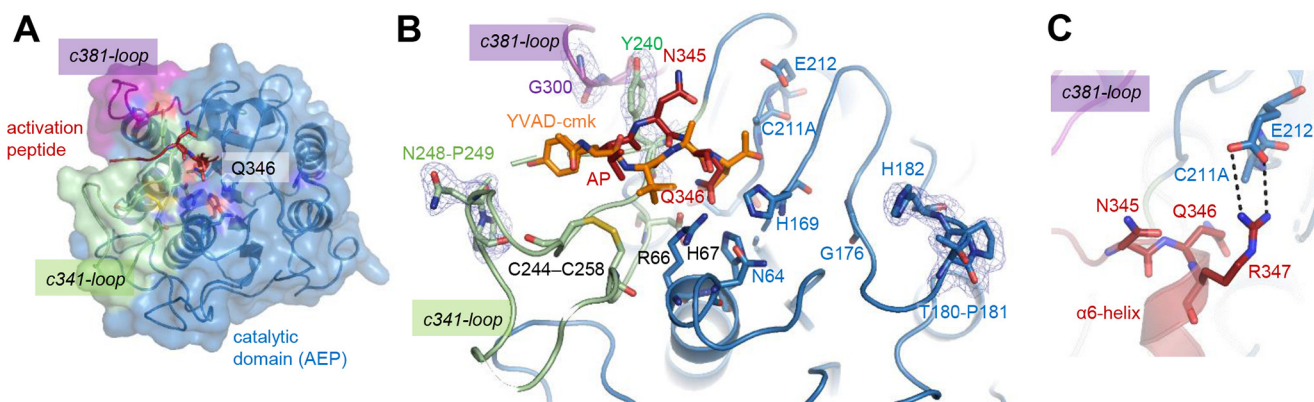
Comparing the crystal structure of proAtLEG $\beta$  with YVAD-cmk-inhibited AtLEG $\gamma$ , we found that the AP binds to the non-prime substrate binding sites in a substrate-like orientation, similar to what we previously observed in mammalian prolegumain (Fig. 1*B* and 3, *A* and *B*). Therefore, the AP is blocking substrate access, keeping the proenzyme in a latent, inactive state. Additionally, we observed a conserved Gln346 (AtLEG $\beta$  numbering) on the N-terminal end of the  $\alpha 6$  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 (Fig. 1*B* and 3*B*). This interaction was similarly observed in the crystal structure of *A. thaliana* legumain isoform  $\gamma$ ; 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 with Glu212, directly next to the catalytic Cys211 (Fig. 3*C*).

Similar to two-chain AtLEG $\gamma$  and mammalian prolegumains, the LSAM domain is further stabilized by two conserved disulfide bonds (Figs. 1*B* and 3*B*). On the C-terminal end of the LSAM domain, AtLEG $\beta$  harbors a potential vacuolar sorting

## Structure and function of AtLEG $\beta$



**Figure 2. proAtLEG $\beta$  is monomeric in solution.** *A*, crystal packing induced proAtLEG $\beta$  dimerization. Monomer 1 is shown in cartoon representation, and monomer 2 is labeled with a prime symbol (AEP' in surface representation). The location of the Asn309 glycosylation site is indicated with an orange diamond. *B*, superposition of *panel A* (proAtLEG $\beta$  dimer observed in the crystals) with dimeric two-chain AtLEG $\gamma$  (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 AtLEG $\gamma$ . Interaction is mediated by a hydrophobic core that is surrounded by electrostatic interactions. *D*, zoom-in view on the 4-helix bundle observed in proAtLEG $\beta$ . Interaction is mediated by a symmetric E390–K383 salt bridge localized on the  $\alpha$ 7-helix and hydrophobic interactions between the LFG motif (Leu396–Gly398) on the helix (or  $\alpha$ 7'-helix) and W355' on the  $\alpha$ 6' helix (or  $\alpha$ 6 helix). Relative to *panel B*, the views in *panels C* and *D* are rotated by 90° along the  $y$  axis. *E*, schematic representation of the 4-helix bundle as observed in AtLEG $\gamma$  and  $\beta$ . *F*, model of an AtLEG $\gamma$ -like dimerization mode in proAtLEG $\beta$ . AtLEG $\gamma$ -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 proAtLEG $\beta$ . 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 proAtLEG $\beta$ . The activation peptide (AP) harboring the autocatalytic Asn345 cleavage site and Gln346 that is occupying the S1 pocket are shown in red. *B*, zoom-in view on the nonprime and prime substrate binding sites with a YVAD-cmk peptide modeled based on the crystal structure of the YVAD-AtLEG $\gamma$  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  $\sigma$ . *C*, zoom-in view on the active site of proAtLEG $\beta$ . The ionic clamp (R347-E312) that links the  $\alpha$ 6-helix to the active site is indicated.

signal, which, however, is 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 proAtLEG $\beta$  activation, we analyzed the interdomain 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 proAtLEG $\gamma$  (R355 $^{\gamma}$ -E220 $^{\gamma}$  and K432 $^{\gamma}$ -D195 $^{\gamma}$ ; Fig. 4A). This is in stark contrast to proAtLEG $\gamma$ , 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 proAtLEG $\gamma$ , but not in proAtLEG $\beta$ , because of a local reorientation of the  $\alpha$ 6 helix. The residues involved in other AtLEG $\gamma$ -specific interdomain salt bridges are not conserved in AtLEG $\beta$ , *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.

Because the interaction between the catalytic domain and the LSAM domain in proAtLEG $\beta$  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 AtLEG $\beta$  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 to gain accessibility to the active site. Indeed, we found complete degradation of the LSAM domain at pH  $\leq$  4.0, indirectly indicating that the LSAM domain is destabilized under acidic pH conditions (Fig. 5A). Interestingly, upon incubation at pH 5.0, proAtLEG $\beta$  was split into catalytic (AEP) and LSAM

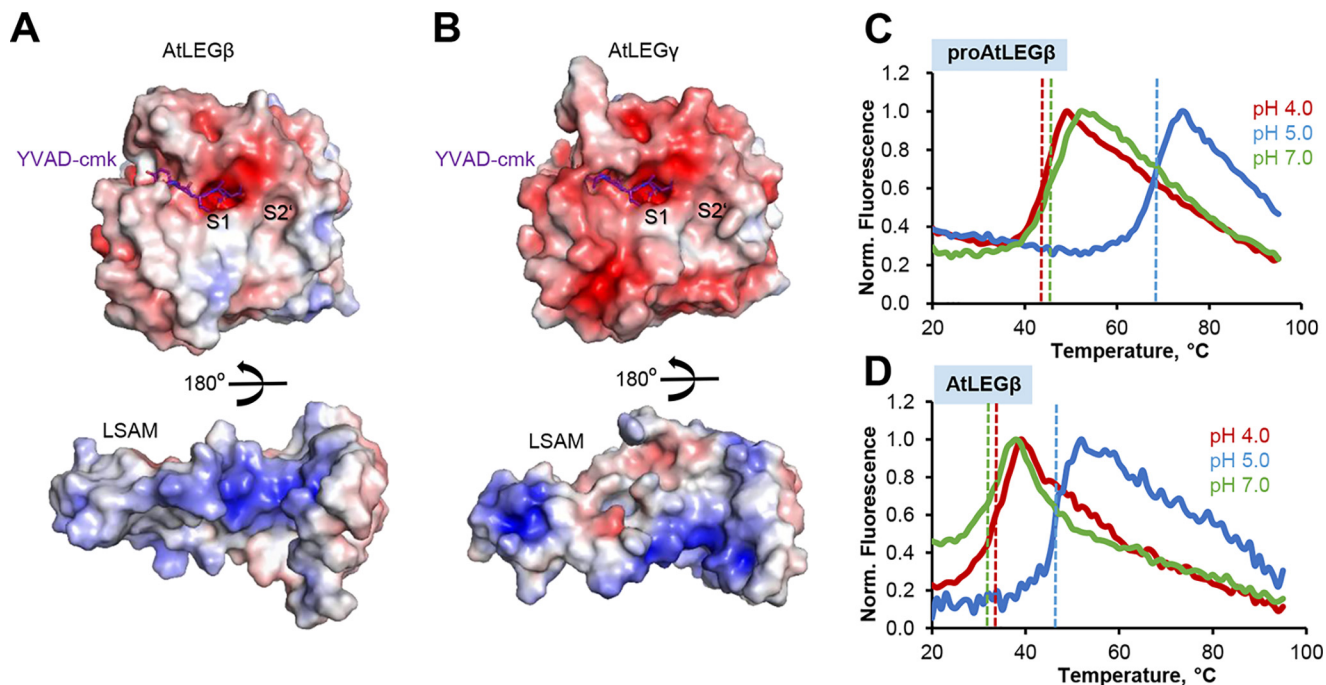
domains. However, the LSAM domain was not degraded but remained stable on SDS-PAGE. This suggested to us that AtLEG $\beta$  forms a two-chain state, where cleavage between LSAM and the catalytic domain occurred but both domains remained bound to each other. To test this, we performed SEC experiments using proAtLEG $\beta$  activated at pH 5.0. Indeed, we found a mixture of the two-chain state and isolated AEP domain (Fig. S3). Importantly, there was no dimeric two-chain intermediate state of AtLEG $\beta$  observed in SEC after activation.

#### Proteolytic activation is initiated by cleavages in the AP

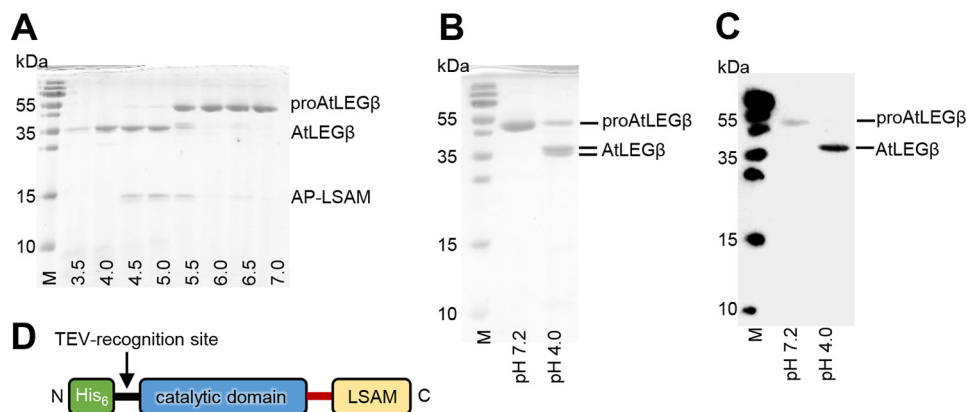
Using MS, we could identify two main autocatalytic cleavage sites, Asn333 and Asn345, on the AP (Fig. 1). These sites were similarly observed in proAtLEG $\gamma$  and seem to be equally accessible to cleavage. Upon incubation at pH  $<$  5.0, we observed additional cleavage sites on the LSAM domain, including Asp363, Asp416, and Asp417 (Fig. 1B). Because of 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  $\alpha$ 6- and  $\alpha$ 7-helices and could, in combination with processing at Asn333/345, allow the selective release of the  $\alpha$ 6-helix (fragment Gln346-Asp363), as observed in mammalian legumain (31). Asp416 and Asp417 are localized within the V<sup>415</sup>DDW<sup>418</sup> motif, right before the  $\alpha$ 9-helix (Fig. 1B and Fig. 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 AtLEG $\beta$  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  $\alpha$ 6-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 propeptide. Here, it is important to note that our proAtLEG expression constructs typically carry an N-terminal His<sub>6</sub> tag followed by a TEV recognition site (ENLYFQG; TEV, tobacco etch virus protease). We found that AtLEG $\beta$  was capable of cleaving after the Asn residue within the TEV recognition site and thereby

## Structure and function of AtLEG $\beta$



**Figure 4. The AEP–LSAM interaction in proAtLEG $\beta$  is mostly hydrophobic.** *A*, color-coded electrostatic surface potential of AtLEG $\beta$  AEP and LSAM domains based on the crystal structure of proAtLEG $\beta$  (blue, positive charge, red, negative charge) calculated at pH 7.0 and contoured at  $\pm 5$  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 crystal structure of the AtLEG $\gamma$  inhibitor complex (PDB entry 5obt). *B*, same as panel *A*, but calculated for AtLEG $\gamma$  in complex with YVAD-cmk inhibitor. *C*, melting curves of proAtLEG $\beta$  at indicated pH values show highest thermal stability at pH 5. Melting points are indicated by dashed lines. *D*, melting curves of active AtLEG $\beta$  showing highest stability at pH 5.0.



**Figure 5. Autocatalytic activation of AtLEG $\beta$  is pH dependent and results in a two-chain intermediate state (pH 5.0) and active AEP state (pH 4.0).** *A*, ProAtLEG $\beta$  after 1-h incubation at indicated pH values. AtLEG $\beta$  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 proAtLEG $\beta$  at pH 7.2 and AtLEG $\beta$  following activation at pH 4.0. Activation results in a double band at around 36 kDa. *C*, Western blot using an anti-His-6 antibody, showing that only one AtLEG $\beta$  activation product harbors the N-terminal His<sub>6</sub>-tag. *D*, scheme illustrating the domain architecture of the recombinant expression construct.

removing the His<sub>6</sub> tag, as evidenced by a Western blot using an anti-His antibody (Fig. 5, *B* and *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, because N-terminal processing within the TEV recognition motif was only observed under very acidic pH conditions, we suggest that the latter cleavage is performed by the fully activated AtLEG $\beta$ . Importantly, N-terminal cleavage is not a physiological event, as the relevant sequence is not present in native proAtLEG $\beta$  (Fig. S1).

### (pro)AtLEG $\beta$ is stable at intermediate pH

Based on the remarkable variances we observed at the AEP–LSAM interfaces of proAtLEG $\beta$  and  $\gamma$ , we hypothesized that they would translate into differences of their pH stability profiles. Indeed, when we measured the thermal stability of proAtLEG $\beta$  using differential scanning fluorimetry, we found a stability optimum at pH 5.0 (Fig. 4C). This is very different from proAtLEG $\gamma$  and mammalian legumain, where the stability optimum of the proenzyme is at neutral pH (21, 32). Even more interestingly, we found that AtLEG $\beta$  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 AtLEG $\gamma$  and also mammalian legumain, where the AEP domain is most stable at pH  $\sim$ 4. However, this difference becomes clear considering the hydrophobic interaction between AEP and LSAM domain in proAtLEG $\beta$ . Mammalian legumain and AtLEG $\gamma$  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 for by the LSAM domain. In human legumain and AtLEG $\gamma$ , the isolated AEP can be stabilized by protonation of the excess acidic residues, hence the maximum stability at pH 4. The AEP in AtLEG $\beta$  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 proAtLEG $\beta$ , AEP–LSAM interaction and stabilization do not depend on neutral pH, whereas the tight electrostatic clamping of these domains in proAtLEG $\gamma$  and human prolegumain depend on neutral pH. Consequently, proAtLEG $\beta$  is most stable at the pH that is also favorable for the isolated AEP $\beta$ .

#### 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 proAtLEG $\beta$  to analyze the active AtLEG $\beta$  state. When we superimposed the AEP domains of AtLEG $\beta$  and  $\gamma$ , we found that their fold is highly conserved (Fig. 6A). AtLEG $\beta$  exhibits a caspase-like topology, *i.e.* a 6-stranded central  $\beta$ -sheet that is surrounded by 5 major  $\alpha$ -helices (Fig. S1 and S4) (35). Furthermore, AtLEG $\beta$  harbors the c341- and c381-loops, which form the nonprime 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 with 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 stable bend and turn formation (Fig. 3B) (36). Interestingly, both turns are located in the substrate binding sites. The Asn248-Pro249 cis-peptide bond is on the c341-loop (nonprime side) and presents the Asn248 carbonyl oxygen as the main-chain recognition site for the P4 amide. Thr180-Pro181 is part of the eastern rim of the S2' pocket.

#### AtLEG $\beta$ has a wide S3-S4 pocket

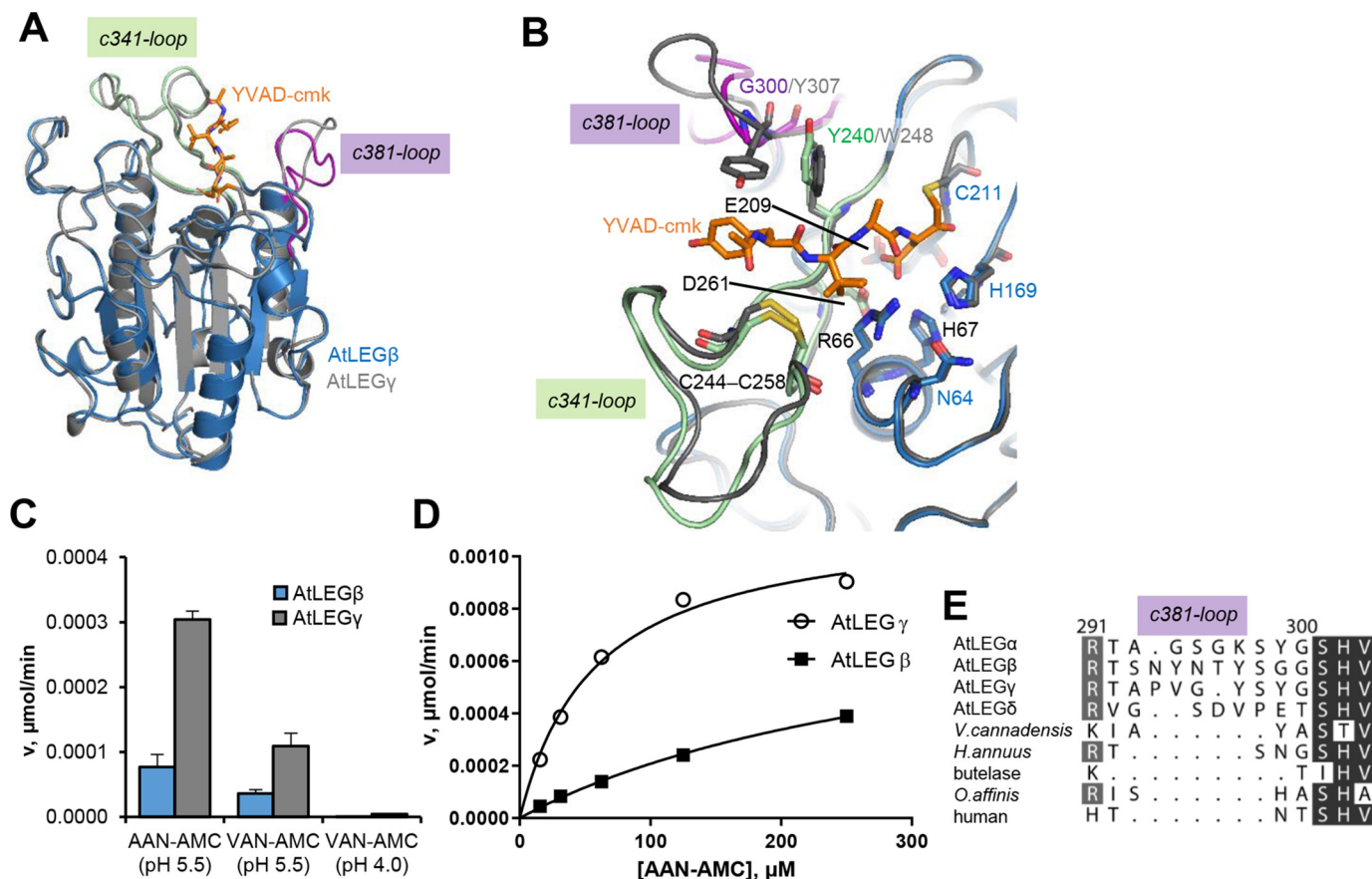
When looking into the active site, we found that the active-site residues Cys211, His168, and Asn64 also superimpose very well with the related AtLEG $\gamma$  (Fig. 6B). Furthermore, the residues forming the S1-specificity pocket, Arg66, His67, E209, and D261, adopt conformations identical to those observed in AtLEG $\gamma$ . The highly conserved architecture of the active site

suggested similar substrate specificity and catalytic activity of AtLEG $\beta$  and  $\gamma$ . However, when we compared the catalytic activity toward the fluorogenic Ala-Ala-Asn-AMC substrate, we observed a surprisingly low catalytic activity for AtLEG $\beta$  compared with  $\gamma$  (Fig. 6C). Because the positioning of the active-site residues were basically identical in  $\beta$  and  $\gamma$ , we did not expect this difference in activity to originate from a  $k_{\text{cat}}$  effect but rather from differences in substrate affinity ( $K_m$ ). Beyond the highly similar S1 pocket, we identified major differences on the c341- and c381-loops on the nonprime side (Fig. 6B and Fig. S5). Variations in sequence and conformation resulted in a narrow S3–S4 pocket in AtLEG $\gamma$  but a rather wide pocket in AtLEG $\beta$  (Fig. 4, A and B, 6, and Fig. S5). To test whether these differences were a result of the induced fit of the YVAD-cmk inhibitor, we superposed the crystal structures of proAtLEG $\beta$ , two-chain (pro)AtLEG $\gamma$ , and active YVAD-AtLEG $\gamma$  and compared their active sites. Interestingly, we found that the conformations of the substrate specificity loops c341 and c381 of proAtLEG $\beta$  most closely resembled the active state of AtLEG $\gamma$ . Therefore, we could exclude that induced fit was a main regulator of substrate affinity (Fig. S5). However, the situation might be different in AtLEG $\gamma$ , 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-AtLEG $\gamma$  crystal structure, we found tight interactions in AtLEG $\gamma$  but fewer interactions to AtLEG $\beta$ . Whereas AtLEG $\beta$  offered an open, broad surface to accommodate the YVAD substrate, AtLEG $\gamma$  was tightly embracing the peptidic substrate, as visible in Fig. 4, A and 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 AtLEG $\beta$  compared with gamma because of missing enzyme-substrate interactions. Indeed, when we determined  $K_m$  values for AtLEG $\beta$  and  $\gamma$  toward the AAN-AMC substrate, we found high-affinity binding ( $K_m = 57 \pm 3 \mu\text{M}$ ) to AtLEG $\gamma$  but low affinity for AtLEG $\beta$  ( $K_m = 337 \pm 3 \mu\text{M}$ ) (Fig. 6D). Importantly, we found similar  $k_{\text{cat}}$  (AtLEG $\beta$ ,  $4.5 \times 10^{-3} \text{ min}^{-1}$ ; AtLEG $\gamma$ ,  $6.3 \times 10^{-3} \text{ min}^{-1}$ ) and  $V_{\text{max}}$  values (AtLEG $\beta$ ,  $0.9 \times 10^{-3} \mu\text{mol/min}$ ; AtLEG $\gamma$ ,  $1.1 \times 10^{-3} \mu\text{mol/min}$ ) for both enzymes. These findings confirmed that the difference in catalytic activity between AtLEG $\beta$  and  $\gamma$  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 for both AtLEG $\beta$  and  $\gamma$  (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-loops serve as a  $K_m$  switch. To analyze this further, we superposed all plant legumain structures available in the PDB. Whereas the main structural elements superimposed

## Structure and function of AtLEG $\beta$



**Figure 6.** AtLEGs differ in their substrate specificity loops. *A*, superposition of AtLEG $\beta$  (blue) and  $\gamma$  (gray) AEP domains. The YVAD-cmk inhibitor bound to AtLEG $\gamma$  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, and residues forming the S1 specificity pocket are labeled in black. *C*, catalytic activities of AtLEG $\beta$  and  $\gamma$  toward peptidic AAN-AMC and VAN-AMC substrates at indicated pH values. *D*,  $K_m$  determination for AtLEG $\beta$  and  $\gamma$  toward the AAN-AMC substrate. *E*, sequence alignment of the c381-loops of indicated (plant) species. Sequences were derived from structures deposited in the PDB, where applicable. AtLEG $\alpha$  (P49047), AtLEG $\beta$  (Q39044), AtLEG $\gamma$  (5nij), AtLEG $\delta$  (Q9LJX8), *V. canadensis* (*Viola Canadensis*; 5zbi), *H. annuus* (*Helianthus annuus*; 6azt), butelase (*Clitoria ternatea*; 6dhi), and *O. affinis* (*Oldenlandia affinis*; 5hoi) were used.

very well in all available structures, we observed big differences on the c381-loops. It is variable in length and sequence and may even contain a glycosylation site (Fig. 6E and Fig. 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).

### AtLEG $\beta$ substrate specificity is pH dependent

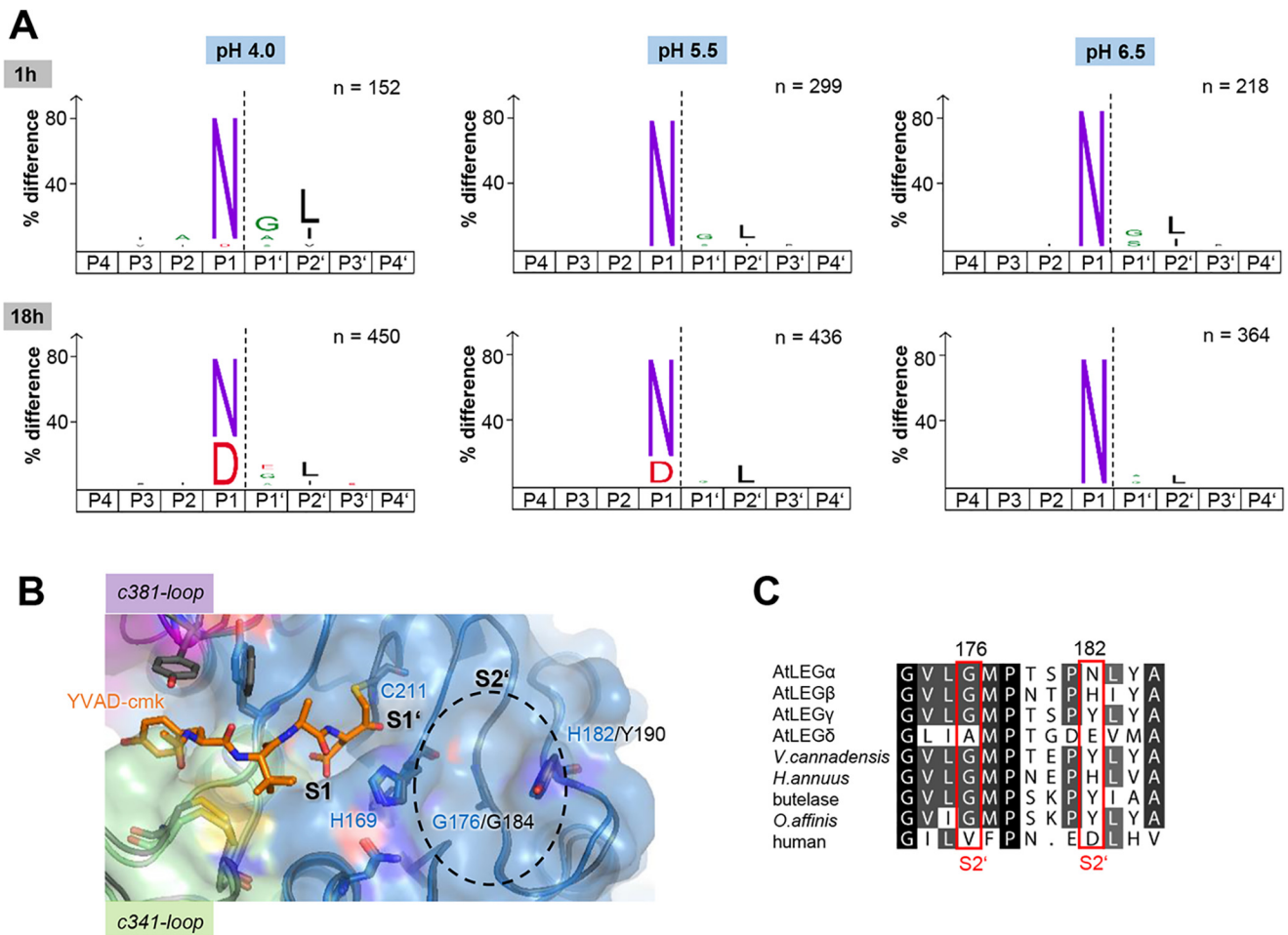
To further analyze the substrate specificity of AtLEG $\beta$ , we carried out PICS experiments, which use proteome-derived peptides as substrate libraries (37, 38). Here, we used a peptide library that was generated from an *E. coli* proteome by digestion with trypsin for AtLEG specificity profiling under three different pH conditions. As expected, we observed a strong preference for Asn in the 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 (18 h). 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

bipolar architecture of the S1 specificity pocket and with previously published data for human legumain (Fig. 6B) (32).

### AtLEG $\beta$ has a strong preference for hydrophobic residues in P2'

Furthermore, we observed a slight preference for small, polar residues in the 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 the P2' position. Leucine has previously been proven to be beneficial at the P2' position in legumain ligase substrates (16). Together, these results are in nice agreement with the architecture of the S1' and S2' binding sites. Whereas 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 the 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,





**Figure 7. AtLEG $\beta$  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 the substrate library. iceLogos visualize the substrate preference surrounding the cleavage sites ( $p = 0.05$ ) based on peptides cleaved by AtLEG $\beta$  after incubation at indicated pH values and times. The numbers of nonredundant cleavage sites used to generate the iceLogos are indicated. *B*, top view of the AtLEG $\beta$  substrate binding site. Binding of the YVAD-cmk inhibitor was modeled based on the crystal structure of the YVAD-AtLEG $\gamma$  complex (PDB entry 5obt). *C*, sequence alignment of the residues forming the prime-substrate binding site. Sequences used are the same as those in Fig. 6.

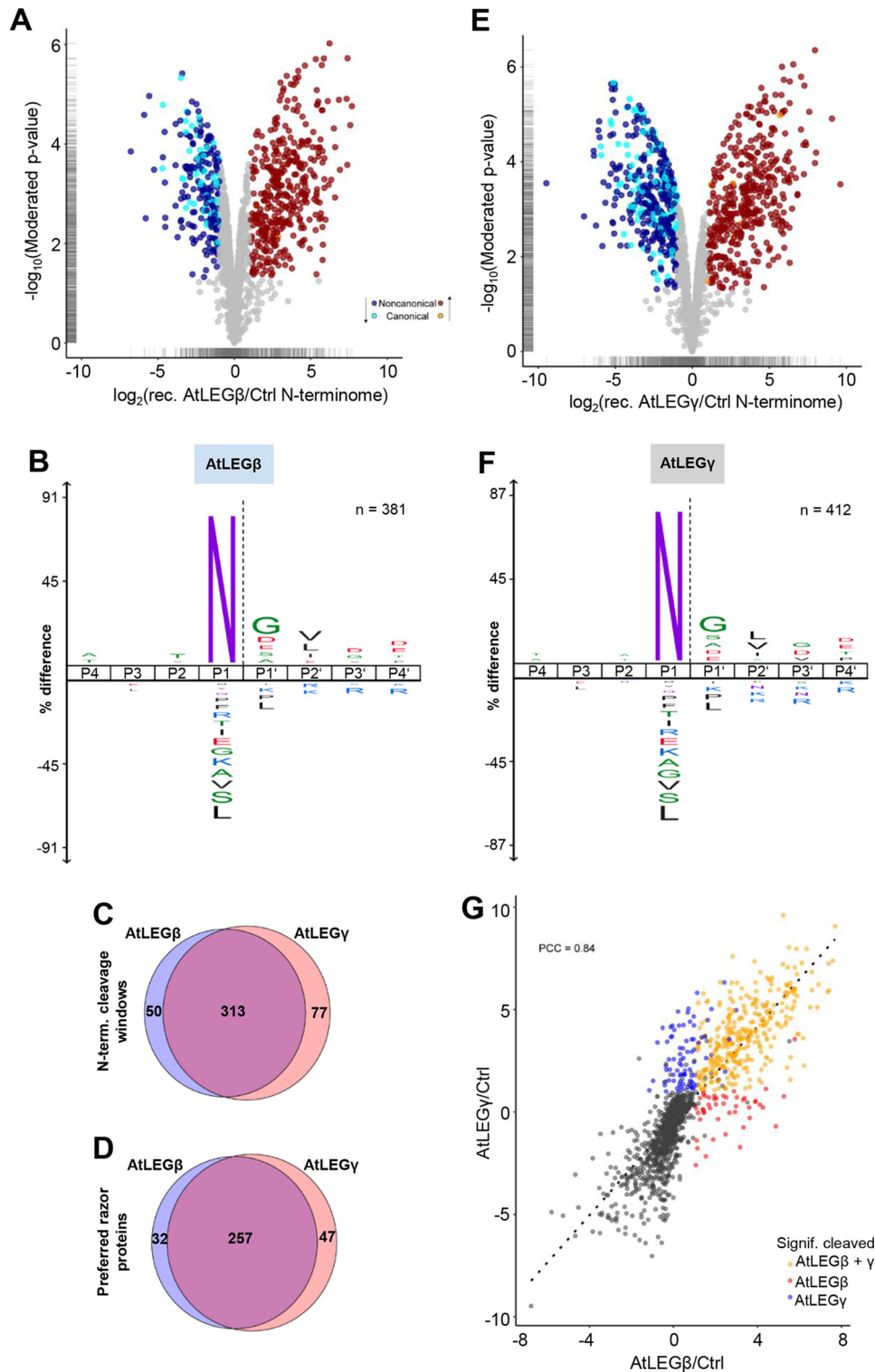
mammalian legumain harbors a valine at position 176, making the S2'-pocket shallower and, thereby, less specific at this position (39). Furthermore, AtLEG $\delta$  has the glycine replaced by alanine (Fig. 7C), suggesting that it also will have a less pronounced specificity at the P2' position. To test the relevance of His182 for prime side substrate specificity, we repeated the PICS experiments using AtLEG $\gamma$ , which has a tyrosine at the equivalent position (Fig. S7). Interestingly, we found highly similar preferences on the nonprime and prime substrate binding sites, further confirming that Gly176 is the main determinant at the S2' site.

#### AtLEG $\beta$ has a strong preference for small residues in P1' position in protein substrates

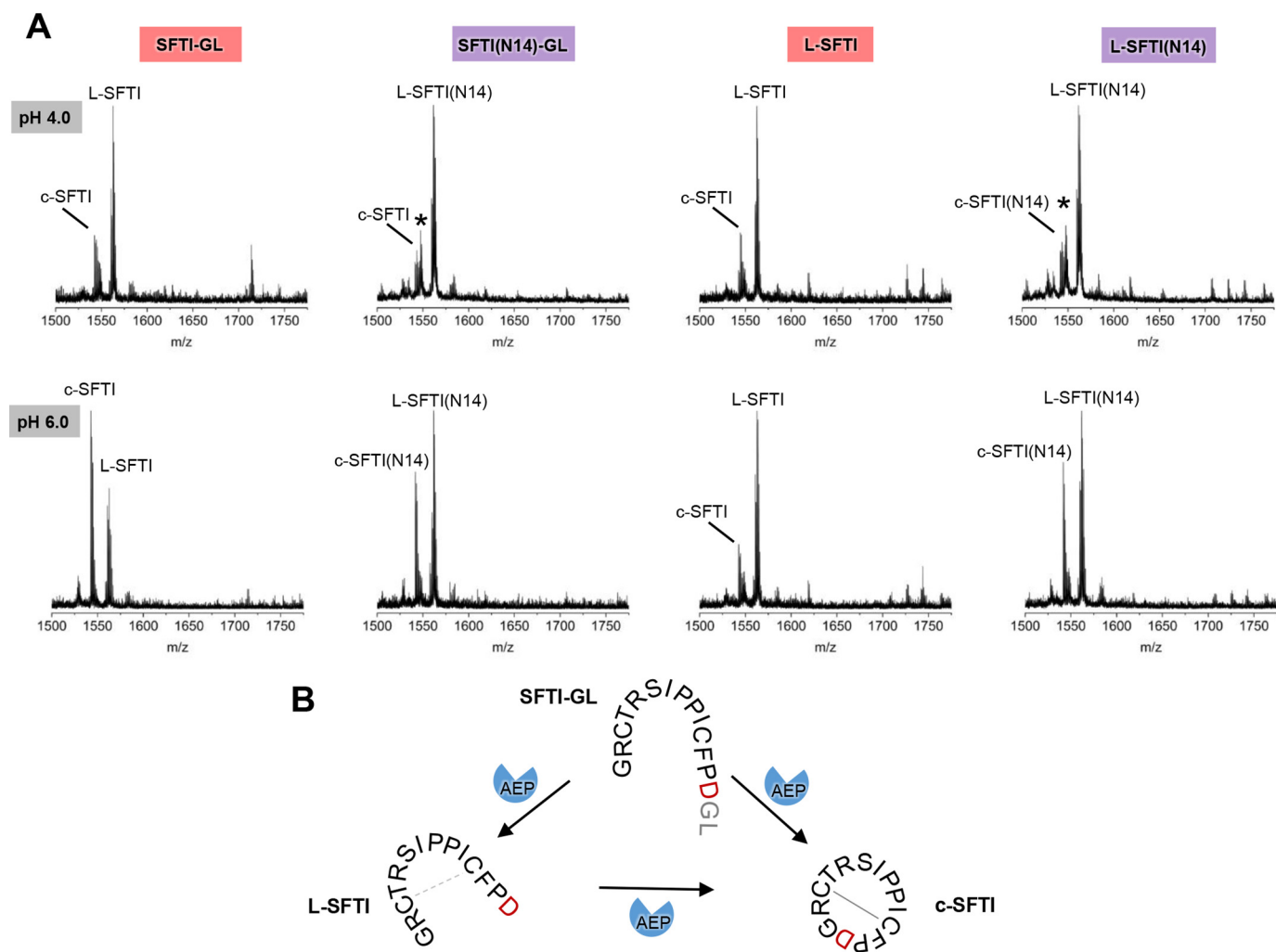
In the next step, we analyzed the substrate specificity of AtLEG $\beta$  toward protein substrates, using proteome extracts isolated under nondenaturing conditions from leaves of the *A. thaliana vpe0* mutant lacking expression of all four VPE isoforms as a substrate library. After incubation with recombinant AtLEG $\beta$ , recombinant AtLEG $\gamma$ , or buffer control, free N-terminal  $\alpha$ -amines were labeled with three different formalde-

hyde isotopologues, and cleavage sites were determined using the HUNTER N-termini enrichment and MS (40). Based on three biological replicates, we identified 381 N-terminal peptides significantly accumulating after incubation with AtLEG $\beta$  at pH 6.0 (Fig. 8, A and B, and Table S1), matching to 363 unique cleavage sites (Fig. 8C) in 289 proteins (Fig. 8D). As expected, we found a pronounced preference for Asn at the P1 position (Fig. 8B). Furthermore, we observed a stronger preference for small and polar residues in the 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 the P2' position. For AtLEG $\gamma$ , we identified 412 significantly accumulating N-terminal peptides (Fig. 8E, Table S1). 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, the vast majority of 313 of the cleavage sites in 257 proteins were cut by both enzymes, although only 50 cleavages in 32 proteins were strongly preferred

## Structure and function of AtLEG $\beta$



**Figure 8. Substrate specificity of AtLEG $\beta$  and  $\gamma$  toward intact proteins extracted from *A. thaliana* leaves.** Volcano plots identify protein N-terminal peptides significantly changing in abundance (greater than 2-fold change in abundance supported by LIMMA-moderated  $t$  test;  $p < 0.05$ ) after *in vitro* incubation of *A. thaliana vpe0* proteome with recombinant AtLEG $\beta$  (A) or AtLEG $\gamma$  (E). Log $_2$  fold change is the mean of 3 biological replicates. Accumulating N-terminal peptides indicative of AtLEG $\beta$ / $\gamma$  cleavage are highlighted red, and depleted peptides cleaved within their sequence are in blue. iceLogos visualize the substrate preference surrounding the cleavage sites for AtLEG $\beta$  (B) and AtLEG $\gamma$  (F) ( $p = 0.05$ ). Numbers of nonredundant cleavage sites used to generate the iceLogos are indicated. Venn diagrams show the overlap of cleavage sites (C) and proteins (D) cleaved in the *vpe0* proteome after incubation with AtLEG $\beta$  or  $\gamma$ . 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 ( $\log_2$  fold change of  $>1$ , LIMMA-moderated  $t$  test  $p$  value of  $<0.05$ ) indicate cleavage by AtLEG $\beta$  (red), AtLEG $\gamma$  (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. Cyclization of SFTI-derived peptides by AtLEG $\beta$  is pH dependent.** *A*, reactions were carried out at indicated pH values. An unidentified species is labeled with an *asterisk*. *B*, reaction scheme of AtLEG $\beta$  catalyzed cyclization of SFTI-GL peptides. The precursor 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.

substrates of AtLEG $\beta$  and 77 cleavage sites of 47 proteins were selectively cut by AtLEG $\gamma$  (Fig. 8, *C*, *D*, and *G*).

#### AtLEG $\beta$ is a broad-spectrum transpeptidase

To characterize the cyclase activity of AtLEG $\beta$ , we coincubated it with different SFTI-derived linear peptides and measured the formation of the cyclic product using MS. Indeed, we found that AtLEG $\beta$  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. 9*A*). Cyclization worked most efficiently 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 the P1 position, we observed a product formation rate of about 60%. This is less than that with AtLEG $\gamma$ , which resulted in approximately 80% product formation (29). Interestingly, when the P1 residue was replaced by Asn, as is the case in SFTI(N14)-GL, AtLEG $\beta$  was still able to catalyze peptide cyclization, in contrast to the situation of AtLEG $\gamma$ . When we replaced the P1'–P2' Gly-Leu with His-Val residues, which is the pre-

ferred sequence found for butelase-1 (*C. ternatea* legumain), we observed a similar cyclization efficiency (Fig. S8) (16), showing us that albeit optimized for butelase-1, the HV-dipeptide is not facilitating peptide ligation in AtLEG $\beta$ .

#### AtLEG $\beta$ is a broad-spectrum ligase

Along these lines, we also tested whether AtLEG $\beta$  would be able to cyclize linear L-SFTI and L-SFTI(N14) peptides, which lack amino acids on P1' and P2' positions of the protease substrate. Surprisingly, AtLEG $\beta$  was indeed able to join the free termini and form the cyclic product, suggesting that AtLEG $\beta$  is not only a transpeptidase but also a real ligase (Fig. 9, *A* and *B*). Using the SFTI peptides carrying Asn at P1 position (N14), cyclization worked equally well with or without the preceding cleavage of prime side residues. In the case of Asp at the P1 position, transpeptidation (cleavage-linked ligation) was preferred 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.

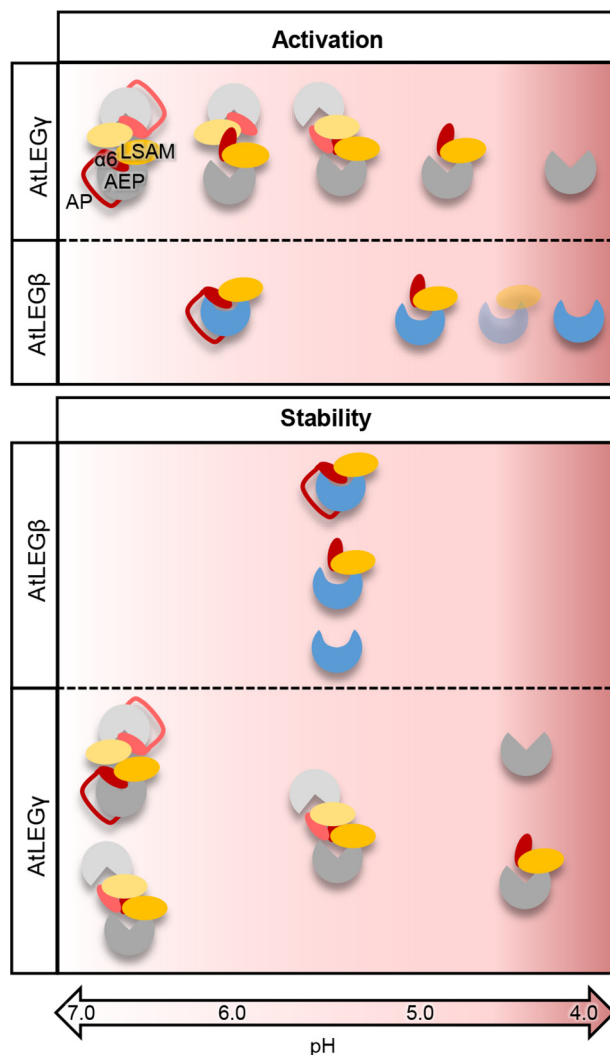
## Structure and function of AtLEG $\beta$

### Discussion

Dimerization is a critical regulatory event for caspase-like proteins. In the case of the apoptotic caspases, dimerization is mediated primarily by the  $\beta 6$  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 AtLEG $\gamma$  (5nij) both showed a dimer state that was mediated by the  $\alpha 6$  and  $\alpha 7$  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 autoprocess 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. First, we observed that proAtLEG $\beta$  is monomeric in solution. In this respect, autocatalytic activation of proAtLEG $\beta$  more resembles 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 proAtLEG $\beta$  dimers per asymmetric unit. Nonetheless, this atypical dimer interaction is transient and short-lived and, hence, could not be observed in solution experiments. Second, we found that the AEP–LSAM interface is rather hydrophobic and not charged in nature. Consequently, the stability profile of AtLEG $\beta$  differs from AtLEG $\gamma$  and mammalian legumains (Fig. 10). Third, AtLEG $\beta$  encodes autocatalytic cleavage sites on both ends of the  $\alpha 6$ -helix (Asn345 and Asp363), which in principle allows the selective removal of the AP, like in mammalian legumain (32). Whereas N-terminal cleavage was observed at pH <6.0, cleavage on the C-terminal end of the  $\alpha 6$ -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  $\alpha 6$ -helix (Arg347) to the active site (Glu212) will loosen (Fig. 3C), which will further facilitate the release of the AP ( $\alpha 6$ -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 a new regulatory mechanism distinct from that of AtLEG $\gamma$ . Different oligomerization states will cause AtLEG $\beta$  to favor activation at high local concentrations but will favor the latent two-chain state in AtLEG $\gamma$ . On the other hand, the transient dimers observed in the AtLEG $\beta$  crystal might play a role in cooperative substrate processing. Together, these findings suggest that AtLEG $\beta$  and  $\gamma$  represent examples of two distinct classes of plant legumains, concerning their physiological function and 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 fea-



**Figure 10. Activation and stability of AtLEG $\beta$  and AtLEG $\gamma$  are pH dependent.** In the zymogen forms of proAtLEG $\beta$  and  $\gamma$ , the LSAM domain (orange) and activation peptide (AP, red line) that harbors the  $\alpha 6$ -helix (red ellipsoid) sit on top of the active site and thereby block access to the substrate binding sites. ProAtLEG $\gamma$  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  $\alpha 6$ -helix. By lowering pH, the interaction of the  $\alpha 6$ -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  $\alpha 6$ -helix and the LSAM domain. In contrast, proAtLEG $\beta$  is a monomer in solution. Activation proceeds via cleavage after (1) Asn333/345 on the N-terminal end of the  $\alpha 6$ -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  $\alpha 6$ -helix selectively removed but the LSAM domain still bound to the AEP domain (indicated by transparent coloring). Whereas all AtLEG $\beta$  activation states show highest conformational stability at intermediate pH, two-chain (pro)AtLEG $\gamma$  is stable at neutral to slightly acidic pH, and monomeric two-chain AtLEG $\gamma$ , as well as the AEP domain, are most stable at acidic pH.

ture 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 because of 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 consequently may 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 proAtLEG $\beta$ , 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 AtLEG $\beta$  and  $\gamma$  cleavage sites in protein substrates *in vitro*.

Previously, the c381-loop was described as a marker of ligase activity (28). More precisely, a deletion in that region was associated with an increase in ligase activity. However, both AtLEG $\beta$  and  $\gamma$  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. Because protease and ligase activities are inversely coupled, the c381-loop may be an indirect marker of ligase activity. If the affinity of the nonprime (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 nonprime substrates may also result in less recleavage of cyclic products and thereby again indirectly favor ligation. This hypothesis also fits with our observation that AtLEG $\beta$ , which has a nonprime 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  $K_m$ ) nonprime 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 nonprime 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. Additionally, the residence time of the ligation product is very short, making recleavage 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. The  $K_m$  likely 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)  $K_m$  than C-terminally free Asp, giving the P1-Asp substrate the

possibility for binding and transpeptidation. Looking at the prime substrate binding sites, we found that AtLEG $\beta$  and  $\gamma$  encode nearly identical substrate binding sites. Taken together, differences in ligation efficiency between AtLEG $\beta$  and  $\gamma$  might be explained by their different nonprime 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 gatekeeper residue for ligase activity (20). Mutation to Ala247 resulted in an enzyme with superior ligase activity. Because all AtLEG isoforms harbor a glycine at the equivalent position (Gly241, AtLEG $\beta$  numbering), this residue cannot explain the observed isoform specific differences in ligase activity. Similarly, the sequence motif Gly171-Pro172 (AtLEG $\beta$  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 nonprime binding site (S2–S3) and which is different in AtLEG $\gamma$  (Trp248). Based on this observation and the differences in the nearby c381-loop, we suggest that it is rather the overall architecture of the nonprime substrate binding sites that affects substrate affinity and might positively affect ligase activity.

The ability of AtLEG $\beta$  to join free ends is also interesting from a biotechnological point of view, as it will allow us to link targets without the necessity of introducing artificial cleavage sites, still with the prerequisite of 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 AtLEG $\beta$ -like enzyme present in every plant.

Previously, we could show that the two-chain state observed in AtLEG $\gamma$  is especially interesting with regard to ligase activity, as it is stable at neutral pH environments where ligase activity is favored. Because two-chain AtLEG $\beta$  has the same pH stability profile as active AEP, with a pH stability optimum at 5.0, two-chain AtLEG $\beta$  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 carboxyl-peptidase activity rather than endopeptidase activity (32). The carboxypeptidase activity is structurally encoded by LSAM-derived arginine residues, which anchor the carboxyl 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 and R), which could similarly indicate carboxypeptidase activity of two-chain AtLEG $\beta$ . 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 AtLEG $\beta$  indeed harbors 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  $\beta$  (AtLEG $\beta$ ) full-length clone U12200 (locus

## Structure and function of AtLEG $\beta$

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<sub>6</sub> 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 that allowed the amplification of the cyclic plasmid template, harboring the proAtLEG $\beta$  WT 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 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 proAtLEG $\beta$  protein was stored in a buffer composed of 20 mM Hepes, pH 7.0, and 50 mM NaCl. ProAtLEG $\gamma$  was prepared by following the same protocol.

### Crystallization, data collection, and refinement

Initial screening was performed using the sitting-drop vapor-diffusion method utilizing a Hydra II Plus one liquid-handling system. Crystals of proAtLEG $\beta$  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 preincubation in a cryoprotectant 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 X-ray 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. 1000 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 AtLEG $\gamma$  combined with the sequence of proAtLEG $\beta$ . 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  $\pm 5$  kT/e.

### Autoactivation

To test the pH dependence of autoactivation of proAtLEG $\beta$ , we incubated it at a concentration of 0.4 mg/ml in buffer com-

posed 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 AtLEG $\beta$  on a 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 autoactivation was analyzed by SDS-PAGE. Active AtLEG $\beta$  was buffer exchanged using a NAP column (GE Healthcare) preequilibrated in a buffer composed of 20 mM citric acid, pH 4.0, and 50 mM NaCl. Active AtLEG $\gamma$  was prepared by following the protocol described in reference 21.

### Enzymatic activity assays

The enzymatic activity of active AtLEG $\beta$  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  $\mu$ M 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 AtLEG $\beta$  or  $\gamma$  with serial dilutions of the AAN-AMC substrate in assay buffer. Kinetic data were processed using GraphPad, and  $K_m$  values were calculated using implemented algorithms.

### Characterization of oligomerization state

To test the oligomerization state of proAtLEG $\beta$ , 200  $\mu$ l of sample was 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 AtLEG $\beta$ , we loaded it on an S200 column preequilibrated 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 proAtLEG $\beta$  and pH 4.0-activated AtLEG $\beta$ , 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 were analyzed as described elsewhere (51).

### Western blotting

Protein samples to be analyzed were separated on an SDS-PAGE gel. Subsequently, proteins were blotted onto an Amersham Biosciences Protran 0.45 NC membrane (GE Healthcare) using a Trans-Blot SD semi-dry transfer cell (Bio-Rad). The membrane was blocked with 1 $\times$  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

Biosciences 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 AtLEG $\beta$  and  $\gamma$ , 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  $\mu$ g/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 ( $^{13}\text{CD}_2\text{O}$ ) and 20 mM sodium cyanoborohydride and control reactions with 20 mM light formaldehyde ( $\text{CH}_2\text{O}$ ) and 20 mM sodium cyanoborohydride for 2 h and quenched with 100 mM Tris, pH 8.0, for 1 h. Protease-treated and control samples were mixed and purified using C<sub>18</sub> StageTips.

### Substrate specificity profiling of AtLEG $\beta$ and $\gamma$ using intact *A. thaliana* leaf proteome

*A. thaliana* VPE quadruple mutant (VPEO [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, 22 °C/18 °C, 120  $\mu$ mol photons  $\text{m}^{-2} \text{s}^{-1}$ ). 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 mixture (ThermoFisher, Dreieich, Germany) on ice. The lysate was then filtered through Miracloth (Merck, Darmstadt, Germany), followed by centrifugation at 4000  $\times g$  at 4 °C for 5 min. The protein concentration was determined by the Bradford assay using BSA as a reference.

Recombinant AtLEG $\beta$ , recombinant AtLEG $\gamma$ , 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 °C for 3 h. The reactions were terminated by addition of 50  $\mu$ M 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 °C for 30 min followed by alkylation with 15 mM iodoacetamide for 30 min at 25 °C and quenched by addition of 15 mM DTT for 15 min. The three samples were differentially dimethyl labeled with 20 mM light formaldehyde ( $^{12}\text{CH}_2\text{O}$ ) and 20 mM sodium cyanoborohydride (light label), 20 mM medium formaldehyde ( $^{12}\text{CD}_2\text{O}$ ) and 20 mM sodium cyanoborohydride (medium label), or 20 mM heavy formaldehyde ( $^{13}\text{CD}_2\text{O}$ ) and 20 mM sodium cyanoborodeuteride. After 16 h of incubation at 37 °C, the same amounts of fresh reagents were added and incubated for another 2 h. The reactions were

quenched with 0.1 M Tris (final concentration) at pH 7.4 and 37 °C for 1 h. 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 °C for 16 h. 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 °C for 45 min. An 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  $\times g$  for 5 min. Next, the supernatant was injected through a pre-activated HR-X (M) cartridge (Macherey-Nagel, Düren, Germany). The flowthrough 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 C<sub>18</sub> Stage-Tip purification prior to mass spectrometric analysis. The assays were performed in three biological replicates.

### MS data acquisition

Samples were analyzed on a two-column nano-HPLC setup (Ultimate 3000 nano-RSLC system with Acclaim PepMap 100 C<sub>18</sub>, ID of 75  $\mu$ m, particle size of 3  $\mu$ m, a trap column of 2-cm length and analytical column of 50 cm length, ThermoFisher) with a binary gradient from 5–32.5% B for 80 min (A, H<sub>2</sub>O + 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 previously (55). Data were 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 top 17 most intense ions were selected for fragmentation with dynamic exclusion of previously selected precursors for the next 30 s, unless intensity increased 3-fold compared with 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 (>25,000 cts) spectra. Fragment spectra were averaged from t-stepped parameters, with 50% of the acquisition time manner with split parameters: 61- $\mu$ s transfer time, 7 eV collision energy, and a collision RF of 1500 Vpp, followed by 100- $\mu$ s transfer time, 9 eV collision energy, and a collision RF of 1800 Vpp

### MS data analysis

Acquired mass spectra were matched to peptide sequences at an 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 November 2015, 4313 entries) with appended common contaminants was used. Search parameters considered semispecific tryptic peptides, light (+28.031300) and heavy (+36.075670) dimethyl labeling at peptide N-termini or Lys side chain amines, and Cys carbamidomethylation as fixed and Met oxidation as

## Structure and function of AtLEG $\beta$

variable modifications. Identified peptides that showed at least a 4-fold increase in intensity after protease treatment compared with the control treatment or were exclusively present in the protease-treated condition were considered 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 previously (38) and visualized as IceLogos using software version 1.3.8 (57).

For HUNTER experiments, the *A. thaliana* UniProt proteome database (downloaded December 2018, 41,592 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 pyro-Glu formation from Glu (−18.010565) or Gln (−17.026549) as variable modifications. Further statistical data analysis, filtering and annotation were performed with the Perl script MANTI.pl 3.9.7 (<https://manti.sourceforge.io>).

### Peptide cyclization assay

SFTI-derived peptides were synthesized and analyzed as described previously (29). Subsequently, cyclization experiments were carried out using 500  $\mu$ M of the respective linear peptide and 0.5  $\mu$ M AtLEG $\beta$  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 C<sub>18</sub> tips (Merck Millipore) and analyzed by MALDI-TOF-MS (Autoflex, Bruker Daltonics, matrix,  $\alpha$ -cyano-4-hydroxycinnamic 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 (58) repository with the accession codes PXD019220 for the PICS data set and PXD019276 for the HUNTER N-terminome data set. All remaining data are contained within the article.

**Acknowledgments**—We thank Sabine Markovic-Ullrich for peptide synthesis.

**Author contributions**—E. D., F. B. Z., C. C., P. F. H., and H. B. conceptualization; E. D., F. B. Z., W. T. S., F. D., C. C., P. F. H., and H. B. data curation; E. D., C. C., P. F. H., and H. B. supervision; E. D., F. B. Z., W. T. S., F. D., S. O. D., P. F. H., and H. B. validation; E. D., F. B. Z., W. T. S., and F. D. investigation; E. D., F. B. Z., W. T. S., and F. D. visualization; E. D., F. B. Z., W. T. S., F. D., S. O. D., C. C., and P. F. H. methodology; E. D. and H. B. writing—original draft; F. B. Z., W. T. S., F. D., S. O. D., C. C., and P. F. H. writing—review and editing; E. D., F. B. Z., W. T. S., F. D., S. O. D.,

C. C., P. F. H. and H. B. formal analysis; H. B. funding acquisition; H. B. project administration.

**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 P. F. H.).

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

**Abbreviations**—The abbreviations used are: AEP, asparaginyl endopeptidase; proAtLEG $\beta$ , *Arabidopsis thaliana* prolegumain isoform beta; AtLEG $\beta$ , *Arabidopsis thaliana* legumain isoform beta; 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; SEC, size exclusion chromatography; TEV, tobacco etch virus; ESS, electrostatic stability switch.

### 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*. *Plant Cell* **5**, 1651–1659 [CrossRef Medline](#)
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. *J. Biol. Chem.* **268**, 3525–3529 [Medline](#)
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 [CrossRef Medline](#)
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 [CrossRef Medline](#)
5. Muntz, K., and Shutov, A. D. (2002) Legumains and their functions in plants. *Trends Plant Sci.* **7**, 340–344 [CrossRef Medline](#)
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. *Plant Cell* **17**, 876–887 [CrossRef Medline](#)
7. Poncet, V., Scutt, C., Tournebize, R., Villegente, M., Cuff, G., Rajjou, L., Balliau, T., Zivy, M., Fogliani, B., Job, C., de Kochko, A., Sarramegna-Burdet, V., and Job, D. (2015) The Amborella vacuolar processing enzyme family. *Front. Plant Sci.* **6**, 618 [CrossRef Medline](#)
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 [CrossRef Medline](#)
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 [CrossRef Medline](#)
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*. *J. Biol. Chem.* **278**, 32292–32299 [CrossRef Medline](#)
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 [CrossRef Medline](#)
12. Gruis, D., Schulze, J., and Jung, R. (2004) Storage protein accumulation in the absence of the vacuolar processing enzyme family of cysteine proteases. *Plant Cell* **16**, 270–290 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  14. Conlan, B. F., Gillon, A. D., Craik, D. J., and Anderson, M. A. (2010) Circular proteins and mechanisms of cyclization. *Biopolymers* **94**, 573–583 [CrossRef](#)
  15. Craik, D. J., and Malik, U. (2013) Cyclotide biosynthesis. *Curr. Opin. Chem. Biol.* **17**, 546–554 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Chem. Biol.* **22**, 571–582 [CrossRef Medline](#)
  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. *Nat. Commun.* **6**, 10199 [CrossRef Medline](#)
  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. *J. Biol. Chem.* **282**, 29721–29728 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Plant Cell* **30**, 686–699 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  23. Craik, D. J. (2012) Host-defense activities of cyclotides. *Toxins* **4**, 139–156 [CrossRef Medline](#)
  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. *Plant J.* **53**, 505–515 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  26. Lesner, A., ěgowska, A., Wysocka, M., and Rolka, K. (2011) Sunflower trypsin inhibitor 1 as a molecular scaffold for drug discovery. *Curr. Pharm. Des.* **17**, 4308–4317 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Nat. Commun.* **9**, 2411 [CrossRef Medline](#)
  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. *J. Biol. Chem.* **293**, 8934–8946 [CrossRef Medline](#)
  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. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11737–11746 [CrossRef](#)
  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. F Struct. Biol. Cryst. Commun.* **68**, 24–31 [CrossRef Medline](#)
  32. 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 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Plant J.* **98**, 988–999 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  36. Stewart, D. E., Sarkar, A., and Wampler, J. E. (1990) Occurrence and role of cis peptide bonds in protein structures. *J. Mol. Biol.* **214**, 253–260 [CrossRef Medline](#)
  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. *Nat. Protoc.* **6**, 111–120 [CrossRef Medline](#)
  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. *Mol. Cell. Proteomics* **15**, 2515–2524 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Mol. Cell. Proteomics* **18**, 2335–2347 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  43. Kabsch, W. (2010) Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 [CrossRef Medline](#)
  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., et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  46. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 [CrossRef Medline](#)
  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 [CrossRef Medline](#)
  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. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10037–10041 [CrossRef Medline](#)
  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–W667 [CrossRef Medline](#)

## Structure and function of AtLEGB

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. *J. Biol. Chem.* **293**, 13151–13165 [CrossRef](#) [Medline](#)
51. Niesen, F. (2010) *Excel script for the analysis of protein unfolding data acquired by differential scanning fluorimetry (DSF)*. Structural Genomics Consortium, Oxford, United Kingdom.
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 [CrossRef](#) [Medline](#)
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*. *J. Biol. Chem.* **280**, 32914–32920 [CrossRef](#) [Medline](#)
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. *Anal. Biochem.* **138**, 141–143 [CrossRef](#) [Medline](#)
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., *et al.* (2017) N-degradomic analysis reveals a proteolytic network processing the podocyte cytoskeleton. *J. Am. Soc. Nephrol.* **28**, 2867–2878 [CrossRef](#) [Medline](#)
56. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 [CrossRef](#) [Medline](#)
57. Colaert, N., Helsen, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 [CrossRef](#) [Medline](#)
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–D456 [CrossRef](#)
59. Karplus, P. A., and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* **336**, 1030–1033 [CrossRef](#) [Medline](#)