Cleaving like a pro: specificity, structure, and function of Pro-Pro endopeptidases

Claushuis, B.

Cleaving like a Pro:

Specificity, structure, and function of Pro-Pro endopeptidases

Bart Claushuis

Colophon

Cover design: Luc van Bokhoven and Bart Claushuis

Layout: Bart Claushuis

Printing: Gildeprint BV, Enschede, the Netherlands

ISBN: 978-94-6496-221-5

This work was supported by an ENW-M grant (OCENW.KLEIN.103) from the Dutch Research Council (NWO).

Cleaving like a Pro:

Specificity, structure, and function of Pro-Pro endopeptidases

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 22 november 2024
klokke 13:00 uur

door

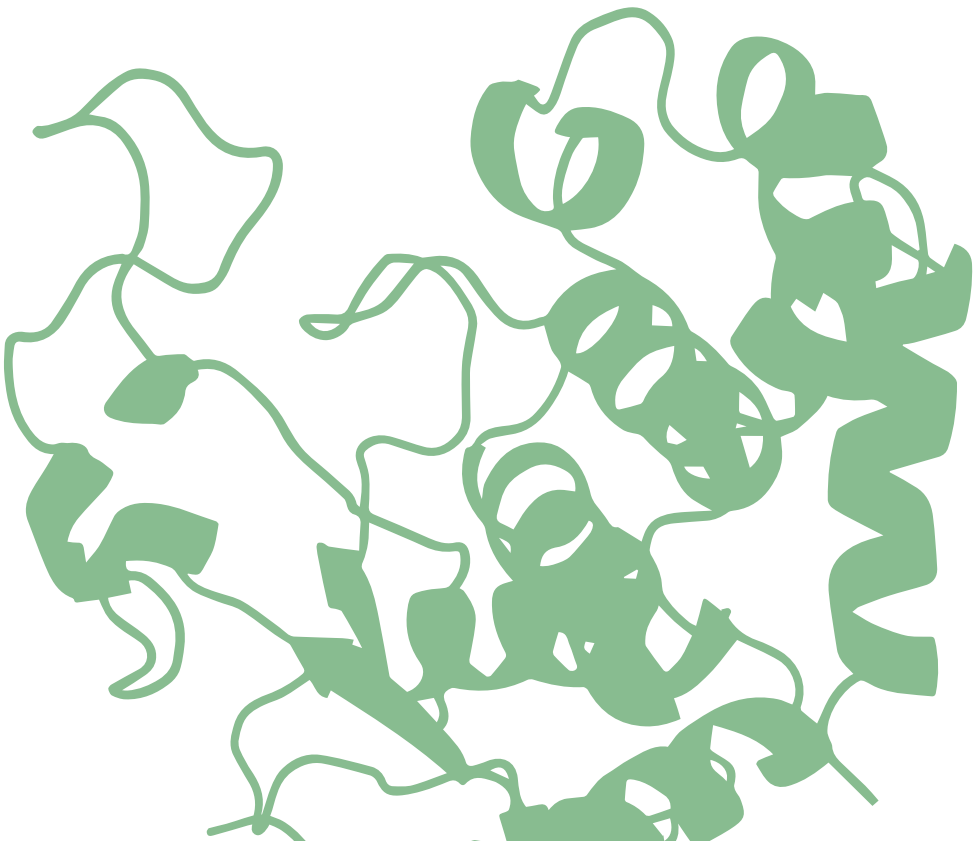
Bart Claushuis
geboren te Den Haag
in 1996

Promotor	Prof. dr. M. Wuhrer
Copromotoren	Dr. P.J. Hensbergen
	Dr. J. Corver
Leden promotiecommissie	Prof. dr. E.J. Snijder
	Prof. dr. D. Claessen (Leiden University)
	Prof. dr. N.M. van Sorge (Amsterdam University Medical Center)
	Prof. dr. U. Baumann (University of Cologne)

"Insight into universal nature provides an intellectual delight and sense of freedom that no blows of fate and no evil can destroy."

— Alexander von Humboldt

Table of contents



1

General Introduction and Thesis Outline

Bacterial life, like all life on earth, cannot exist without the group of biomolecules known as proteins. In bacteria, proteins serve as structural elements, participate in metabolic processes, transport molecules across the cell membrane, regulate gene expression, provide defense mechanisms, and are involved in communication, signaling, adhesion, motility, and pathogenesis. The versatility of proteins originates from the inherent diversity of the molecules themselves. Although all proteins are composed of the same proteinogenic amino acids, of which 20 exist (or 22 when including selenocysteine and pyrrolysine), proteins exhibit immense structural variability due to the countless possible combinations of amino acid residues. For instance, a stretch of merely ten amino acids could give rise to over 1 x 1013 different peptides. This diversity is further increased by post-translational modifications (PTMs), which are enzyme-mediated alterations of a protein. Two of these PTMs, glycosylation and, especially, proteolysis, will play a central role in this thesis. We will focus on their role in bacterial adhesion and motility, particularly in the bacterium *Clostridioides difficile*.

Glycosylation is the enzymatic process by which carbohydrates are covalently attached to proteins, lipids, or other organic molecules. This post-translational modification plays a crucial role in various bacterial functions, including cell wall formation, biofilm development, and interactions with host organisms [1]. In pathogenic bacteria, glycosylation can significantly impact virulence by modifying surface structures such as pili and flagella, aiding in immune evasion and adherence to host tissues [1,2]. Bacterial glycosylation, although less complex than those in eukaryotes in terms of branching and spatial separation of the synthetic pathway, exhibits considerable diversity involving unique glycosyltransferases and oligosaccharide structures.

Proteases

Proteases, also known as peptidases, are enzymes that catalyze the hydrolysis of peptide bonds. This enzymatic activity, also known as proteolysis, is often overlooked as a PTM. However, proteolytic processing can be crucial for the correct functioning of proteins. In addition, proteases are essential for degrading misfolded, damaged, and unwanted proteins and are therefore essential for protein homeostasis, quality control, and turnover [3,4]. In pathogenic bacteria, proteases can aid in colonization of the host or cause disease [5].

Proteases can be categorized based on their functional group at their catalytic site [6], i.e., the type of amino acid residue or co-factor involved in the hydrolytic reaction. Proteases were originally classified as serine (Ser) proteases, cysteine (Cys) proteases, aspartic (Asp) proteases, or metalloproteases. In addition, a fifth group was later

recognized that consists of threonine (Thr) proteases [7]. Ser, Cys, and Thr proteases employ their respective Ser, Cys, and Thr residues in the catalytic site as the nucleophilic residue that attacks the carbonyl group of the peptide bond [8–11], while Asp proteases activate a water molecule that in turn acts as a nucleophile and hydrolyzes the peptide bond [12]. Metalloproteases require a divalent cationic metal ion, commonly zinc, bound to amino acid ligands in the active site. Two histidine residues coordinate the metal ion while an additional glutamic acid (Glu) is necessary for the catalytic activity, collectively forming the HEXXH (sometimes HEXXXH) motif that is indicative of metalloprotease activity [13,14]. The bound metal ion, in combination with the electrophilic Glu residue, activates a water molecule that in turn acts as the nucleophile for the hydrolysis of the peptide bond [15].

A second discrimination is made between proteases that function either as endopeptidase or as exopeptidase. Endopeptidases cleave peptide bonds within a protein, while exopeptidases cleave the peptide bonds of the terminal amino acid residues. Proteolysis by an exopeptidase releases a single amino acid, dipeptide, or tripeptide from the N- or C-terminus of the substrate.

Proteases exhibit specificity for certain amino acid sequences near the cleavage site. This specificity can be broad, allowing the enzyme to hydrolyze a range of substrates, or it can be highly specific, recognizing only a few or even a single protein. For example, the promiscuous protease trypsin is known for its specificity for the carboxyl side of lysine and arginine residues, while the highly specific TEV protease recognizes a stretch of seven residues (EXXYXQ↓(G/S), ↓ indicates cleavage site) [16].

Schechter and Berger (1967) developed a nomenclature to describe the positions surrounding the cleavage site for both the protease and the substrate [17]. In this system, the substrate residues that interact with the protease are named according to their position relative to the cleavage site. The amino acid residue located directly N-terminal to the cleavage site is named P1, and the residues upstream of the P1 are named P2, P3, etc., collectively forming the non-prime-side (**Figure 1**). Similarly, the C-terminal residues are named P1’, P2’, P3’, etc., forming the prime-side (**Figure 1**). Correspondingly, the protease has binding sites or pockets that recognize these substrate residues. These pockets are named S1, S2, etc., for the non-prime-side and S1', S2', etc., for the prime-side **(Figure 1**). The specificity of a protease is largely determined by the properties of these S pockets.

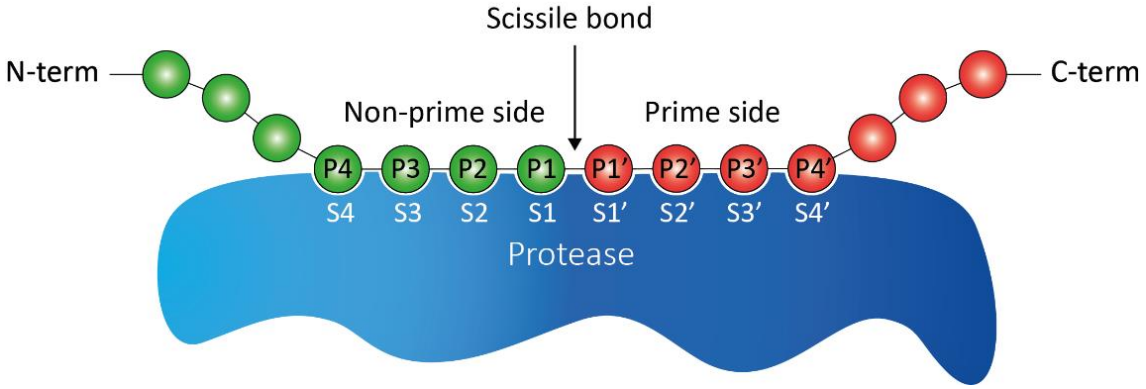


Figure 1. Nomenclature for the positions in the protease and substrate surrounding the cleavage site. The residues N-terminal of the scissile bond in the substrate form the non-prime-side and are named P1, P2, P3, etc. The C-terminal residues form the prime-side and are named P1', P2', P3', etc. Similarly, the binding sites in the protease are named S1, S2, etc. for the non-prime-side and S1', S2, etc. for the prime-side.

There are several methods to study protease activity and specificity. A straightforward approach involves the use of Fluorescence Resonance Energy Transfer (FRET) peptides that possess a fluorescent group and a quencher [18]. Cleavage of these peptides by a protease releases the quencher from the fluorescent group, which can be detected by measuring the increase in fluorescence. This method allows for detection and quantification of activity, but also for monitoring the activity real-time when measuring fluorescence during a period of time. However, the major drawbacks of this method are that the FRET peptides have to be synthesized and tested separately and that this method does not determine the cleavage site, i.e., it only indicates if a peptide is cleaved or not.

To determine the cleavage site in a highly specific manner, mass spectrometry (MS)-based approaches can be used. For example, cleaved (FRET) peptides can be analyzed by matrix-assisted laser desorption/ionization time of flight (MALDI-ToF) MS, which provides information on the cleavage site. Other methods used to study protease activity and specificity involve analysis using liquid chromatography with tandem mass spectrometry (LC-MS/MS). LC-MS/MS has been very valuable in protease research and many methods to study protease activity and specificity have been developed that use this technique including a subtiligase-based method [19], COFRADIC [20], PS-SCL [21], MSP-MS [22], TAILS [22,23], and PICS [24]. These methods generally involve chemically labeling protease-generated peptides to distinguish them from other peptides or to enrich them through selection procedures. For example, in subtiligase-based methods, a biotin is attached to (neo)-N-termini, which allows the capture of these peptides using avidin or streptavidin [19]. Analysis using LC-MS/MS offers numerous advantages. Unlike FRET peptides, which require separate testing, LC-MS/MS is capable of analyzing highly complex samples. The most commonly used chromatography method for this purpose

is reversed-phase high-performance liquid chromatography (RP-HPLC), which separates components based on their hydrophobicity. Although the separation of complex mixtures is often incomplete, the first mass analyzer (MS1) separates peptides based on their mass-to-charge (m/z) ratio, while also determining their intensity (quantity). Typically, data-dependent acquisition (DDA) is employed, where selected peptides (precursors) undergo fragmentation through collisions with gas molecules. These fragments are then analyzed in a second mass spectrometry event (MS2), providing detailed information about the sequence of the precursor peptide. In protease research, LC-MS/MS enables highly specific determination of the cleavage site and the surrounding N- or C-terminal residues. Additionally, LC-MS/MS is extremely sensitive, capable of detecting peptides at attomolar levels [25].

Clostridioides difficile infection and physiology

Clostridioides difficile, formerly known as *Clostridium difficile* or *Peptoclostridium difficile* [26], is the leading causative agent of antibiotic-associated diarrhea and colitis [27,28]. The bacterium is an opportunistic pathogen of the gastrointestinal (GI) tract, and patients with a disturbed microbiome, often due to treatment with antibiotics, are especially susceptible to *C. difficile* infection (CDI) [29,30]. *C. difficile* is a gram-positive, rod-shaped bacterium, and, due to its obligate anaerobic character, primarily infects the oxygen-poor environment of the mammalian colon [31–33], although infection of avian species has also been reported [34,35].

The life cycle of *C. difficile* starts as a dormant spore that, upon ingestion by the host, traverses the GI tract to the terminal region of the small intestine (**Figure 2**). Here, several germinants stimulate the germination of spores [36], a process that is inhibited by a healthy gut microbiota [37,38]. Once germinated, the vegetative *C. difficile* cells colonize the intestinal tract by adhering to the epithelial cells. Once *C. difficile* has established a population of proliferating cells, the bacteria start to produce toxins that damage the host's colonic epithelium, resulting in symptoms that include diarrhea, inflammation of the colon, pseudomembranous colitis, and toxic megacolon, which can be fatal [27]. Additionally, during the vegetative growth phase, the cells initiate spore production, enabling *C. difficile* to transmit to new hosts through the fecal-oral route.

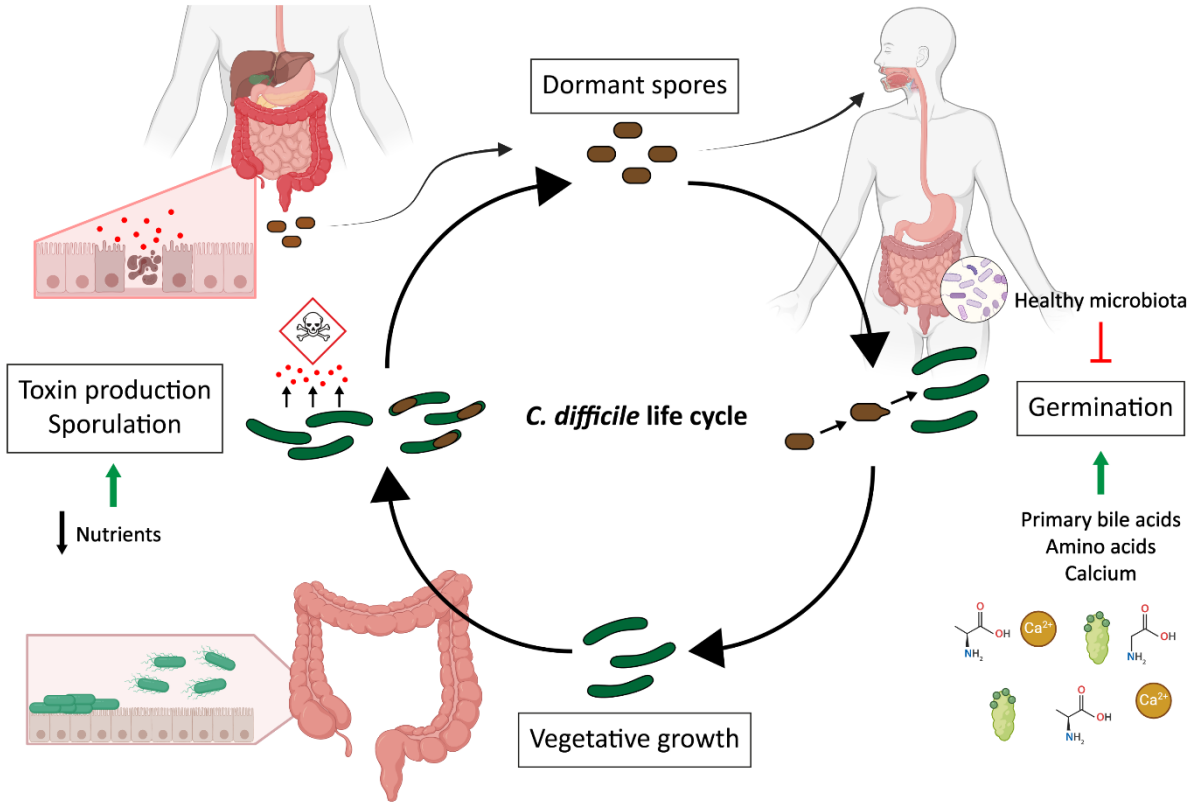


Figure 2. The life cycle of *C. difficile*. Spores enter the host by ingestion and traverse the GI tract to the distal part of the small intestine. Germination of spores is stimulated by germinants but is inhibited by a healthy microbiota. Following germination, vegetative growth of *C. difficile* commences in the colon. Environmental signals such as nutrient limitation induce the processes of toxin production and sporulation. The toxins disrupt the epithelial layer and cause apoptosis. Spores of *C. difficile* exit via the fecal route for transmission to new hosts. The image was created with BioRender.com.

Toxins

C. difficile produces two main toxins, toxin A (TcdA) and toxin B (TcdB), which are crucial for its pathogenicity [39–41]. However, some strains encode only a single toxin [42–45]. Some strains encode a third toxin, the binary toxin *C. difficile* transferase (CDT) [46,47]. Additionally, non-toxinogenic strains lacking the genes for toxin have been identified, which do not cause the symptoms associated with CDI [48,49].

The major virulence factors TcdA and TcdB function by inactivation of Rho-family GTPases through glucosylation, which leads to disruption of the actin cytoskeleton in epithelial cells. This disruption leads to an alteration of cell morphology, breakdown of tight junctions, and cell death [50]. In addition, the loss of the epithelial barrier function promotes further bacterial invasion into the tissue and provokes inflammation [51–53].

The toxins are encoded within a chromosomal region known as the pathogenicity locus (PaLoc) [54]. The other genes located in the PaLoc are *tcdC*, *tcdE*, and *tcdR*. The product of *tcdR* is an alternative sigma factor that regulates the expression of *tcdA*, *tcdB*, *tcdE,* and *tcdR* itself [55]. TcdE is a holin-like protein needed for the secretion of the glucosylating toxins [56]. TcdC is thought to inhibit toxin expression [57], but its role as a negative regulator is under debate [58–60].

The regulation of toxin expression through TcdR is complex, involving various transcriptional regulators that respond to environmental signals and physiological stages of the bacterium [61–65]. One of the environmental factors that regulate toxin production is nutrient availability. Expression of *tcdA* and *tcdB* is regulated through the global transcriptional regulators CcpA and CodY, which respond to the environmental nutrient availability [66–68]. When nutrients are available, CcpA represses toxin production by binding the regulatory regions of the *tcdA* and *tcdB* genes [69], while CodY binds the promoter region of *tcdR* [66].

Sporulation

Since *C. difficile* is not able to survive in the presence of oxygen, the formation of spores is essential for transmission to new hosts [70]. In addition, spores promote the recurrence of CDI within the same host after treatment [70–72].

The initiation of sporulation is tightly controlled by various regulators and is largely dependent on nutrient availability. The master regulator of sporulation, Spo0A, controls the transcription of hundreds of genes, including the genes necessary to initiate sporulation [73,74]. Spo0A-mediated transcription is controlled by the phosphorylation of Spo0A [75], however, the mechanisms that activate Spo0A are poorly understood [62,75–77]. Two other regulators of sporulation are the nutritional regulators CcpA and CodY, which also repress toxin production. CcpA, the regulator of carbohydrate metabolism, represses the expression of Spo0A and SigF in the presence of glucose [68]. CodY also represses sporulation, since the deletion of *codY* causes an earlier onset and higher frequency of sporulation [78]. Another regulator of sporulation is RstA, a bifunctional protein that promotes sporulation while repressing toxin production and motility [79].

C. difficile produces endospores, a process that starts with asymmetric cell division after the initiation of sporulation. Following DNA replication, both the mother cell and the forespore obtain a copy of the chromosome. Then, the transcription of compartment-dependent sporulation genes proceeds first by activity of SigF and subsequently SigG in the forespore and through SigE and then SigK in the mother cell [80,81]. This ultimately

results in the engulfment of the forespore by the mother cell, followed by the addition of layers of protein around the spore that form the cortex, coat, and exosporium [80,81].

The core of the spore contains the DNA, RNA, and most enzymes needed for germination. These molecules are protected by a low water content, small acid-soluble proteins (SASPs), and dipicolonic acid chelated with calcium (Ca-DPA) [80,82–85]. The core is enveloped by a rigid cell membrane that exhibits very low permeability [86]. The cortex surrounding the membrane is composed of a modified peptidoglycan layer and is involved in spore dehydration and resistance [87,88]. On top of the cortex, a multilayered, proteinaceous structure is deposited that forms the spore coat. This spore coat consists of morphogenetic proteins that confer resistance to various environmental stresses [89–91]. The outermost layer of *C. difficile* spores is the exosporium layer, whose proteins are involved in assembling the spore coat and exosporium, conferring resistance to environmental challenges, and adhesion to the colonic epithelium [92,93].

Germination

Metabolically dormant spores can resume vegetative growth by a process called germination. Spores that have entered the GI tract of the host germinate when the conditions are favorable, primarily in the distal part of the small intestine [36]. The environmental conditions and position in the GI tract are sensed through receptors that recognize germinants, the small molecules that stimulate germination. These include bile acids, different types of amino acids, and calcium [36,94,95]. In addition, factors such as temperature and pH influence the onset of germination [36,94].

The recognition of bile acids, primarily taurocholate, by the pseudoprotease CspC triggers a series of events that ultimately lead to the activation of SleC, which hydrolyzes the cortex [96,97]. This causes both the release of Ca-DPA from the spore core and the rehydration of the core [97]. The process of germination is further dependent on the presence of Ca^{2+} and other proteins including GerS, YabG, and SpoVAC [98–101]. Collectively, these mechanisms allow the cell to resume metabolic activity and vegetative cell growth.

Adhesion and motility

Adhesion of *C. difficile* to the host's colonic epithelium is critical for successful colonization of the colon. Adhesion allows the cells to stay in place and divide until they reach higher cell numbers. In addition, these higher cell densities allow for intercellular

communication, i.e., quorum-sensing, that regulates several virulence factors such as biofilm formation, sporulation, and toxin production [102–104].

At the start of the *C. difficile* life cycle, the spores possess adhesive properties that allow them to adhere to epithelial cells [105]. The exosporium protein CdeC stimulates adhesion to the gut wall since CdeC-deficient spores exhibit a decreased adherence [106]. In addition, spores have been shown to adhere to fibronectin and vitronectin by the exosporium protein BclA3, which contributes to the recurrence of CDI [107,108], and to E-cadherin, which becomes accessible after cell-cell disruption by TcdA and TcdB [109].

The processes of adhesion and motility in vegetative cells of *C. difficile* are inversely regulated through cellular levels of the second messenger cyclic diguanylate (c-di-GMP) [110]. C-di-GMP regulates gene expression of genes by binding segments of mRNA called riboswitches [111,112]. Two types of riboswitches exist that either prematurely terminate transcription (Type I) or promote transcription (Type II) at elevated levels of c-di-GMP [111,113]. The levels of c-di-GMP are influenced by nutrient availability and are regulated through CodY-mediated transcription of genes involved in c-di-GMP synthesis and degradation [64,114–116]. Consequently, the presence of nutrients increases c-di-GMP levels in the cell, thereby stimulating the expression of proteins involved in adhesion and biofilm formation through Type II riboswitches, while reducing the expression of the genes promoting motility with Type I riboswitches [110,116–119]. Conversely, when nutrient availability is limited, the lower levels of c-di-GMP promote motility while inhibiting adhesion and biofilm formation.

Several cell wall proteins of *C. difficile* possess adhesive properties against various host ligands. The most abundant cell wall protein that forms the surface layer (S-layer) of *C. difficile*, SlpA, binds various components of the intestinal epithelium [120–122]. SlpA is a member of a family of 29 cell wall proteins (CWPs) that possess three cell wall binding 2 (CWB2) domains [123]. The CWPs are located in the S-layer and several possess adhesive properties [124–126]. In addition, other proteins have been involved in adherence, including Type IV pili which promote adherence and autoaggregation [117,127], the fibronectin-binding protein Fbp68 [128], the collagen-binding CbpA [129], and the tyrosine-transporting lipoprotein CD0873 that also displays adhesive properties [130,131].

Limited nutrient availability and noxious environments necessitate switching from a sessile to a motile state in order to migrate toward more favorable environments. For this, *C. difficile* possesses peritrichous flagella that facilitate swimming motility in liquid environments. In addition, flagella have been implicated in processes such as adherence [132] and immunomodulation [133]. The structure of these flagella consists of a basal

body (the flagellar motor), a hook, and the filament. The largest part of the flagellum consists of repeating units of the filament protein FliC, which forms the complete filament except for the tip, which is formed by FliD [134]. In many species, e.g., *Campylobacter jejuni* and *Helicobacter pylori*, glycosylation of the FliC units is important for flagellar assembly and function [135–138]. Also in *Pseudomonas aeruginosa*, FliC is modified with a glycan structure that is strain-dependent [139,140]. Two types are known, Type A and Type B, which differ largely in structure, size, and biosynthesis [139,140].

Also *C. difficile* has been shown to glycosylate FliC in a strain-dependent manner. Like in *P. aeruginosa*, two types have been identified; a smaller and relatively simple Type A structure is present in several *C. difficile* strains including 630Δ*erm* and a larger, more complex Type B structure is observed in hypervirulent strains including R20291 [141–143]. The Type A glycan structure consists of an *O*-linked *N*-acetylglucosamine (GlcNAc) that is linked to N-methyl-L-threonine through a phosphodiester bond [142], reminiscent of the Type B structure in *P. aeruginosa* [140]. *C. difficile* 630Δ*erm* mutants that lack the Type A glycan are impaired in motility and have been shown to aggregate [141,142]. The more complex structure of the Type B glycan has also been elucidated, and mutants that fail to produce this glycan exhibit diminished motility and increased cell aggregation, biofilm formation, and adherence to Caco-2 cells [143,144].

A gene cluster of five genes is thought to be responsible for the biosynthesis of the Type A glycan on FliC [141,142]. Insertional mutation of the individual genes in this cluster results in a non-motile phenotype, showing the importance of this modification for bacterial motility. A model has been proposed for the biosynthesis pathway of this modification but the exact functions of the individual genes are unclear [142].

Regulation of adhesion and motility by PPEP-1 and its substrates CD2381 and CD3246

Importantly, another well-defined mechanism exists that regulates adhesion and motility in *C. difficile*. In addition to the proteins with adhesive properties discussed above, *C. difficile* also produces the collagen-binding protein CD2831. This extracellular protein is tethered to the peptidoglycan through its LPXTG-like motif, which acts as a cell wall sorting signal, and contains a collagen hug domain (**Figure 3**) [145–147]. Likewise, another adhesion protein that possesses an LPXTG-like motif, CD3246, exists in *C. difficile*. However, CD3246 possesses a class 2 thioester domain (TED) which can facilitate covalent attachment to host cells through a cross-linking reaction (**Figure 3**), though its specific ligand remains unidentified [148,149].

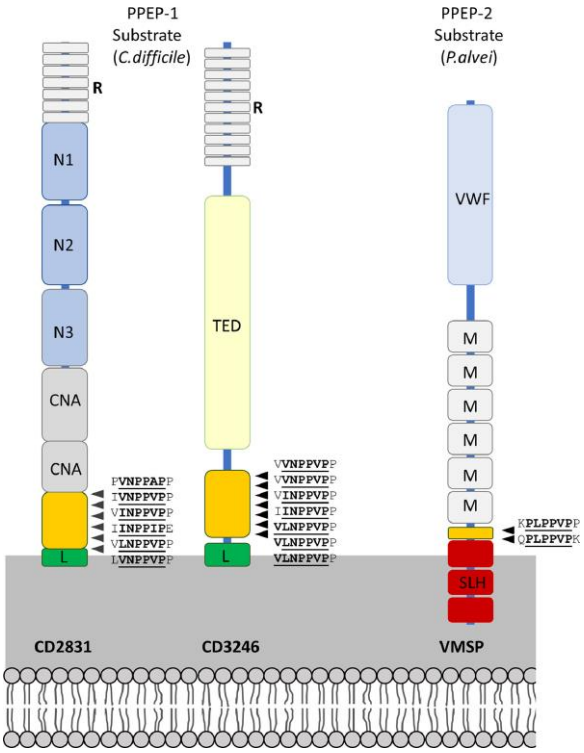


Figure 3. Domain organization of the substrates of PPEP-1 and PPEP-2. The PPEP-1 substrates CD2831 and CD3246 from *C. difficile* are attached to the peptidoglycan through their LPXTG-like motif (green). Both substrates possess multiple PPEP-1 cleavage sites close to the cell wall (recognition site is underlined). The N-terminal domains N1 and N2 in CD2831 are predicted to form a collagen hug domain, while the CNA domains likely serve as a stalk. CD3246 is predicted to possess a TED domain. The substrate of PPEP-2, VMSP, is tethered to the cell wall via three SLH domain repeats and contains two PPEP-2 cleavage sites close to the cell wall. VMSP is predicted to bind ligands via the VWF domain. R=repetitive sequence, M=MucBP-repeats. Figure was copied from van Leeuwen *et al.* (2021) [148], which is available under Creative Commons Attribution 4.0 International License CC BY-NC-ND 4.0 DEED.

The expression of both CD2831 and CD3246 is regulated by c-di-GMP levels since a Type II riboswitch is found upstream of *cd2831* and *cd3246* [119]. Consequently, high levels of c-di-GMP, and thus nutrients, promote adhesion through these proteins. The gene directly upstream of *cd2831*, i.e., *cd2830*, possesses a Type I riboswitch and is therefore inversely regulated [119]. The *cd2830* gene encodes a metalloprotease now known as PPEP-1, which has been shown to cleave both CD2831 and CD3246 [146].

In the model for the regulation of adhesion and motility by PPEP-1 and its substrates, high c-di-GMP levels stimulate adhesion to the host cells by the production of CD2831 and CD3246. Lower levels of c-di-GMP, e.g., due to nutrient stress, induce the expression of PPEP-1 (**Figure 4**). The protease cleaves its substrates, i.e., CD2831 and CD3246, close to the cell surface, thereby releasing the cells from the intestinal epithelium. Lower levels of c-di-GMP likewise promote the production of flagella, thus enhancing the motility of *C. difficile* (**Figure 4**).

A PPEP-1 ClosTron mutant (*ppep-1*::CT) has been shown to display attenuated virulence in a hamster model, as the inability to produce PPEP-1 at least doubled the survival time in hamsters [147]. The attenuated virulence of the *ppep-1*::CT strain is likely due to decreased motility of *C. difficile* and thus less efficient colonization of the colon.

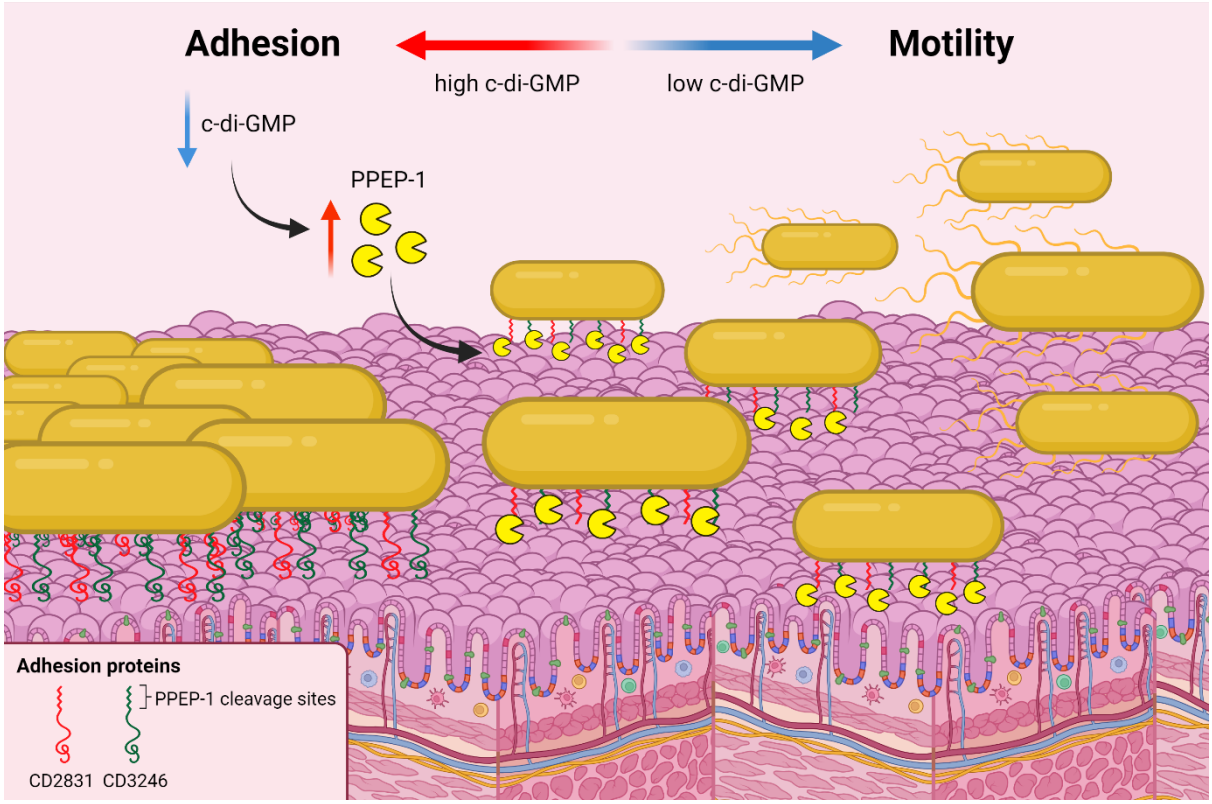


Figure 4. Model for the regulation of adhesion and motility by PPEP-1 and the substrates CD2831 and CD3246. *C. difficile* adheres to the intestinal epithelium through the adhesion proteins CD2831 and CD3246. A decrease in c-di-GMP stimulates the production of PPEP-1, which cleaves the protein anchors close to the cell wall, thereby releasing the cells. Similarly, lower c-di-GMP levels promote the production of flagella. Image was created with BioRender.com.

Pro-Pro endopeptidases (PPEPs)

PPEP-1, formerly known as CD2830 or Zmp1 [146,150], was the first member of a family of metalloproteases known as the Pro-Pro endopeptidases. First identified in a *C. difficile* secretome study, PPEP-1 was noticed due to its homology with the anthrax toxin lethal factor (ATLF) catalytic domain which is essential for the lethal activity of the toxin [146,151]. With PPEP-1 being a secreted protein, a search for substrates in colonic epithelial cells revealed that heat shock protein (HSP) 90β was cleaved between two alanine residues (PNA↓AVP, P3-P3') by PPEP-1 [146]. Further investigations into the specificity of PPEP-1 using a library of synthetic peptides revealed that this enzyme

preferred to cleave substrates between proline residues instead of alanines [146]. This observation was remarkable since the cyclic structure of proline produces conformational constraints that often prevent proteolysis at proline-containing sites [152–154]. Although X-Pro and Pro-X endopeptidase activity had been observed before, no proteolytic enzyme was known to hydrolyze substrates between two proline residues [152,155,156]. Therefore, the enzyme with the unique ability to cleave proline-proline bonds was named Pro-Pro endopeptidase 1 (PPEP-1) (**Figure 5**).

Apart from the prolines at the P1 and P1' positions, a Pro at P3' was shown to be an important determinant for activity by PPEP-1 [146]. A search for the sequence PPXP (P1-P3') in the proteome of *C. difficile* revealed that the PPEP-1 substrates CD2831 and CD3246 contain multiple PPEP-1 cleavage sites, six in CD2831 and seven in CD3246. Based on these data, the PPEP-1 consensus cleavage motif (V,I,L)NP↓P(V,I,A)P (P3-P3') could be determined (**Figure 5**), indicating that there is some flexibility at the P3 and P2' position.

Since the discovery of PPEP-1, PPEP homologs have been identified in other species of bacteria. A phylogenetic analysis revealed the presence of proteins containing a PPEP domain in over 130 species spread over 9 genera, namely the genera *Clostridioides*, *Clostridium*, *Paenibacillus*, *Bacillus*, *Parageobacillus*, *Geobacillus*, *Anoxybacillus*, *Salinicoccus*, and *Jeotgalicoccus* [148]. The species differ largely in lifestyle and include enteric pathogens, commensal microbiota, plant root-associated bacteria, and soil-dwelling species [148]. For the latter, PPEP domain-containing proteins are found in halo-, alkalo-, and thermophilic bacteria [148].

A

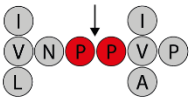
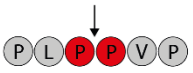
Protein name/ Ordered locus name	Organism	Substrate	Cleavage motif
PPEP-1	*Clostridioides difficile*	CD2831 and CD3246 (*C. difficile*)	
PPEP-2	*Paenibacillus alvei*	VMSP (*P. alvei*)	
CD1597/ CD630_15970	*Clostridioides difficile*	Unknown	Unknown
GTNG_1672	*Geobacillus thermodenitrificans*	Unknown	Unknown
HNQ34_002771	*Anyxobacillus tepidamans*	Unknown	Unknown

Figure 5. Overview of PPEPs. A) An overview of several PPEPs. For PPEP-1 and PPEP-2, the endogenous substrates have been identified. The consensus cleavage motif based on these substrates is shown. The PPEP homologs CD1597, GTNG_1672, and HNQ34_002771 have not been characterized. **B)** The experimentally determined cocrystal structure of PPEP-1 (PDB: 6R5C) and the apo structure of PPEP-2 (PDB: 6FPC) and the predicted structures of CD1597 (Uniprot ID: Q186F3), GTNG_1672 (Uniprot ID: A4INY2), and HNQ34_002771 (Uniprot ID: A0A7W8IRZ3). All structures (for CD1597 only the PPEP-like domain) have been superimposed to highlight the similarities in protein folds.

The second PPEP that was characterized was PPEP-2 from *Paenibacillus alvei* (**Figure 5**) [157]. *P. alvei* is an anaerobic, Gram-positive, and endospore-forming bacterium known as a secondary invader of honeybees and associated with foulbrood [158]. Similar to *C.*

difficile, PPEP-2 and its substrate, VMSP, are encoded by adjacent genes. VMSP is an extracellular protein that is tethered to the cell wall through its surface-layer homology (SLH) domain repeats [159] and is likely involved in binding extracellular matrix proteins through its Von Willebrand factor type A (VWFA) domain (**Figure 2**), yet the binding ligand is unknown [157]. VMSP contains two PPEP-2 cleavage sites close to the SLH domain repeats (**Figure 2**) with the sequence PLP↓PVP (P3-P3') [157]. Although the PPEP-2 cleavage site resembles that of PPEP-1, i.e., a PPXP (P1-P3') motif, PPEP-1 is not able to cleave a PPEP-2 substrate and vice versa, indicating a clear difference in the specificity of the two PPEPs [157].

For PPEP-1 and PPEP-2, protein structures have been experimentally determined [157,160–162]. Although the amino acid identity is only 50%, the proteases share a highly similar fold [157]. The active site α4-helix that contains the HEXXH motif separates the N-terminal domain (NTD) and the C-terminal domain (CTD) (**Figure 6**). The NTD possesses a flexible S-loop that is involved in substrate recognition and closes upon substrate binding (**Figure 6**) [160]. The NTD also features the so-called "diverting loop" that restricts the substrate from exiting the active site [160]. To overcome this steric hindrance, the prime-side of the substrate needs to adopt a unique double-kinked conformation that is produced by the Pro at P1' and the Val at P2', which is one of the determinants of the Pro-Pro specificity (**Figure 6**) [160]. The largest structural difference between PPEP-1 and PPEP-2 is found at the β3/β4-loop which is in close proximity to the P3 and P2 positions in the substrate [157]. In PPEP-2, the Glu-113 forms a salt bridge with Arg-145, thereby causing a steric hindrance at the P3 position [157]. The Pro at the P3 positions in the substrate peptide PLP↓PVP (P3-P3') of PPEP-2 produces a kink that directs the substrate away from the salt bridge [157]. The difference in β3/β4-loop between PPEP-1 and PPEP-2 explains the distinct substrate specificity of the two proteases [146,157]. Substitution of the PPEP-2 β3/β4-loop for that of PPEP-1 shifts the specificity for the P3 residue from a Pro to a Val, thus making the specificity more PPEP-1-like [157].

Although PPEP-1 and PPEP-2 are very similar in domain organization, other PPEP homologs possess additional domains [148]. The PPEPs from several species of *Bacillus* contain a predicted SH3b domain, which is known to bind peptidoglycan [148,163]. In species of *Salinococcus* and *Jeotgalicoccus*, PPEP homologs harbor a fibronectin type 3 (FN3) domain [148], although the function of FN3 domains in bacteria remains ambiguous [164,165]. Furthermore, two members of the genus *Paenibacillus* possess an SLH domain that allows them to bind non-covalently to the cell surface [159]. The absence of this domain in other species of *Paenibacillus* indicates an accessory rather than an essential element of these PPEP homologs [148].

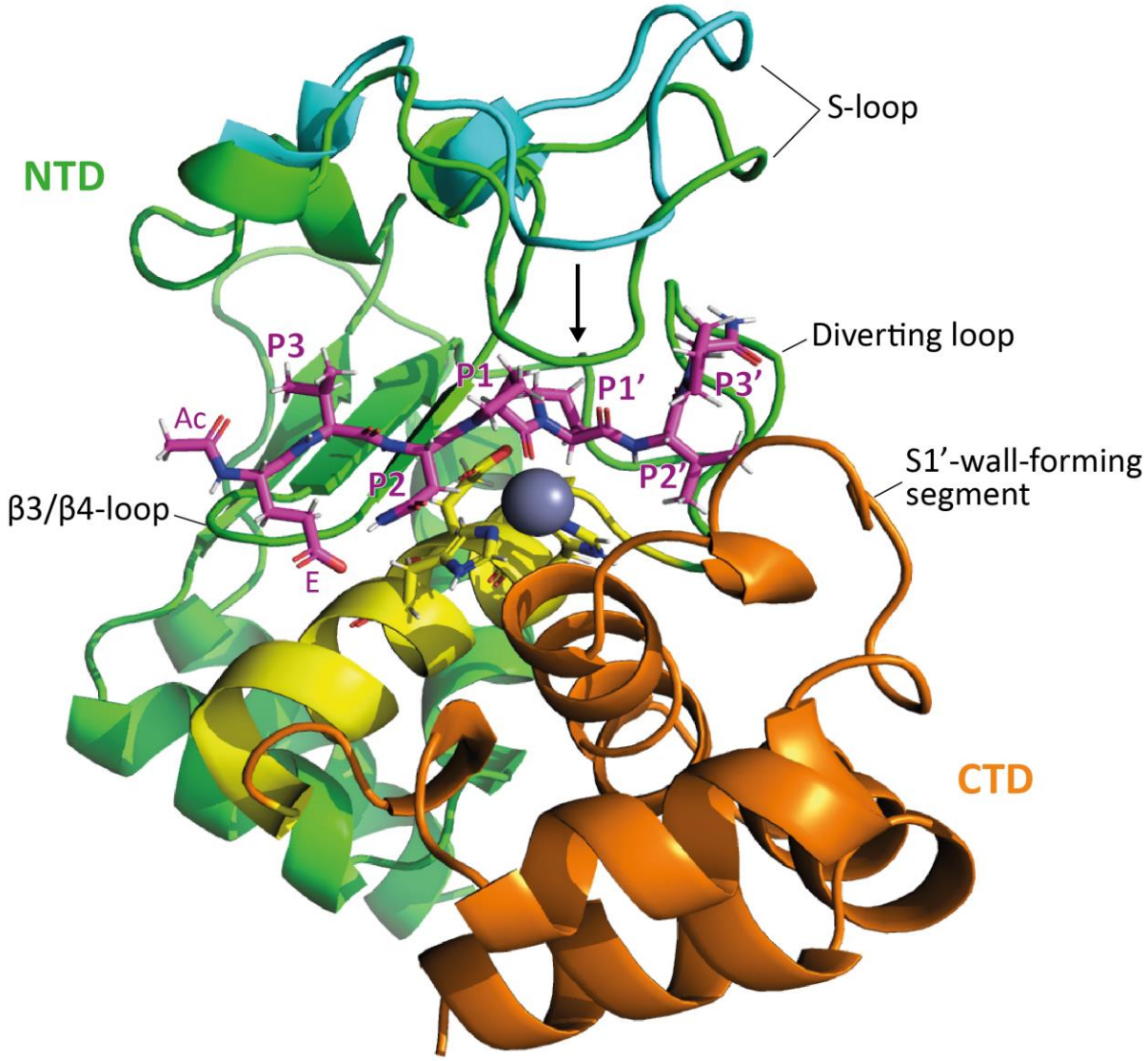


Figure 6. Cocrystal structure of PPEP-1. Structure of PPEP-1 with the substrate Ac-EVNPPVP-NH_2 (PDB: 6R5C). The N-terminal domain (NTD) is shown in green, the active-site helix α4 in yellow, the N-terminal domain (NTD) in orange, and the substrate in magenta. The S-loop from the PPEP-1 apo-structure (PDB: 5A0P) is shown in cyan. The zinc ion binding residues and the substrate are shown in sticks. The arrow indicates the S-loop movement upon substrate binding.

Another PPEP homolog that is distinct from the typical PPEPs is CD1597 from *C. difficile* (**Figure 5**). Although the protein possesses a PPEP-like domain, this PPEP homolog does not contain a signal peptide for secretion, suggesting a fundamentally different function from other PPEPs. Moreover, this PPEP homolog contains a substantial N-terminal domain of unknown function, accounting for approximately half of the protein **(Figure 5B**). Its unique non-secretory nature, the presence of an uncharacterized domain, and being a second PPEP homolog in *C. difficile* make CD1597 an interesting subject for further investigation.

PPEP-1 and PPEP-2 both cleave adhesion proteins encoded by adjacent genes. For other PPEPs, substrate prediction seems less straightforward as the adjacent genes do not encode adhesion proteins or canonical PPEP cleavage sites, i.e., PPXP (P1-P3') [148]. For example, the PPEP homolog that is phylogenetically the most distant from PPEP-1, a PPEP from *Geobacillus thermodenitrificans* (GTNG_1672, **Figure 5**), is encoded by a gene adjacent to a gene encoding a putative adhesin, a YpjP-like protein, which lacks a canonical PPEP cleavage site [148]. If this putative adhesin represents the endogenous substrate of this PPEP, the specificity should differ from that of PPEP-1 and PPEP-2. The PPEP homolog from *Anoxybacillus tepidamans*, an organism closely related to *G. thermodenitrificans*, has no secreted proteins encoded adjacent to the gene encoding the PPEP homolog (HNQ34_002771, **Figure 5**). Therefore, the substrate for this PPEP is likely located elsewhere on the genome or, alternatively, encoded by another organism than *A. tepidamans*. Although substrate prediction for the PPEP homologs from *G. thermodenitrificans* and *A. tepidamans* is less straightforward than for PPEP-1 and PPEP-2, their similar folds (**Figure 5B**) suggests a PPEP-like substrate specificity, i.e., a PPXP motif (P1-P3').

To identify substrates for orphan PPEPs that lack an obvious substrate candidate, a similar approach can be taken as was done for PPEP-1. In this approach, the cleavage specificity of PPEP-1 was determined using a library of synthetic peptides and a search for the resulting cleavage motif in the *C. difficile* proteome led to the identification of the two substrates CD2381 and CD3246 [146]. Similar approaches exist that profile the specificity of proteases and might prove valuable in uncovering the substrates for other PPEP homologs. For example, the branch of MS-based proteomics termed N-terminomics focuses on identifying neo-N-termini generated by proteases and provides information on the substrate specificity [20,23,166–168]. Alternatively, proteolytic digestion of peptide libraries that can be proteome-derived, produced synthetically, or generated by phage-display technologies can be employed to profile the specificity of PPEPs [22,24,169–172]. Apart from the potential for substrate identification, data on PPEP specificity, when combined with structural data, can aid in understanding the structure-function relationships of PPEPs.

Thesis outline

The research described in this thesis aims to uncover the roles of bacterial enzymes involved in the processes of adhesion and motility in bacteria, with an emphasis on the enzymes' substrate specificities. The studies presented hereafter can be divided into three parts. In the first part, we aimed to elucidate the biosynthetic pathway for a glycan structure essential for motility in *C. difficile* (**Chapter 2**). In the second part, we investigated the function and enzymatic activity of a PPEP homolog, CD1597, present in *C. difficile* (**Chapter 3**). Lastly, we developed a method that allows for an in-depth characterization of PPEP substrate specificity that we employed to study the structure-function relationships of PPEPs and predict substrates for previously uncharacterized PPEPs (**Chapters 4, 5 and 6**).

Chapter 2 aimed to elucidate the biosynthetic pathway for the Type A glycan modification found on FliC in *C. difficile* 630Δ*erm*. Previous studies have shown the importance of the Type A glycan for motility, but failed to specify a role for one of the proteins involved in the biosynthesis, i.e., CD0244, since this protein was deemed non-essential for producing the Type A glycan. Furthermore, no detailed enzymatic functions and biosynthetic intermediates have been predicted for the biosynthetic pathway. Using quantitative MS-based proteomics analyses, we investigated the importance of CD0244 for Type A synthesis. Moreover, we predicted detailed enzymatic activities and biosynthetic intermediates in the Type A biosynthetic pathway, using bioinformatic analyses and structural comparisons. We proposed a revised model for the Type A glycan biosynthesis that serves as a basis for future research.

In **Chapter 3** we investigated the role of CD1597 in *C. difficile*. Based on the homology with other PPEPs, it was hypothesized that CD1597 displays similar endoproteolytic activity. However, the distinct characteristics of this protein suggested a markedly different function for CD1597. Using purified recombinant CD1597, we tested the proteolytic activity of CD1597 for several potential substrates. In addition, the function of CD1597 was investigated by generation of a *cd1597* insertional mutant to determine the effects of the inability to produce CD1597. By employing various phenotypical assays, microscopy, and comparative proteomics analyses, we aimed to uncover the role of this enigmatic protein.

Information on the substrate specificity of a protease can aid in identifying biologically relevant substrates. In addition, a thorough understanding of substrate specificity is needed for the application of proteases in research, healthcare, and industry. In **Chapter 4** we sought to develop a novel method for profiling PPEP specificity in high detail. For this, we combined the advantages of a synthetic combinatorial peptide library, i.e., high diversity and equimolar peptide concentrations, with the sensitivity and

specificity of MS detection. We used this approach to characterize the prime-side specificity of several PPEPs, which included PPEP-1, PPEP-2, and a novel PPEP from *Geobacillus thermodenitrificans*.

The substrate specificity of PPEPs does not only depend on the prime-side residues, but also on those at the non-prime-side. Therefore, in **Chapter 5,** we expanded our newly developed method for profiling PPEP specificity by synthesizing a complimentary combinatorial peptide library. Using this new library, we investigated the non-prime-side specificity of PPEPs. Moreover, we aimed to characterize the full PPEP specificity in a single experiment. We used the expanded method to determine the specificity of known PPEPs, PPEP mutants, and a novel PPEP from *Anoxybacillus tepidamans*. In addition, we tested our libraries with CD1597. By combining the data on PPEP specificity with structural information, we shed more light on the structure-function relationship of PPEPs.

The crystal structures of PPEP-1 and PPEP-2 allowed us to explain their substrate specificity on an atomic level. For other PPEPs, however, structures have not been experimentally determined. In **Chapter 6** we unraveled the atomic cocrystal structure of PPEP-3 from *G. thermodenitrificans* as determined by X-ray crystallography. We investigated the protease-substrate complex and highlighted the similarities and differences to other PPEPs for which an experimental structure is available. In addition, we characterized the PPEP-3 specificity using the combinatorial synthetic library method. Together, these data were used to explain the preference of PPEP-3 for certain residues surrounding the cleavage site.

Chapter 7 reflects on our findings, discusses the topics presented in this thesis, and provides a framework for future research.



2

A revised model for the Type A glycan biosynthetic pathway in *Clostridioides difficile* strain 630Δ*erm* based on quantitative proteomics of *cd0241-cd0244* mutant strains

Bart Claushuis[1], Arnoud H. de Ru[1], Sarah A. Rotman[1], Peter A. van Veelen[1], Lisa F. Dawson[2], Brendan W. Wren[2], Jeroen Corver[3], Wiep Klaas Smits[3], Paul J. Hensbergen[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom
[3] Department of Medical Microbiology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

Published in ACS Infectious Diseases, 2023, 9, 12, 2665-2674

DOI: 10.1021/acsinfecdis.3c00485

Abstract

The bacterial flagellum is involved in a variety of processes including motility, adherence and immunomodulation. In the *Clostridioides difficile* strain 630Δ*erm*, the main filamentous component, FliC, is post-translationally modified with an *O*-linked Type A glycan structure. This modification is essential for flagellar function since motility is seriously impaired in gene mutants with improper biosynthesis of the Type A glycan. The *cd0240-cd0244* gene cluster encodes the Type A structure, but the role of each gene, and the corresponding enzymatic activity, has not been fully elucidated. Using quantitative mass spectrometry-based proteomics analyses, we determined the relative abundance of the observed glycan variations of the Type A structure in *cd0241*, *cd0242*, *cd0243* and *cd0244* mutant strains. Our data not only confirm the importance of CD0241, CD0242 and CD0243, but, in contrast to previous data, also show that CD0244 is essential for the biosynthesis of the Type A modification. Combined with additional bio-informatic analyses, we propose a revised model for Type A glycan biosynthesis.

Introduction

Many bacteria are flagellated, i.e. they have at least one flagellum. Rotation of the flagellar filament allows directed motility towards beneficial conditions (e.g. nutrient-rich) and away from noxious environments [173,174]. In addition, flagella mediate processes like adherence [132] and immunomodulation [133]. The flagellar filament is composed of repeating units of flagellin C (FliC) [175,176]. FliC *O*-glycosylation is essential for flagellar assembly and/or function in many species, e.g. *Helicobacter pylori* and *Campylobacter jejuni* [135,136]. Often, the glycan structures are unique and dependent on biosynthetic pathways with unusual enzyme activities [137,177].

In the major human gut pathogen *Clostridioides difficile*, FliC is also modified with glycan structures. In *C. difficile*, FliC glycosylation is pivotal for flagellar function because motility is seriously impaired in gene mutants with improper biosynthesis of the flagellar glycan [142,178]. So far, two different strain-dependent glycan structures have been described, Type A and Type B [142,179], which only have in common the core monosaccharide that is *O*-linked to multiple serine and threonine residues of FliC. The Type A glycan, which is found in the *C. difficile* strain 630Δ*erm*, consists of an *O*-linked *N*-acetylglucosamine (GlcNAc), that is linked to *N*-methyl-L-threonine through a phosphodiester bond (**Figure 1A**). This structure was fully characterized by a combination of mass spectrometry (MS) [141] and nuclear magnetic resonance spectroscopy (NMR) [142].

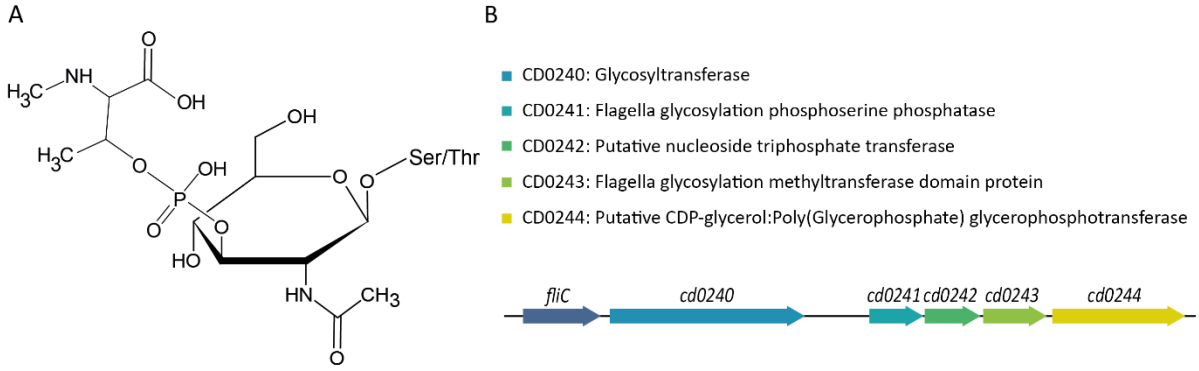


Figure 1. Reported structure of the Type A glycan modification and the gene cluster responsible for its biosynthesis. A) Structure of the FliC Type A glycan. The structure consists of an O-linked GlcNAc that is linked to N-methyl-L-threonine through a phosphodiester bond. **B)** The gene cluster responsible for the Type A glycan modification and the functions of the protein products as annotated in the UniProt *C. difficile* 630Δ*erm* reference proteome (Taxon ID: 272563).

In *C. difficile* 630Δ*erm*, a cluster of five genes (encoding CD0240-CD0244, **Figure 1B**) is linked to the biosynthesis of the Type A glycan [141,142]. This cluster is found downstream of the *fliC* gene (*cd0239*), as part of the larger flagellar gene cluster. CD0240 is a glycosyltransferase and disruption of this gene led to non-glycosylated FliC [141].

The role of the other genes within the cluster is less clear, but one study looked at alterations in the Type A glycan structure in mutants with insertions in individual genes using MS analyses of FliC glycopeptides from purified flagella [142]. In two of the mutants (*cd0241*::CT and *cd0242*::CT), flagellin was modified with only the core GlcNAc (i.e. lacking the *N*-methyl-phosphothreonine moiety). Whereas, in the *cd0243*::CT mutant strain, the Type A glycan structures lacked the *N*-methyl group on the threonine (only GlcNAc modifications were also observed), which was in line with the putative methyltransferase activity of CD0243 (**Figure 1B**). Surprisingly, no clear alterations in the Type A glycan structure were observed in the *cd0244*::CT strain (a mix of the full Type A glycan and GlcNAc on FliC was found), suggesting that CD0244 is redundant for Type A glycan biosynthesis. However, in the same study, it was observed that bacterial motility in the *cd0244*::CT strain was highly impaired. The reason for this apparent inconsistency has hitherto remained elusive. Nonetheless, a model for the biosynthesis of the Type A glycan structure in *C. difficile* was proposed [142], in which no role for CD0244 was defined.

Interestingly, in addition to *C. difficile* (a Gram-positive bacterium), a Type A-like glycan is also found in the Gram-negative bacterium *Pseudomonas aeruginosa,* for example in the reference strain PAO1. In this structure, the monosaccharide is a deoxyhexose which is linked to an unknown moiety through a phosphodiester bond [140]. The similarity between the structures is also apparent from the gene cluster observed in *P. aeruginosa* (**Supplemental Figure 1A**). However, this cluster consists only of four genes (*pa1088-pa1091*, homologs of *cd0240-cd0243*) and lacks a gene similar to *C. difficile cd0244* [140]. This supported the absence of an essential role for CD0244 in the model for the Type A glycan biosynthetic pathway in *C. difficile* as described above. However, bioinformatic analyses show that *pa1091* (*fgtA*) encodes a protein with both putative glycosyltransferase activity (similar to CD0240) as well as phosphotransferase activity (similar to CD0244). When mapping the predicted structures of CD0240 and CD0244 to the predicted structure of PA1091 (FgtA), the enzymatic domains of these proteins align with the predicted glycosyltransferase and phosphotransferase domains of PA1091, respectively (**Supplemental Figure S1B,C**). Hence, this also challenges the current model for the Type A glycan biosynthesis in *C. difficile* and led us to reinvestigate the alterations of the Type A glycan on FliC in the individual *C. difficile* mutant strains. In contrast to the previous study that used qualitative analysis of FliC glycopeptides from purified flagella, we used an overall quantitative mass spectrometry-based proteomics approach. Importantly, and in contrast to the previous data, we show that CD0244 is essential for full Type A glycan biosynthesis in *C. difficile.* Based on our data, we propose a revised model for the Type A glycan biosynthesis, providing testable hypotheses on the activity of individual enzymes encoded in the gene cluster.

Results

Relative abundance of CD0241-CD0244 in *C. difficile* WT, mutant, and complemented strains

To determine the relative abundance of the Type A biosynthetic proteins, we performed a mass spectrometry-based quantitative proteomics analysis of the *C. difficile* strains from the previous study as listed in **Table 1** [142] (i.e. wild-type (WT), *cd0241*::CT, *cd0242*::CT, *cd0243*::CT, *cd0244*::CT, *cd0241*::CT comp., *cd0242*::CT comp., *cd0244*::CT comp., in duplicate) using TMTpro 16plex labeling (no complemented strain for *cd0243*::CT was available).

Table 1. Overview of the *C. difficile* strains used in this study.

Description	Strain	Genotype	Plasmid	Reference
Wild-Type	WKS2044	*C. difficile* strain 630Δ*erm*	None	Ref [142]
cd0241::CT	WKS2047	*C. difficile* strain 630Δ*erm-cd0241*::CT	None	Ref [142]
cd0241::CT complemented	WKS2048	*C. difficile* strain 630Δ*erm-cd0241*::CT	pMTL84153 with *cd0241* cloned behind the *fdx* promoter	Ref [142]
cd0242::CT	WKS2049	*C. difficile* strain 630Δ*erm-cd0242*::CT	None	Ref [142]
cd0242::CT complemented	WKS2050	*C. difficile* strain 630Δ*erm-cd0242*::CT	pMTL84153 with *cd0242* cloned behind the *fdx* promoter	Ref [142]
cd0243::CT	WKS2051	*C. difficile* strain 630Δ*erm-cd0243*::CT	None	Ref [142]
cd0244::CT	WKS2052	*C. difficile* strain 630Δ*erm-cd0244*::CT	None	Ref [142]
cd0244::CT complemented	WKS2053	*C. difficile* strain 630Δ*erm-cd0244*::CT	pMTL84153 with *cd0244* cloned behind the *fdx* promoter	Ref [142]

Overall, 2187 *C. difficile* proteins with at least two peptides were identified (**Supplemental Table S1**). To our knowledge, this represents one of the most in-depth proteomics analyses of *C. difficile* cells. Given the aim of our study, we focused on the proteins encoded by the genes in the Type A glycan biosynthesis cluster (CD0240-CD0244) and all of them were readily identified with a high number of peptides. The data clearly showed that the levels of CD0241, CD0242 and CD0244 in the respective complemented strains were much higher than in the WT strain (**Supplemental Figure S2**), likely as a result of the plasmid-mediated expression under the control of a constitutive promotor from the *fdx* gene of *Clostridium pasteurianum* [180].

Unexpectedly, the relative protein abundance in the insertion mutants compared to the WT of CD0241 and especially CD0244 did not reflect a knockout phenotype of the

individual genes (**Supplemental Figure S2 and Supplemental Table S1**). To rule out any unexpected issues with the strains, we performed whole genome sequencing of all *C. difficile* 630Δ*erm* strains in **Table 1**, which confirmed that the strains were isogenic and that the ClosTron insertions were as reported previously [142] (data not shown). We argue that the seemingly high levels of CD0241 and especially CD0244 in their KO strains result from the unusually high expression of these proteins in their respective complemented strains, thereby compromising the correct quantification of these proteins using TMT labels (i.e. the levels of reporter ions are outside the dynamic range for accurate TMT quantification) [181]. This was supported by the data for CD0243 in the *cd0243*::CT strain, for which no complemented strain was available, and which reflected a knockout phenotype (**Supplemental Figure S2**).

To increase the accuracy of the quantification of the proteins involved in Type A biosynthesis, a second quantitative proteomics analysis was performed in which the complemented strains were excluded. The data from this experiment confirmed the knockout phenotype of the individual insertion mutants (**Figure 2**). The minor residual signals can be explained by either co-isolation or impurities in the TMTpro labels, i.e. each label contains a small percentage of different isotopologues (TMT Reporter Ion Isotope Distributions for TMTpro 16plex batch WK334339, Thermo Fisher Scientific). However, **Figure 2** also clearly shows an effect of the gene disruption by ClosTron mutagenesis [182] on the downstream genes. For example, a ClosTron insertion in *cd0242* influenced protein expression from the downstream *cd0243* and *cd0244* genes yet did not influence the upstream *cd0241* gene to a similar extent. Obviously, these downstream polar effects could not be restored by complementation (**Supplemental Figure S2**). It is unsurprising that the ClosTron insertion caused the polar effects, given the fact that *cd0241-cd0244* are part of a single operon in which transcription occurs from *cd0241* throughout the rest of the genes. Since *cd0244* is the last gene of the operon, this knockout did not show disruptive polar effects on the upstream genes in the operon (**Figure 2**).

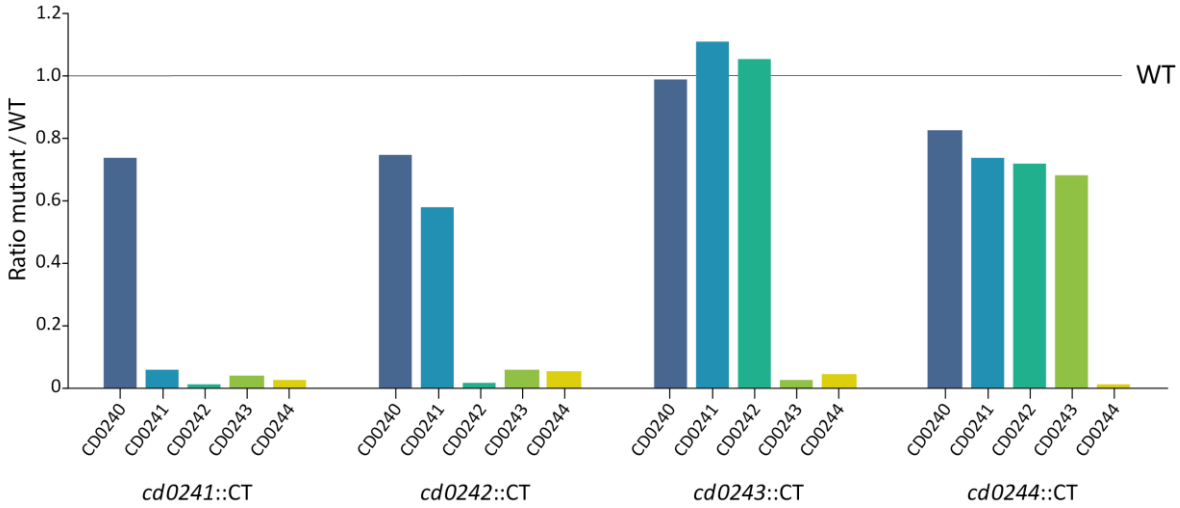


Figure 2. The relative levels of the Type A biosynthetic proteins in mutants with ClosTron insertions in the individual genes. A quantitative proteomics experiment was performed using TMTpro 15plex labeling (each strain in triplicate) and analyzed using LC-MS/MS on an Orbitrap Fusion Lumos Tribrid mass spectrometer. The protein levels of the Type A biosynthetic proteins in each of the individual gene mutant strains relative to the WT are shown. Ratios are calculated based on the average absolute abundance of a protein from three replicates per strain.

Alterations of the Type A glycan in the *cd0241-cd0244* mutant strains

To study the role of the individual genes in the *cd0241-cd0244* cluster on the Type A glycan biosynthesis, we explored our data from the TMTpro 16plex experiment, including all strains, for the presence of Type A glycan-modified tryptic peptides from *C. difficile* FliC (UniProt ID: Q18CX7). We focused on four different tryptic peptides of FliC that were modified with a Type A glycan structure. (LLDGTSSTIR, aa 135-144; AGGTTGTDAAK aa 191-201; TMVSSLDAALK, aa 202-212; LQVGASYGTNVSGTSNNNNEIK, aa 145-166). For each of these peptides, we concentrated on three scenarios, i.e. modification with the full Type A modification, a GlcNAc, or a Type A lacking the methyl group. MS/MS spectra corresponding to these peptides were observed in the proteomics data described above (**Supplemental Table S1**). However, to provide the best quantitative information, we performed additional targeted HCD MS/MS analyses of these peptides, which allowed us to sum the intensities of the TMT signals over the full peak, instead of using the TMT signals from a single MS/MS scan. In addition, this generated good-quality fragmentation spectra of our peptides of interest and their (altered) respective Type A structures.

The MS/MS spectrum of the Type A modified tryptic peptide LLDGTSSTIR is shown in **Figure 3A**. In this spectrum, Type A glycan-specific fragments at *m/z* 214.048 (*N*-methylthreonine-phosphate, $[M+H]^+$) and *m/z* 284.053 (phospho-GlcNAc, $[C_8H_{15}NO_8P]^+$)

are apparent. Moreover, the major peptide fragments have lost the Type A glycan modification. For the other three peptides containing a full Type A modification, similar fragmentation characteristics were observed (**Supplemental Figure S3-S5**).

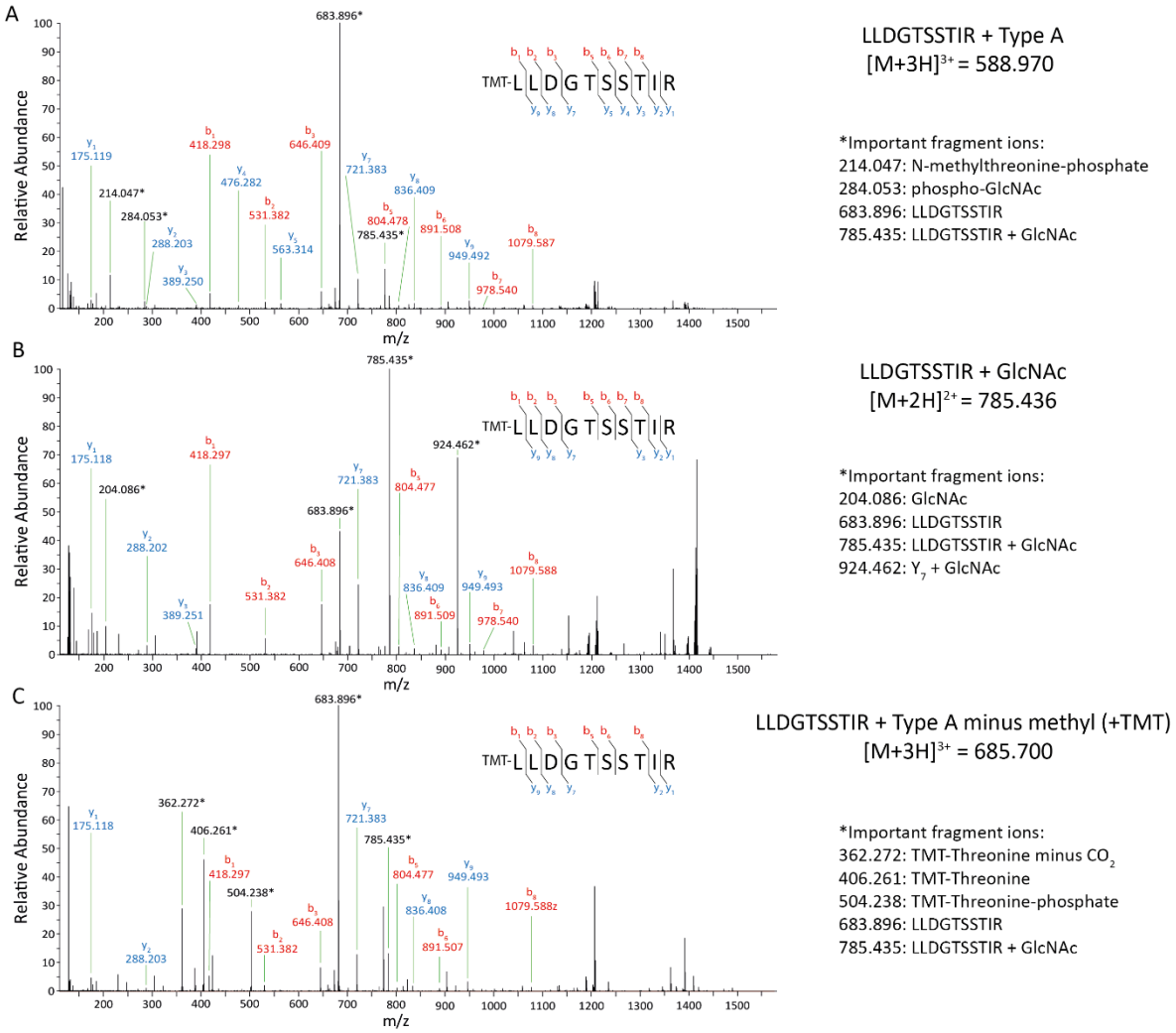


Figure 3. Summed MS/MS spectra of the LLDGTSSTIR peptide displaying the Type A glycan and variants thereof. Targeted HCD MS/MS analysis of the TMTpro 16plex labeled strains was performed. MS/MS spectra were summed over the full peak corresponding to the LLDGTSSTIR peptides displaying the complete Type A (**A**), only the GlcNAc (**B**) or Type A minus the methyl group having an extra TMT label (**C**). The theoretical precursor masses and the experimental masses of important fragment ions are shown on the right. All indicated b- and y-ions are from the unmodified TMT-labeled peptide.

The MS/MS spectrum of the tryptic peptide LLDGTSSTIR modified with a single GlcNAc is shown in **Figure 3B** and **Supplemental Figure S3-S5** for the other peptides. The MS/MS spectra of these species more clearly showed the GlcNAc oxonium ions, e.g. at *m/z* 204.087, as compared to Type A glycan-modified peptides (**Figure 3A and Supplemental Figure S3-S5**). The ratio of the oxonium ions at *m/z* 138.055 and 144.066 is consistent

with a GlcNAc [183,184]. Of note, a signal at *m/z* 126.055 was observed that corresponds to a GlcNAc oxonium ion, which is distinct from the 126C TMT reporter ion (*m/z* 126.128).

Interestingly, in the case of the absence of the *N*-methyl on the threonine as part of the Type A structure, our experimental setup would allow for TMT labeling of this extra amine group. Indeed, such FliC tryptic peptides containing the Type A glycan lacking the methyl group but with an additional TMT label were observed (**Figure 3C and Supplemental Figure S3-S5**). The fragmentation spectra were dominated by ions at *m/z* 504.239 (TMT-threonine-phosphate, $[M+H]^+$ and *m/z* 406.262 (TMT-threonine, $[M+H]^+$).

Next, we determined the relative abundance of the differently modified FliC peptides in each of the strains from **Table 1**. In **Figure 4**, the TMT signals from the MS/MS spectra of these modified FliC tryptic peptides are depicted. In line with what was shown previously [142], the Type A glycan-modified FliC peptides were absent in the *cd0241*::CT, *cd0242*::CT and *cd0243*::CT strains. However, in contrast to what was shown previously, Type A glycan-modified peptides were also absent in the *cd0244*::CT strain (**Figure 4**). As described above, the minor TMT signals that were observed for *cd0241*::CT and *cd0244*::CT in **Figure 4** can be explained by impurities in the TMT labels. As expected, in addition to the WT strain, Type A glycan-modified peptides were also detected in the complemented strains, although the level of complementation varied per strain and peptide.

In line with previous data [142], FliC tryptic peptides with a single GlcNAc were observed in the *cd0241*::CT and *cd0242*::CT strains (**Figure 4**). Importantly, we clearly show that also in the *cd0244*::CT strain, FliC tryptic peptides with a single GlcNAc are highly abundant, again demonstrating that the modification of FliC in this strain is radically different from the WT strain. FliC peptides with only GlcNAc moieties were also detected in the *cd0243*::CT strain, which is in line with what has been observed before (3). We propose that this is due to the polar effects on *cd0244* expression in the *cd0243*::CT strain (**Figure 2**).

As expected, peptides containing the Type A modification lacking the methyl group but with an additional TMT label were predominantly observed in the *cd0243*::CT strain (**Figure 4**). However, TMT reporter ions that could not be explained by impurities in the TMT labels were also observed in the *cd0241*::CT comp. and *cd0242*::CT comp. strains. We find it likely that this is due to a decreased efficiency in Type A biosynthesis due to the polar effects of the ClosTron insertion on *cd0243* in the *cd0241*::CT and *cd0242*::CT strains (**Figure 2**).

Overall, our new data not only confirms the importance of CD0241-CD0243 for full Type A glycan biosynthesis in *C. difficile* but also demonstrates that CD0244 is pivotal for full

Type A glycan biosynthesis. In the *cd0244*::CT strain, loss of the Type A glycan structure coincides with the appearance of peptides displaying only a GlcNAc.

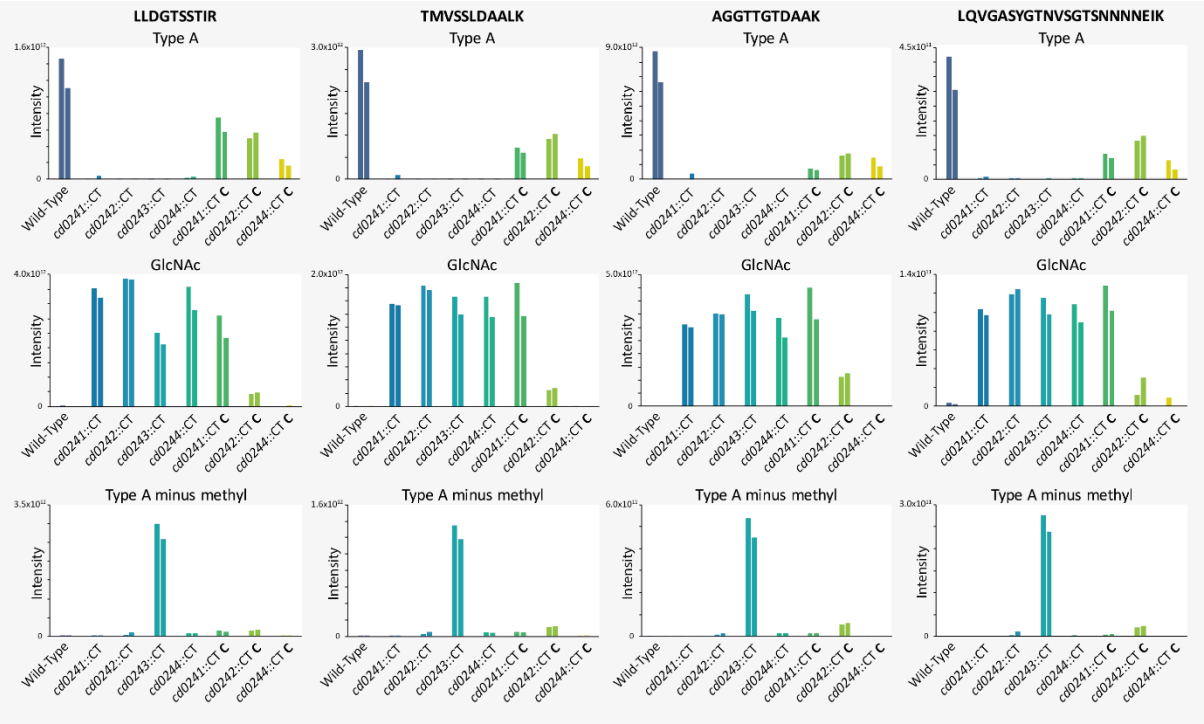


Figure 4. Relative levels of peptides containing the Type A variants in individual gene mutants and complemented strains. Targeted HCD MS/MS analysis of the TMTpro 16plex labeled strains was performed. MS/MS spectra were summed over the full peak corresponding to the peptides displaying the complete Type A, only the GlcNAc or Type A minus the methyl group having an extra TMT label. The bars represent the absolute intensities of the TMT reporter labels for each strain, analyzed in duplicate. The complemented strains are indicated with a "C".

New model for the Type A glycan biosynthetic pathway in *C. difficile*

Our results are not compatible with the current model for the Type A glycan biosynthesis, which did not include a role for CD0244. In the previous model, it was proposed that CD0241 catalyzes the addition of phosphate to threonine, followed by CD0242 mediating the transfer of the phosphothreonine to the GlcNAc [142]. Finally, CD0243 catalyzes the *N*-methylation of the threonine, although it is unclear during which step this occurs. In addition to the lack of a role for CD0244, the previous model also did not predict how the phosphothreonine is activated as a biosynthetic intermediate that can act as a donor. Hence, the above prompted us to formulate new hypotheses about the activities of the different enzymes in this important biosynthetic pathway.

Bio-informatic analyses show that CD0242 belongs to the family of nucleotidyl transferases, which transfer a nucleoside monophosphate moiety to an accepting

molecule. For example, a Phyre2 homology search models 97% of the sequence with 99.8% confidence to GDP-mannose pyrophosphorylase (a nucleotidyl transferase) from *Leishmania donovani* (PDB: 7whs, 21% i.d.). Indeed, the *C. difficile* reference genome (strain 630) from UniProt (Taxon ID: 272563) annotates CD0242 as a nucleoside triphosphate transferase (**Figure 1**, ID: Q18CY2). One of the proteins that is similar to CD0242, and was mentioned in the previous study [142], is CTP:phosphocholine cytidylyltransferase. This cytidylyltransferase is a key enzyme in the synthesis of phosphatidylcholine referred to as the Kennedy pathway [185]. Based on this amino acid similarity, we hypothesize that CD0242 is a CTP:phosphothreonine cytidylyltransferase that transfers CMP to phosphothreonine. The end product of the reaction is expected to be CDP-L-threonine.

CD0244, for which no role has previously been predicted, shows similarity to the CDP-glycerol:Poly(glycerophosphate) glycerophosphotransferase TagF from *Staphylococcus epidermidis* (Phyre2 models 77% of the sequence with 100% confidence, PDB: 3l7m, 16% i.d.), which has similar enzymatic activity as the phosphotransferase in the Kennedy pathway [185]. In line with this and our new data for the *cd0244*::CT strain, we hypothesize that CD0244 is a CDP-threonine:GlcNAc threoninephosphotransferase that transfers the phosphothreonine moiety from CDP-L-threonine to the core GlcNAc on FliC.

The most challenging prediction is the role of *C. difficile* CD0241. In the previous model, it was thought to be involved in the synthesis of phosphothreonine. However, bioinformatic analyses showed the homology of CD0241 with a phosphoserine phosphatase (PSP), not a kinase. A Phyre2 homology search models 96% of the sequence with 100% confidence to the PSP from *Methanocaldococcus jannashii* (PDB: 1j97, 29% i.d.). Also, the counterpart of CD0241 in *P. aeruginosa*, PA1089, is predicted to exhibit a similar activity. However, in that same organism, a different PSP-like enzyme is present that not only shows phosphatase activity but also phosphotransferase activity [186]. This enzyme, ThrH, is a phosphoserine:homoserine phosphotransferase. Interestingly, both CD0241 and PA1089 are ThrH homologs and are predicted to adopt a similar fold to that of ThrH (**Supplemental Figure S6**), while many other similar PSP-like proteins display more differences in size and or fold. In addition, homoserine is an isomer of threonine, indicating that there are only minor differences in substrates. Based on the above, we hypothesize that CD0241 (and PA1089) possesses a phosphoserine:threonine phosphotransferase activity.

Based on our proteomics data and bioinformatic analyses, we propose a revised model for the Type A glycan biosynthesis on FliC as shown in **Figure 5**. Here, CD0241 transfers the phosphate group from a phosphoserine to a threonine, forming phosphothreonine. Next, CD0242 transfers the phosphothreonine to CTP, thereby forming CDP-threonine,

while releasing an inorganic pyrophosphate (PPi). Then, CD0244 transfers the phosphothreonine to the GlcNAc moiety on FliC, which is attached by the glycosyltransferase CD0240, and this causes the release of CMP. At an unknown point during these steps, CD0243 mediates the *N*-methylation of the threonine to form the complete Type A glycan modification. A likely donor is S-adenosyl methionine, that is converted to S-adenosyl homocysteine when donating its methyl group.

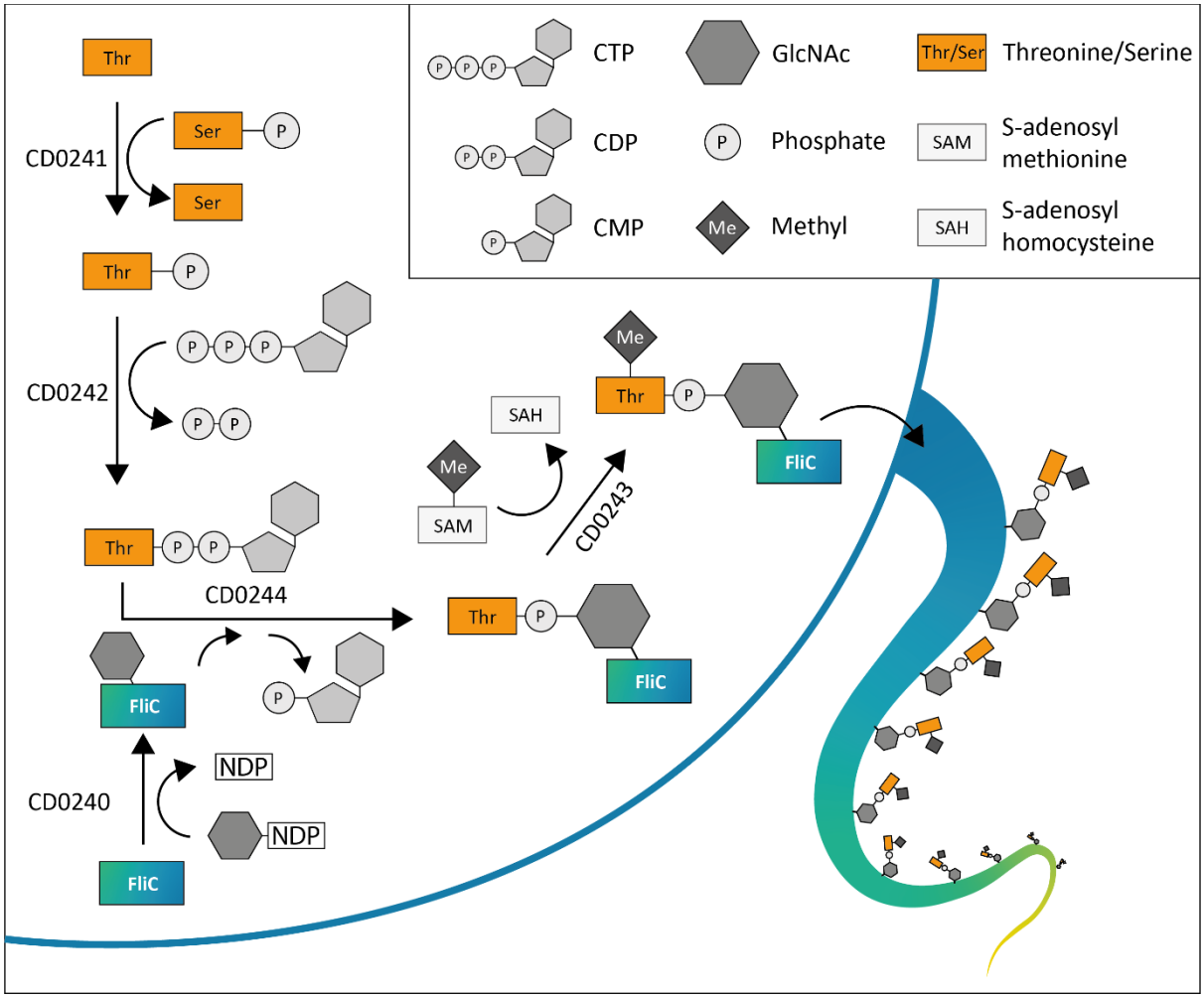


Figure 5. Schematic overview of the revised model for the Type A glycan biosynthetic pathway. First, CD0241 transfers the phosphate group from a phosphoserine to a threonine, forming phosphothreonine. Next, CD0242 transfers the phosphothreonine to CTP, thereby forming CDP-threonine, while releasing inorganic pyrophosphate. Then, CD0244 transfers the phosphothreonine to the GlcNAc moiety on FliC, which is attached by the glycosyltransferase CD0240, and this causes the release of CMP. The GlcNAc moiety on FliC is most likely donated by a nucleoside-diphosphate-GlcNAc (NDP-GlcNAc). At an unknown point during these steps, CD0243 mediates the N-methylation of the threonine to form the complete Type A glycan modification. A likely donor is S-adenosyl methionine, that is converted to S-adenosyl homocysteine when donating its methyl group.

Discussion

Flagella and their role in motility, adherence and other host-pathogen interactions vary across different *C. difficile* lineages. For the *C. difficile* 630Δ*erm* strain, flagella are not essential for colonization of the host by the bacteria [187–189]. However, flagellated strains display an increased fitness *in vivo* [189], greater caecal adherence [189] and induce a more intense inflammation [190] than their non-flagellated counterparts. On the other hand, strains that are impaired in FliC production, the major component of the flagellar filament, have been shown to produce more exotoxins and are more virulent [188,189]. Previous studies have shown that post-translational modification of FliC is important for flagellar function [135,136]. Also in *C. difficile* 630Δ*erm*, disruption of several genes involved in the biosynthesis of the Type A glycan structure that is present on FliC, i.e. *cd0241*, *cd0242* and *cd0244*, resulted in impaired mobility [142]. Moreover, a strain that was only able to modify FliC with the core GlcNAc moiety of the Type A glycan (*cd0241*::CT) showed attenuated initial colonization and recurrence in mice [142]. However, the proposed model for the biosynthesis of the Type A glycan defined no role for CD0244 and lacked detailed prediction on enzymatic activities and biosynthetic intermediates [142].

Our results demonstrate a clear role for CD0244 in the biosynthesis of the Type A glycan. In the *cd0244*::CT strain, the loss of Type A coincided with the appearance of the core GlcNAc of the Type A structure. Previously, a mixed population of both structures was observed [142]. We currently have no explanation for the discrepancy between our results and the previously reported data, especially since we used the same set of strains. However, the quantitative nature of the current study, as compared to the earlier qualitative analyses, may partially explain this. Nevertheless, the current data would explain the apparent inconsistency that was found between the impaired motility that was observed in the *cd0244*::CT strain as opposed to the absence of clear alterations in the Type A glycan structure in the earlier study.

Based on our bioinformatic analyses, we hypothesize that CD0242 mediates the synthesis of CDP-L-threonine, which would be a key biosynthetic intermediate of the Type A biosynthesis. To our knowledge, CDP-L-threonine would be a so far not described cellular metabolite. However, several studies have shown the existence of amino acid residues linked to CDP in other prokaryotes, namely CDP-L-glutamine and CDP-L-serine [191]. Furthermore, we predict CD0244 to be a CDP-threonine:GlcNAc threoninephosphotransferase. However, CD0244 also shows similarity to UDP-N-acetylglucosamine 2-epimerase. This enzyme catalyzes the reversible epimerization of UDP-GlcNAc into UDP-ManNAc, the activated donor of ManNAc. Yet, this function is not supported by the data. First of all, the Type A modification has been shown to contain a

GlcNAc and not a ManNAc [142]. Second, the lack of CD0244 in the *cd0244*::CT strain does not prevent the glycosylation of FliC.

Recently, we showed that a phosphoproteomics workflow could be used to enrich Type A modified peptides [192], probably due to the phospho moiety of the Type A glycan. For the current study, such an approach was not suitable because we would lose the GlcNAc-modified peptides. In our previous phosphoproteomics data, we also observed a fraction of FliC tryptic peptides that were modified with a phospho-GlcNAc [193]. However, we have not observed the accumulation of such peptides in any of our knockout strains. Therefore, we find it likely that these peptides were the results of breakdown processes. This is supported by the fact that phospho-GlcNAc peptides could be identified in our database searches, but they all co-eluted with the full Type A-modified peptides, indicating in-source decay of the Type A peptides (data not shown). Moreover, species only originated from the WT and the complemented strains that produce the full Type A glycan, further supporting the idea that the phospho-GlcNAc moiety is not an intermediate in the biosynthesis of the Type A glycan.

Disruption of any of the genes in the *cd0241-cd0244* cluster using the ClosTron method completely prevents the formation of the Type A glycan. Although the levels of CD0241, CD0242, and CD0244 are unnaturally high in their respective complemented strains, this overexpression did not restore the levels of Type A containing peptides to the WT level. For the *cd0241*::CT complemented and *cd0242*::CT complemented strains, we argue that this is due to the polar effects on the downstream genes caused by the gene insertions, which appeared to be quite strong. Nonetheless, the fact that partial complementation was possible shows that, despite the strong polar effects, active enzymes from the affected genes are still present. For the *cd0244*::CT strain, no disruptive polar effects on the upstream genes in the cluster were observed, which was also apparent from the lack of peptides with a single GlcNAc in the *cd0244*::CT complemented strain. Still, we found lower levels of Type A modified peptides in the *cd0244*::CT complemented strain, as compared to the WT strain. This, however, might be explained by low levels of FliC itself in the *cd0244*::CT complemented strain (**Supplemental table S1**). FliC levels in the *cd0244*::CT complemented strain appeared to be around six times lower compared to the WT, and if we corrected for these differences in FliC levels, the levels of Type A modified peptides in the *cd0244*::CT complemented strain would approach the WT levels. The nature of the lower levels of FliC in this strain remains unclear. Possibly, a feedback loop is present that responds to the overexpression of *cd0244* in the complemented strain.

Conclusion

In conclusion, based on quantitative proteomics and bio-informatic analyses, we propose a revised model for the biosynthesis of the Type A glycan modification on FliC in *C. difficile* and predict enzymatic activities for each of the involved proteins. Further experiments using these enzymes should shed more light on their activities. Our findings and model for post-translational glycan modification of flagellin in *C. difficile* will be relevant to the similar locus in *P. aeruginosa* PA01 and other bacterial species with similar flagellin modifications.

Experimental procedures

Bacterial strains and growth conditions

The *C. difficile* strains used in this study are listed in **Table 1** [142] and were cultured at 37 °C in a Don Whitley A55 HEPA anaerobic workstation. The cells were grown in brain heart infusion (BHI, Oxoid) broth supplemented with 5 g/liter yeast extract (BHIY) or on BHIY agar plates. When appropriate, 15 µg/ml thiamphenicol was added.

Sample preparation for the quantitative proteomics analysis of *C. difficile* strains

Single colonies of *C. difficile* were picked and were precultured for 24 h in 5 ml prereduced BHIY. Next, the precultures were used to inoculate 5 ml of prereduced BHIY broth at a starting OD_{600} of 0.05 and cells were grown for 16 h. Then, cells were pelleted by centrifugation (3220 x g, 20 min, 4 °C). Pellets were resuspended in 1 ml ice-cold PBS and washed three times (8000 x g, 5 min, 4 °C). After the last wash, pellets were resuspended in 1 ml ST lysis buffer (5% SDS, 0.1 M Tris-HCL pH 7.5). Tubes were incubated for 20 min on ice prior to lysis by sonication and cells were subsequently lysed by sonication for five bursts of 10 s with cooling on ice in between rounds. After lysis, tubes were centrifuged (15 min, 15000 x g, RT). Supernatants were transferred to new tubes and stored at -20 °C until further use.

For each strain, 100 µg of protein in 100 µl ST buffer was used as the starting material. Proteins were reduced using 5 mM TCEP for 30 min, alkylated with 10 mM iodoacetamide for 30 min, and quenched with 10 mM DTT for 15 min, all at room temperature. Proteins were precipitated by chloroform-methanol precipitation. For this, 400 µl methanol, 100 µl chloroform, and 300 µl dH_2O were added with vortexing in between each step. Following centrifugation (21130 x g, 2 min, RT), the pellet was washed two times with 500 µl methanol. The protein pellet was subsequently

resuspended in 100 µl 40 mM HEPES pH 8.4 containing 4 µg trypsin and incubated overnight at 37 °C. Again, 4 µg trypsin was added and incubated for 3 h.

TMT labeling was performed on 10 µg tryptic peptides using TMTpro 16plex labeling (Thermo Fisher Scientific, lot no. WK334339) for 1 h at RT. Excess TMT label was quenched with 5% hydroxylamine for 15 min at RT. The labeled peptides from each sample were mixed and freeze-dried. The peptides were resuspended in 10 mM ammonium bicarbonate pH 8.4 and separated in 12 fractions on an Agilent Eclipse Plus C18 column (2.1 x 150 mm, 3.5 µM). Half of the labeled peptides (80 µg) were injected. Mobile phase A: 10 mM ammonium bicarbonate (pH 8.4). Mobile phase B: 10 mM ammonium bicarbonate in 80% acetonitrile (pH 8.4). The gradient was as follows: 2% B, 0-5 min; 2%-90% B, 5-35 min; 90% B, 35-40 min; 90%-2% B, 40-41 min; 2% B, 41-65 min. The 12 collection vials were rotated every 30 s during sample collection. The 12 fractions were freeze-dried and stored at -20 °C prior to LC-MS/MS analysis. Two TMT experiments were performed: one with 16 samples (16plex), one with 15 samples (15plex). The overview of the TMTpro labels for each strain is shown in **Supplemental Table S2**.

LC-MS/MS analysis

LC-MS/MS analyses were performed as previously described with minor adjustments [194]. TMT-labeled peptides were dissolved in 0.1% formic acid and subsequently analyzed by online C18 nano-HPLC MS/MS with a system consisting of an Easy nLC 1200 gradient HPLC system (Thermo, Bremen, Germany), and an Orbitrap Fusion LUMOS mass spectrometer (Thermo). Fractions were injected onto a homemade precolumn (100 µm × 15 mm; Reprosil-Pur C18-AQ 3 µm, Dr Maisch, Ammerbuch, Germany) and eluted via a homemade analytical nano-HPLC column (30 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm). The analytical column temperature was maintained at 50 °C with a PRSO-V2 column oven (Sonation, Biberach, Germany). The gradient was run from 2% to 40% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 120 min. The nano-HPLC column was drawn to a tip of ~5 µm and acted as the electrospray needle of the MS source. The LUMOS mass spectrometer (Thermo) was set to use the MultiNotch MS3-based TMT method [195]. The MS spectrum was recorded in the Orbitrap (resolution 120,000; m/z range 400–1500; automatic gain control (AGC) target was set to 50%; maximum injection time 50 ms). Dynamic exclusion was after n = 1 with an exclusion duration of 45 s with a mass tolerance of 10 ppm. Charge states 2–5 were included. Precursors for MS2/MS3 analysis were selected using “TopSpeed” with a cycle time of 3 sec. MS2 analysis consisted of collision-induced dissociation (quadrupole ion trap analysis; AGC was set to “standard”; normalized collision energy (NCE) 35; maximum injection time 50 ms). The isolation window for MS/MS was 1.2 Da. Following the

acquisition of each MS2 spectrum, the MultiNotch MS3 spectrum was recorded using an isolation window for MS3 of 2 Da. Ten MS2 fragments were simultaneously selected for MS3 and fragmented by high energy collision-induced dissociation (HCD) at 65% at a custom AGC of 200% and analyzed using the Orbitrap from *m/z* 120 to 500 at a maximum injection time of 105 ms at a resolution of 60,000).

To obtain more accurate ratios for selected species a separate targeted MS2 (tMS2) run was recorded for the following peptides and their selected *m/z*: LLDGTSSTIR with: Type A, 588.97 ($[M+3H]^{3+}$); GlcNAc, 785.44 ($[M+2H]^{2+}$); Type A minus methyl, 685.70 ($[M+3H]^{3+}$); AGGTTGTDAAK with: Type A, 652.67 ($[M+3H]^{3+}$); GlcNAc, 587.66 ($[M+3H]^{3+}$); Type A minus methyl, 749.40 ($[M+3H]^{3+}$); TMVSSLDAALK with: Type A, 714.71 ($[M+3H]^{3+}$); GlcNAc, 649.70 ($[M+3H]^{3+}$); Type A minus methyl, 811.44 ($[M+3H]^{3+}$); LQVGASYGTNVSGTSNNNNEIK with: Type A, 819.16 ($[M+4H]^{4+}$); GlcNAc, 770.40 ($[M+4H]^{4+}$); Type A minus methyl, 891.71 ($[M+4H]^{4+}$). tMS2 spectra were recorded with a precursor isolation width of 0.7 Da, at an HCD collision energy of 36% at resolution 30,000 and an AGC target “standard”. The maximum injection time was set to 54 ms. MS2 spectra of each selected species were summed.

LC-MS/MS data analysis

In a post-analysis process, raw data were converted to peak lists using Proteome Discoverer version 2.5.0.400 (Thermo Electron), and submitted to the UniProt *C. difficile* 630Δ*erm* database (3752 entries) (Taxon ID: 272563) using Mascot v. 2.2.07 (www.matrixscience.com) for peptide identification. Mascot searches were performed with 10 ppm and 0.5 Da deviation for precursor and fragment mass, respectively, and trypsin was selected as enzyme specificity with a maximum of 2 missed cleavages. The variable modifications included Type A (ST), Type A minus methyl plus TMT (ST), HexNAc (ST), Oxidation (M), and Acetyl (protein N-term). For the TMTpro 15plex search, also phosphoHexNAc was included. The static modifications included TMTpro (N-term, K) and Carbamidomethyl (C). Peptides with an FDR < 1% based on Percolator [196] were accepted. Quantification of peptides was performed on MS3 spectra with an SPS Mass Matches threshold of 100%.

Whole genome sequencing

For identity confirmation, mutant strains were subjected to whole genome sequencing, according to standard procedures [197]. In short, total genomic DNA was isolated from a single colony resuspended in PBS on a QiaSymphony platform (Qiagen). Purified DNA was sequenced on the Illumina Novaseq6000 platform with a read length of 150 bp in

paired-end mode. The resultant FASTQ files were used in a reference assembly against the *C. difficile* 630 reference genome (GenBank AM180355) in Geneious software (Biomatters Ltd); Clostron insertions were confirmed by visual identification of clusters of nucleotide polymorphisms, and computational identification of high-quality single nucleotide polymorphisms using the Find Variations/SNPs algorithm in Geneious (minimum coverage 10, minimum variant frequency 0.8).

Bioinformatic analyses

To search for protein homologs and predict functions, the Phyre2 web portal [198] and the InterPro website for classification of protein families (www.ebi.ac.uk/interpro/search/sequence) were used. Predicted protein structures were retrieved from the Alphafold database (alphafold.ebi.ac.uk/) or modeled using Alphafold2 [199]. Analyses of protein structures were performed in PyMOL 2.5.5.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [200] via the PRIDE [201] partner repository with the dataset identifier PXD045152.

Acknowledgments

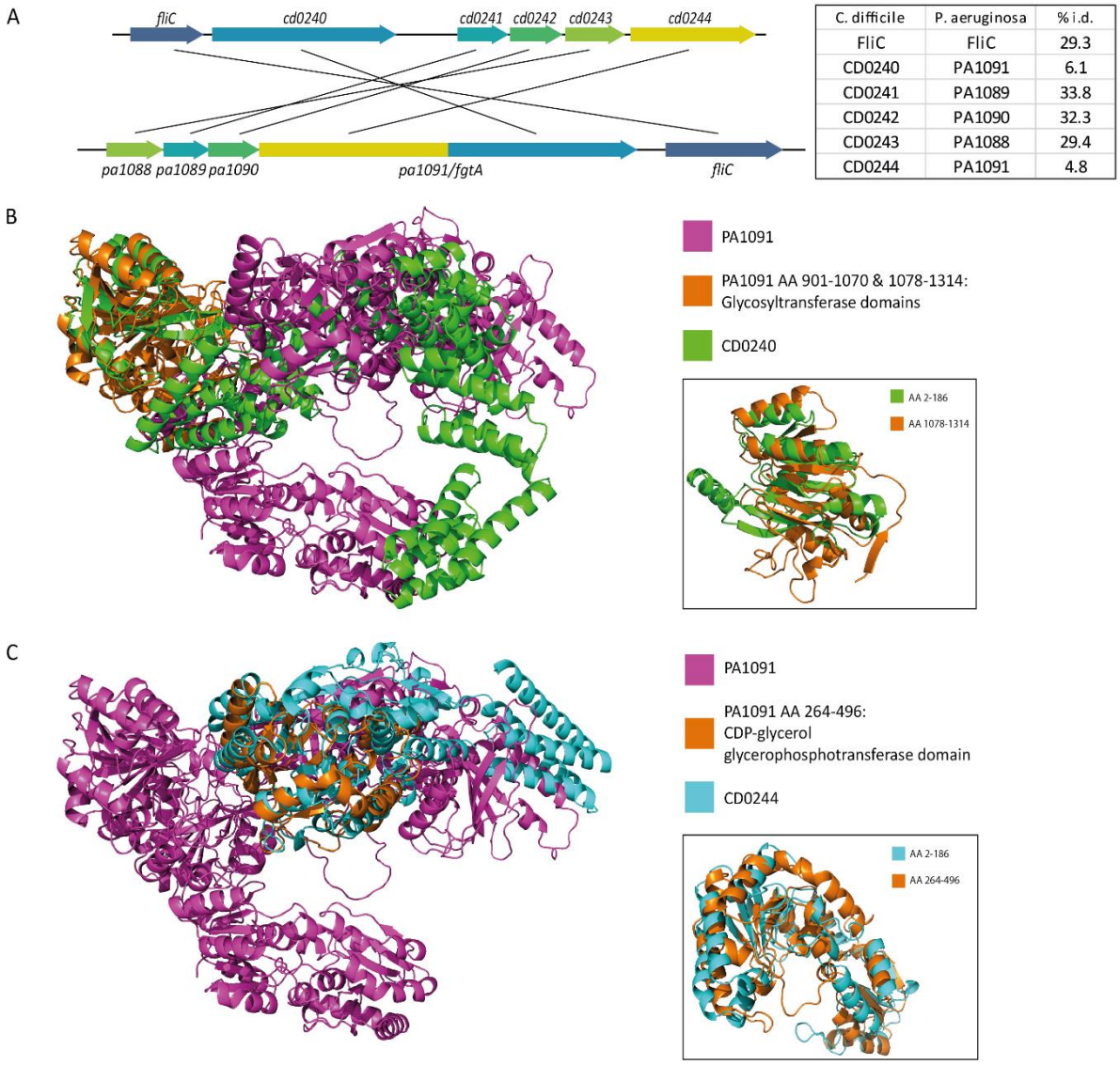
This research was supported by an ENW-M grant (OCENW.KLEIN.103) from the Dutch Research Council (NWO).

Supporting information

Supplemental Table S1: Results from database searches of LC-MS/MS data (XLSX) can be found online

https://pubs.acs.org/doi/full/10.1021/acsinfecdis.3c00485

Supporting information



C. difficile	P. aeruginosa	% i.d.
FliC	FliC	29.3
CD0240	PA1091	6.1
CD0241	PA1089	33.8
CD0242	PA1090	32.3
CD0243	PA1088	29.4
CD0244	PA1091	4.8

Figure S1. PA1091 from *P. aeruginosa* contains multiple domains that are similar to both CD0240 and CD0244 from *C. difficile* 630Δ*erm*. A) A schematic representation of the FliC post-translational modification gene clusters in both *C. difficile* and *P. aeruginosa*. The lines connect the genes which products have similar predicted functions. The percentage of identities of the homologous proteins are shown in the table on the right. **B)** CD0240 superimposed on PA1091. The predicted glycosyltransferase domain of CD0240 (AA 2-186) maps to the predicted glycosyltransferase domains of PA1091 (AA 901-1070 and 1078-1314, in orange). Inset: The predicted glycosyltransferase domain of CD0240 and one of the predicted domains of PA1091. **C)** CD0244 superimposed on PA1091. The predicted CDP-glycerol glycerophosphotransferase domain of CD0244 (AA 244-486) maps to the predicted CDP-glycerol glycerophosphotransferase domain of PA1091 (AA 264-496, in orange). Inset: The predicted CDP-glycerol glycerophosphotransferase domains of PA1091 and CD0244.

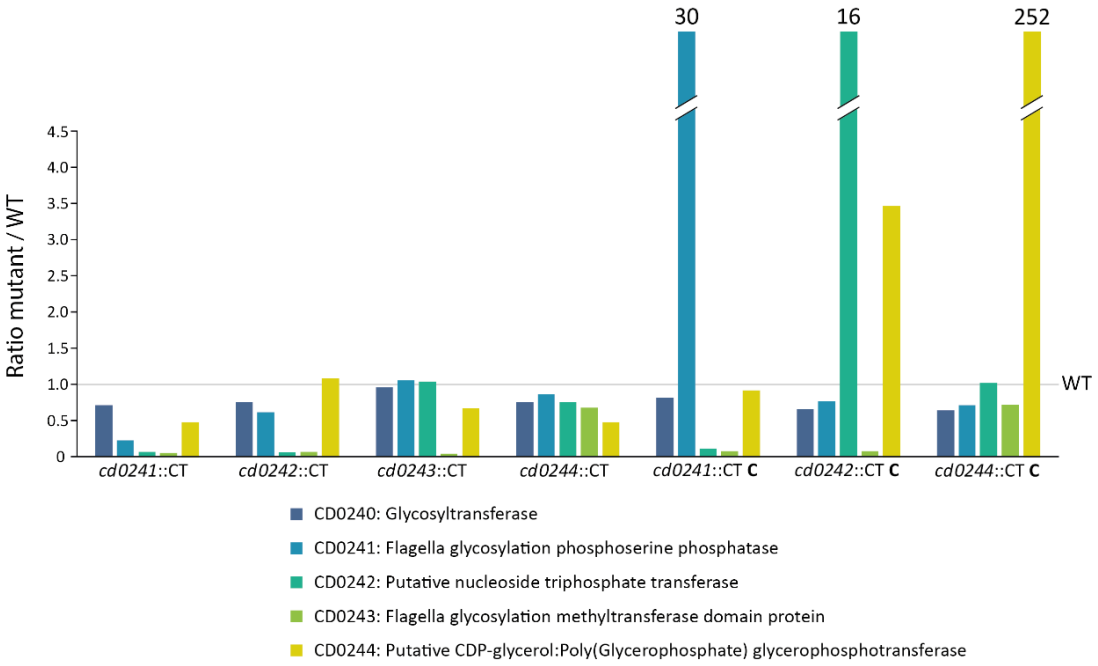


Figure S2. The relative levels of the Type A biosynthetic proteins in mutants with ClosTron insertions in the individual genes and their complemented strains. A quantitative proteomics experiment was performed using TMTpro 16plex labeling (each strain in duplicate). The protein levels of the Type A biosynthetic proteins in each of the individual strains relative to the WT are shown. For CD0241, CD0242, and CD0244 in their respective complemented strains, the ratios are depicted above the bars. Ratios are calculated based on the average absolute abundance of a protein from two replicates per strain.

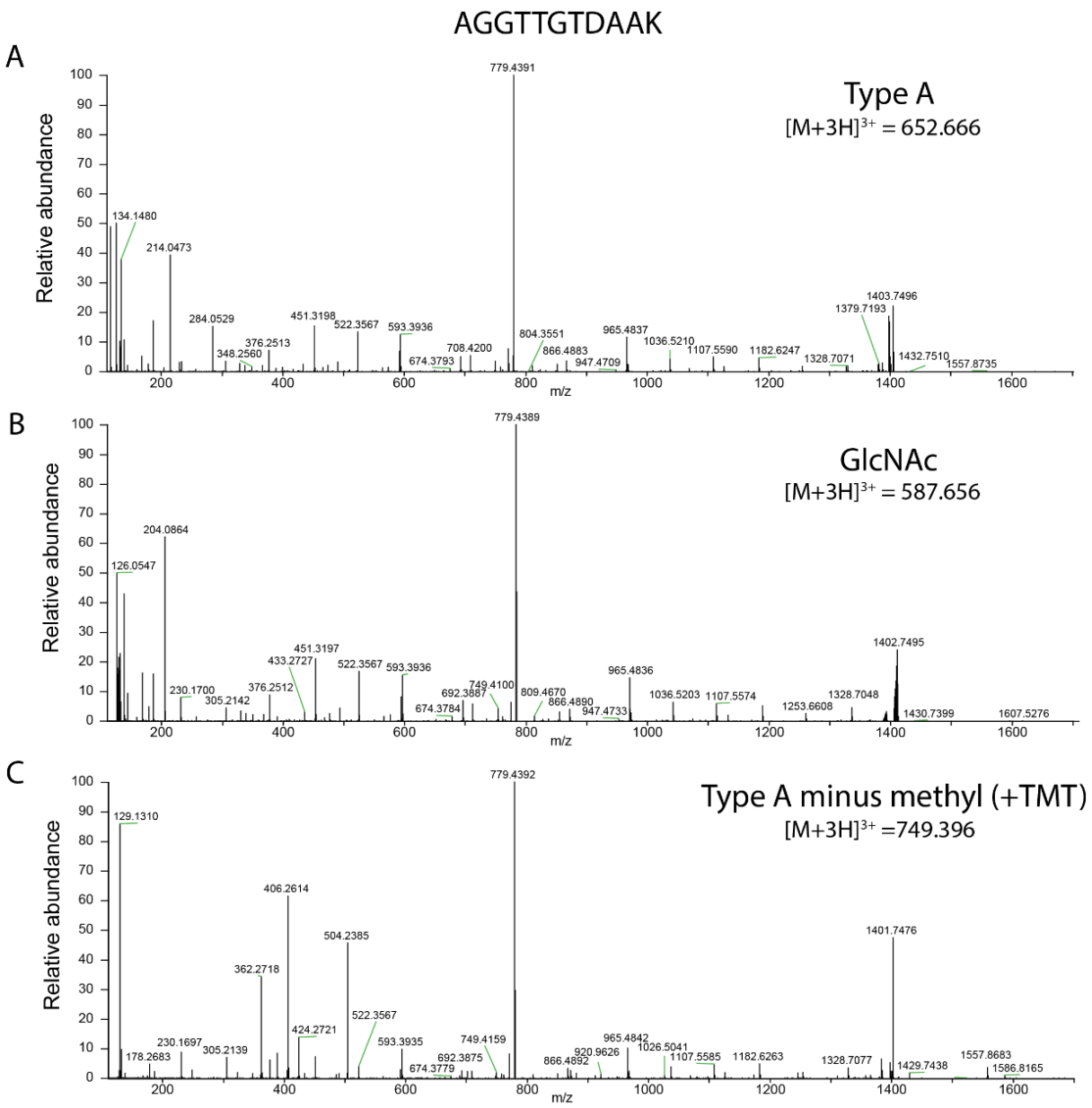


Figure S3. Summed MS/MS spectra of the AGGTTGTDAAK peptide displaying the Type A variants. Targeted HCD MS/MS analysis of the TMTpro 16plex labeled strains was performed. MS/MS spectra were summed over the full peak corresponding to the AGGTTGTDAAK peptides displaying the complete Type A (**A**), only the GlcNAc (**B**) or Type A minus the methyl group but having an extra TMT label (**C**). The theoretical precursor masses are shown on the right.

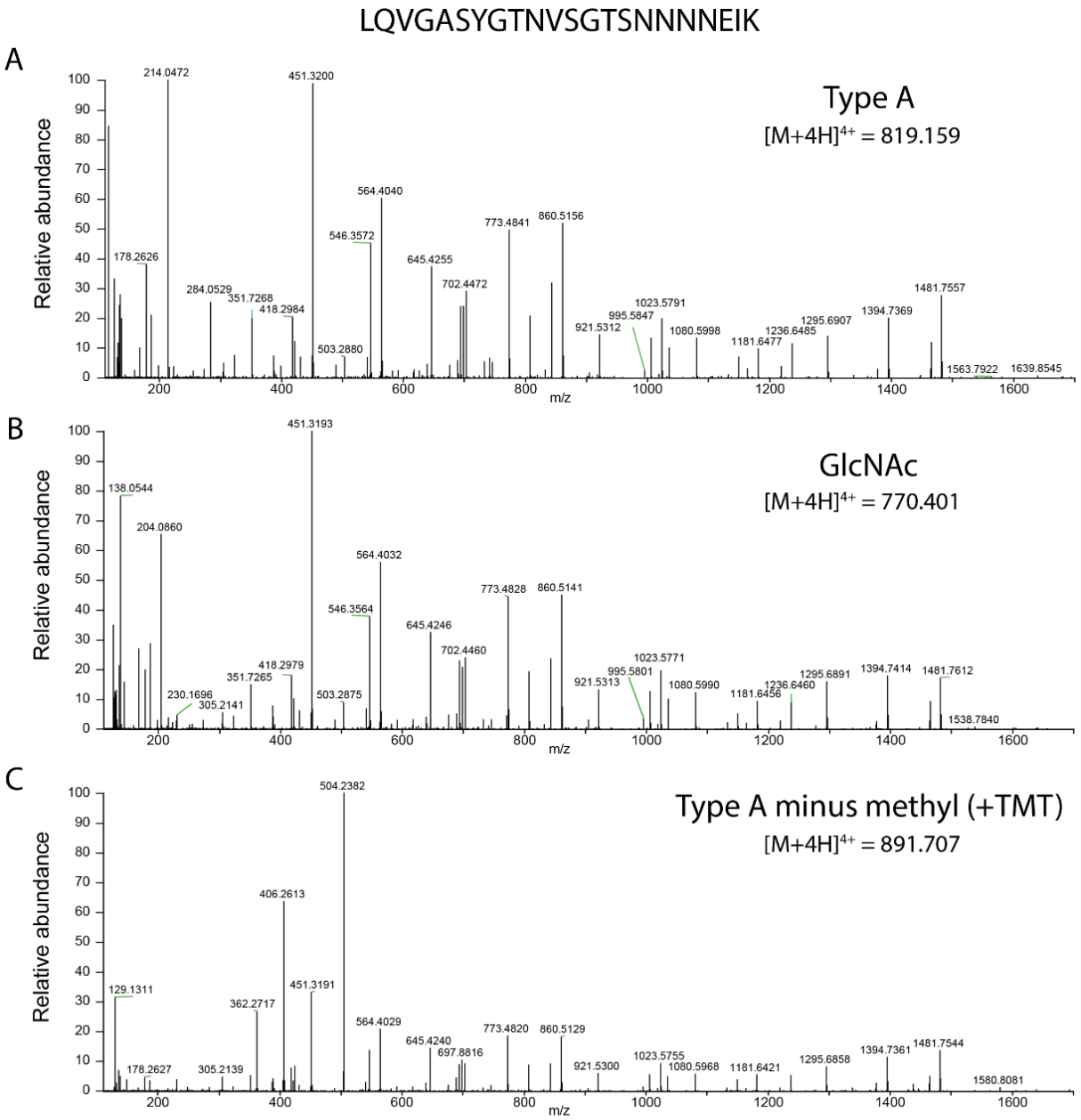


Figure S4. Summed MS/MS spectra of the LQVGASYGTNVSGTSNNNNEIK peptide displaying the Type A variants. Targeted HCD MS/MS analysis of the TMTpro 16plex labeled strains was performed. MS/MS spectra were summed over the full peak corresponding to the LQVGASYGTNVSGTSNNNNEIK peptides displaying the complete Type A (**A**), only the GlcNAc (**B**) or Type A minus the methyl group but having an extra TMT label (**C**). The theoretical precursor masses are shown on the right.

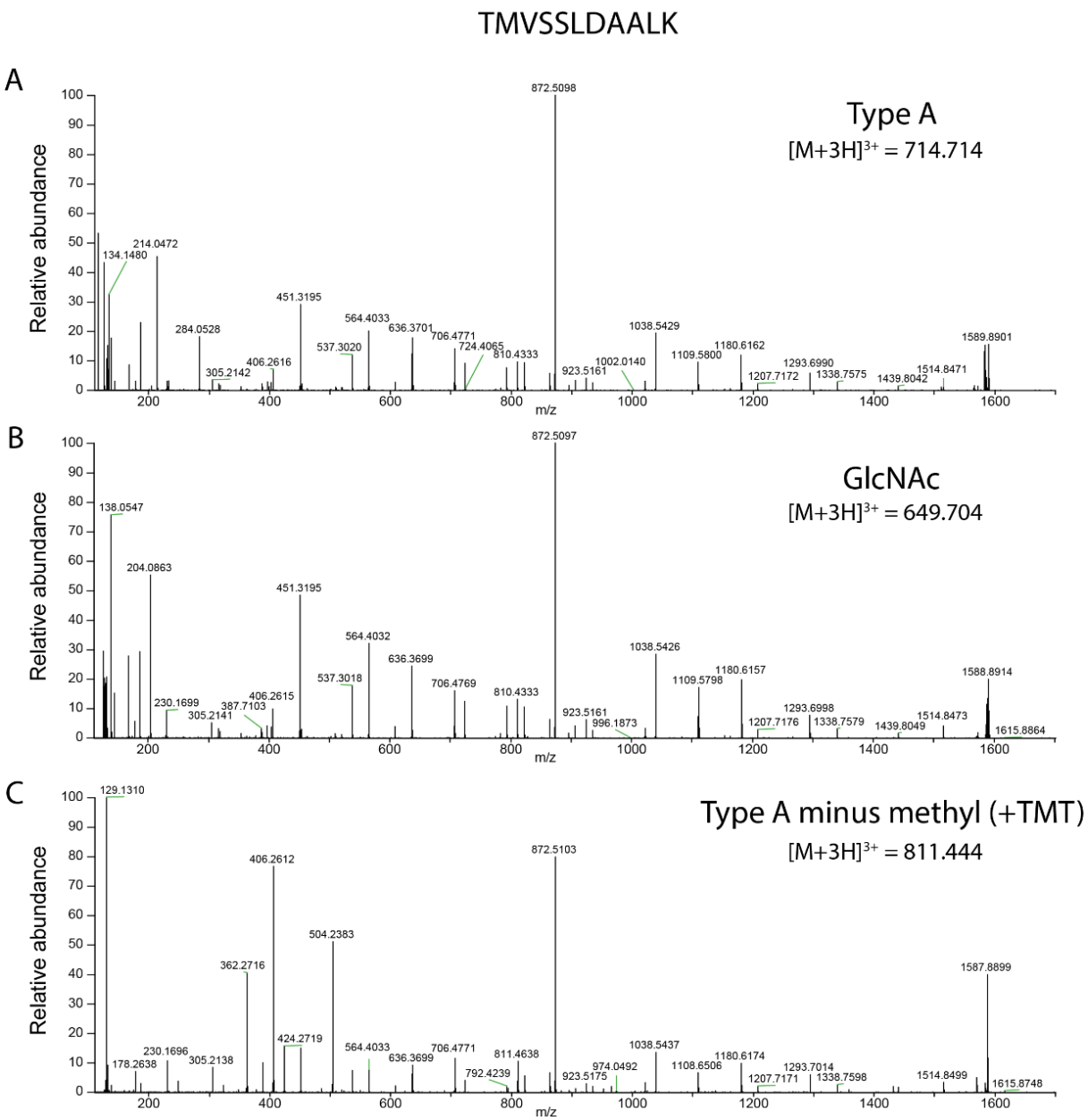


Figure S5. Summed MS/MS spectra of the TMVSSLDAALK peptide displaying the Type A variants. Targeted HCD MS/MS analysis of the TMTpro 16plex labeled strains was performed. MS/MS spectra were summed over the full peak corresponding to the TMVSSLDAALK peptides displaying the complete Type A (**A**), only the GlcNAc (**B**) or Type A minus the methyl group but having an extra TMT label (**C**). The theoretical precursor masses are shown on the right.

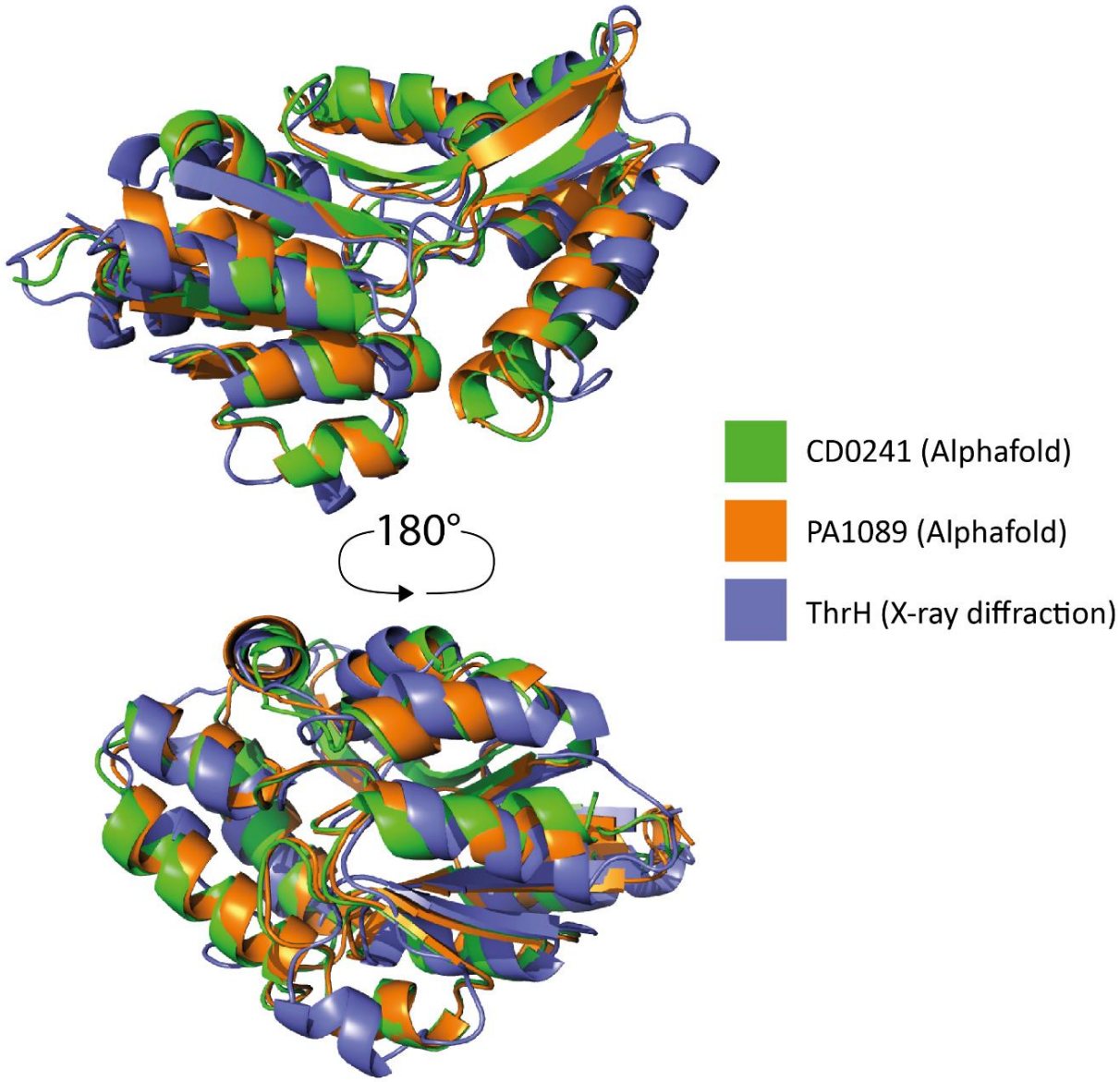


Figure S6. The predicted structures of CD0241 (*C. difficile*) and PA1089 (*P. aeruginosa*) are similar to the experimentally determined structure of ThrH from *P. aeruginosa*. The predicted structures for CD0241 and PA1089 were retrieved from the Alphafold database and superimposed on ThrH (PDB: 1RKU).

3

Characterization of the *Clostridioides difficile* 630Δ*erm* putative Pro-Pro endopeptidase CD1597

Bart Claushuis[1], Arnoud H. de Ru[1], Peter A. van Veelen[1], Paul J. Hensbergen[1§] & Jeroen Corver[2§]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] Leiden University Center of Infectious Diseases, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

§ Authors contributed equally.

Published in Access Microbiology

DOI: https://doi.org/10.1099/acmi.0.000855.v2

Abstract

Clostridioides difficile is the leading cause of antibiotic-associated infections worldwide. Within the host, *C. difficile* can transition from a sessile to a motile state by secretion of PPEP-1, which releases the cells from the intestinal epithelium by cleaving adhesion proteins. PPEP-1 belongs to the group of Pro-Pro endopeptidases, which are characterized by their unique ability to cleave proline-proline bonds. Interestingly, another putative member of this group, CD1597, is present in *C. difficile*. Although it possesses a domain similar to other PPEPs, CD1597 displays several distinct features that suggest a markedly different role for this protein.

We investigated the proteolytic activity of CD1597 by testing various potential substrates. In addition, we investigated the effect of the absence of CD1597 by generating an insertional mutant of the *cd1597* gene. Using the *cd1597* mutant, we sought to identify phenotypic changes through a series of *in vitro* experiments and quantitative proteomic analyses. Furthermore, we aimed to study the localization of this protein using a fluorogenic fusion protein.

Despite its similarities to PPEP-1, CD1597 did not show proteolytic activity. In addition, the absence of CD1597 caused an increase in various sporulation proteins during the stationary phase, yet we did not observe any alterations in the sporulation frequency of the *cd1597* mutant. Furthermore, a promoter activity assay indicated a very low expression level of *cd1597* in vegetative cells that was independent of the culture medium and growth stage. The low expression was corroborated by our comprehensive proteomics analysis of whole cell cultures, which failed to identify CD1597. However, an analysis of purified *C. difficile* spores identified CD1597 as part of the spore proteome. Hence, we predict that the protein is involved in sporulation, although we were unable to define a precise role for CD1597 in *C. difficile*.

Introduction

Clostridioides difficile, a Gram-positive opportunistic gut pathogen, is recognized as a leading cause of healthcare-associated infections worldwide [28,202,203]. A *C. difficile* infection (CDI) manifests primarily as antibiotic-associated diarrhea, but symptoms range from mild, self-limiting disease to severe and life-threatening pseudomembranous colitis and toxic megacolon [27]. The symptoms of CDI are attributed to the production of potent exotoxins, namely Toxin A (TcdA) and Toxin B (TcdB), which disrupt intestinal epithelial integrity [204]. As an obligate anaerobe, *C. difficile* relies on the production of spores for transmission to new hosts via the fecal-oral route [70].

However, the movements of *C. difficile* extend beyond the transmission to new hosts since the bacteria also travel within the host. In the host, *C. difficile* can exist in a sessile state, adhering to the gut epithelium through adhesion proteins, of which CD2831 and CD3246 are two important players [146,147]. Environmental cues, such as nutrient deprivation, can induce a transition to a motile state, characterized by the release from the gut wall and the onset of flagella production [115]. To detach from the gut wall, *C. difficile* secretes the protease PPEP-1, which cleaves the anchoring substrates CD2831 and CD3246 and thereby releases the cell [146,147].

PPEP-1 belongs to the group of Pro-Pro endopeptidases (PPEPs), comprising secreted zinc metalloproteases with the unique ability to cleave proline-proline bonds [146,157]. Beyond *C. difficile*, PPEP homologs have been predicted in several other bacterial species [148]. The second PPEP that was characterized, PPEP-2 from *Paenibacillus alvei*, displays a distinct specificity from PPEP-1, since both proteases cannot hydrolyze each other's substrate [157]. On the other hand, PPEP-2 also cleaves a bacterial cell surface protein that is likely involved in adhesion [157], indicating a common function for PPEPs, although alternative roles are conceivable [148].

Interestingly, a second putative PPEP, CD1597 (UniProt ID: Q186F3), was identified in *C. difficile*. This homolog is distinct from other PPEPs in several ways. First, this putative PPEP lacks a signal peptide for secretion and is presumed to function intracellularly, suggesting a markedly different role for this protein. Second, CD1597 possesses an N-terminal domain of unknown function, constituting approximately half of the protein's structure (**Figure 1A**). This domain is predicted to be linked to the PPEP-like domain through an unstructured (flexible) stretch of residues. Although the presumed catalytic C-terminal domain of CD1597 closely resembles that of PPEP-1 (**Figure 1B**), several amino acid substitutions and insertions are observed (**Figure 1C**). However, the presence of a zinc-binding HEXXH motif in CD1597 (**Figure 1C**) indicates metalloprotease

activity [205]. Therefore, we hypothesized that CD1597 is a zinc metalloprotease with PPEP-like specificity [206].

In this study, we sought to uncover the function of CD1597 in *C. difficile* and thereby explore the diversity of roles played by PPEPs in bacteria. To test CD1597 for proteolytic activity, we tested recombinant CD1597 with potential substrates. Moreover, we investigated an insertional *cd1597* gene mutant for an altered phenotype through a series of *in vitro* experiments and quantitative proteomic analyses.

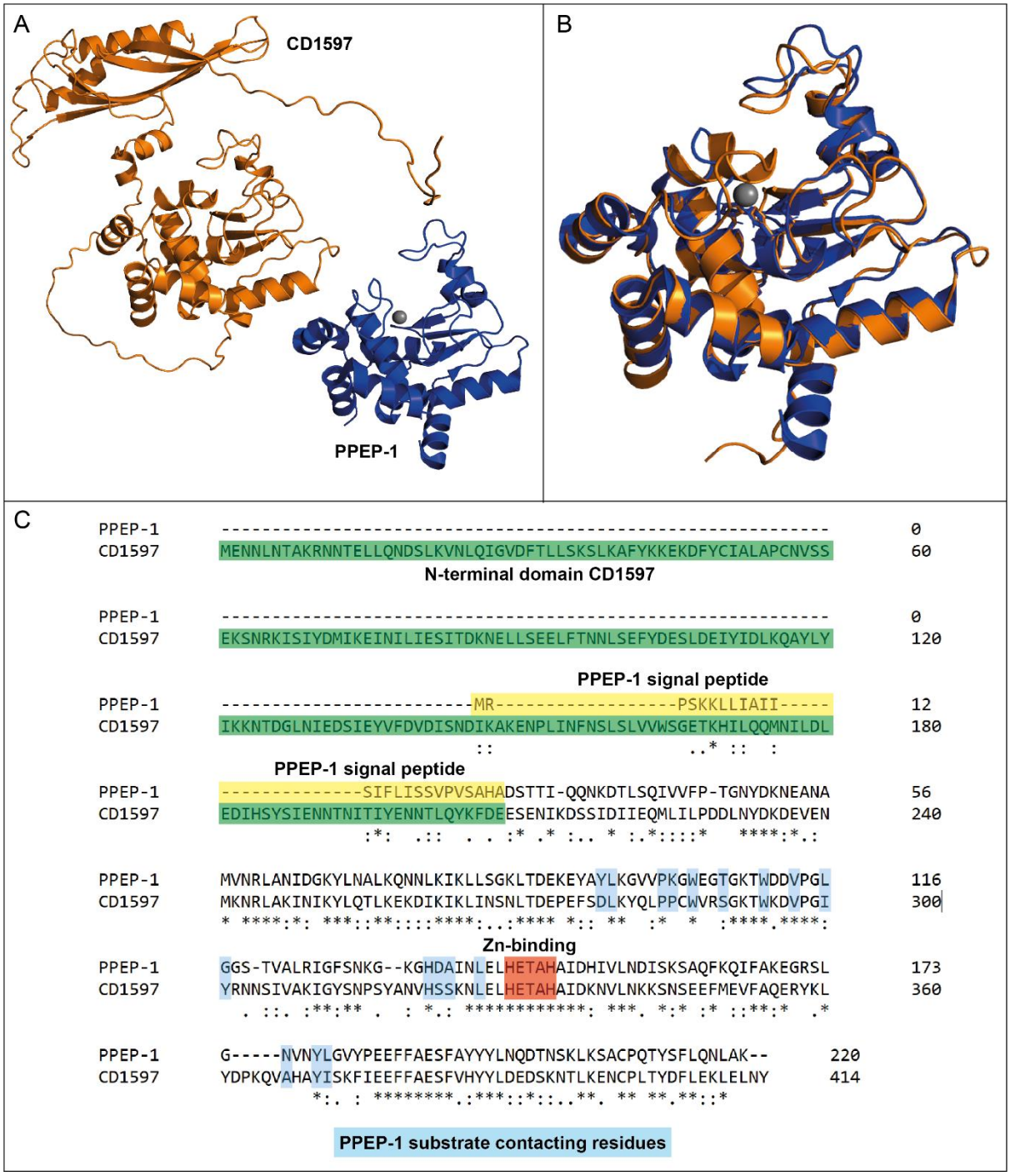


Figure 1. Structural comparison of CD1597 and PPEP-1 from *C. difficile*. A) Predicted structure of CD1597 (Alphafold, orange) and the crystal structure of PPEP-1 (5A0P, blue). The zinc in PPEP-1 is shown in gray. **B)** Superimposition of the catalytic domain of CD1597 (AA211-416, orange) and PPEP-1 (blue). **C)** Sequence alignment of CD1597 and PPEP-1. Sequence alignment was performed using Clustal Omega [207]. The N-terminal domain of CD1597, the PPEP-1 signal peptide, the HEXXH motif and the substrate contacting residues of PPEP-1 are highlighted. Of note, the first two amino acid residues were removed, since the translation start site is wrongly annotated in the UniProt database (UniProt ID: Q186F3).

Results

Purification of recombinant CD1597

To perform *in vitro* assays using CD1597, both the full-length protein and the predicted catalytic domain (AA 211-416) were recombinantly expressed and purified by Immobilized Metal Affinity Chromatography (IMAC) His-tag purification (**Figure 2**). Although the purified protein samples predominantly contained the CD1597 constructs, the full-length purified protein included other proteins as observed by the faint smear on the Coomassie-stained gel (**Figure 2**). To ascertain whether these co-purified proteins included other proteases, an LC-MS/MS analysis was performed, and raw data were searched against both an *E. coli* and *C. difficile* database (**Supplemental Table S1**). Despite the identification of 150 other proteins other than CD1597, none of them were annotated as proteolytic enzymes. In addition, a search for the HEXXH motif, characteristic of metalloproteases, did not reveal any other metalloproteases among the uncharacterized proteins. Based on the LC-MS/MS analysis, we concluded that the purified protein was >90% pure (**Supplemental Table S1**).

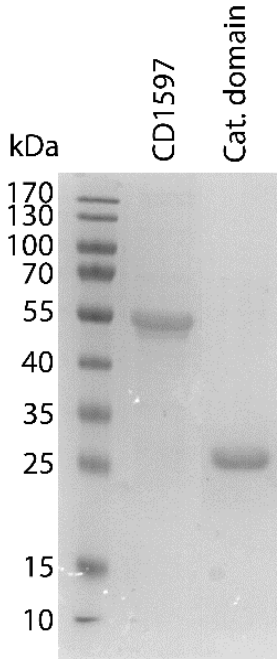


Figure 2. SDS-PAGE analysis of purified CD1597. After IMAC His-tag purification, the fractions from the elution peak were pooled, analyzed by SDS-PAGE, and visualized by Coomassie staining. Lanes from left to right: ladder, full-length CD1597, catalytic domain (AA 211-416). The full-length CD1597 and the catalytic domain have a MW of 50.6 kDa and 26.5 kDa, respectively.

Investigation of the proteolytic activity of CD1597

To evaluate the proteolytic activity of CD1597, we initially incubated CD1597 with a BODIPY TR-X casein substrate (**Figure 3**). Casein, known for its lack of defined tertiary structure, is considered a generic substrate for many proteases [208]. In addition, one of the constituents, β-casein, contains numerous proline residues and also a PPQP sequence, reminiscent of the PPEP-1 cleavage motif [146]. During incubation of BODIPY TR-X casein with CD1597, no increase in fluorescence was observed for both the full-length protein and its catalytic domain, indicating the absence of proteolytic activity toward BODIPY TR-X casein. However, PPEP-1 also did not show any activity toward BODIPY TR-X casein, indicating that casein might not be an appropriate substrate for the highly specific PPEPs.

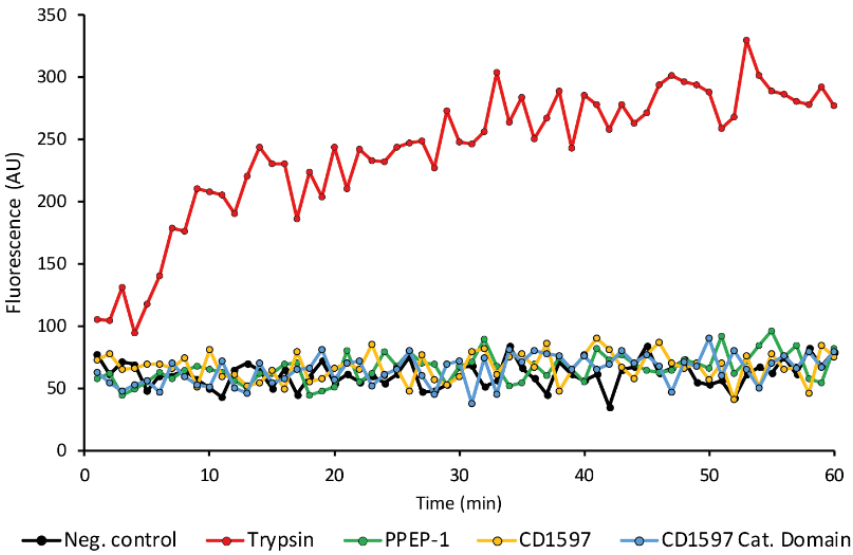


Figure 3. Incubation of BODIPY TR-X casein with CD1597. Proteolysis of the BODIPY TR-X casein substate relieves quenching of the red fluorescent dye and is observed as an increase in fluorescence. The substrate was incubated with CD1597, the CD1597 catalytic domain, PPEP-1 (PPEP control), Trypsin (positive control), and without enzyme (negative control).

To assess the activity of CD1597 against Pro-Pro-containing oligopeptides, we incubated CD1597 with a collection of 38 FRET-quenched peptides that were previously used for the characterization of PPEP-1 and PPEP-2 specificity [146,147,157] (**Figure 4**). While PPEP-1 demonstrated proteolytic activity, as evidenced by the increase in fluorescence for multiple peptides, CD1597 exhibited no activity toward any of the peptides. This observation suggests that either CD1597 does not possess PPEP activity or that it might be inactive toward the specific peptides used.

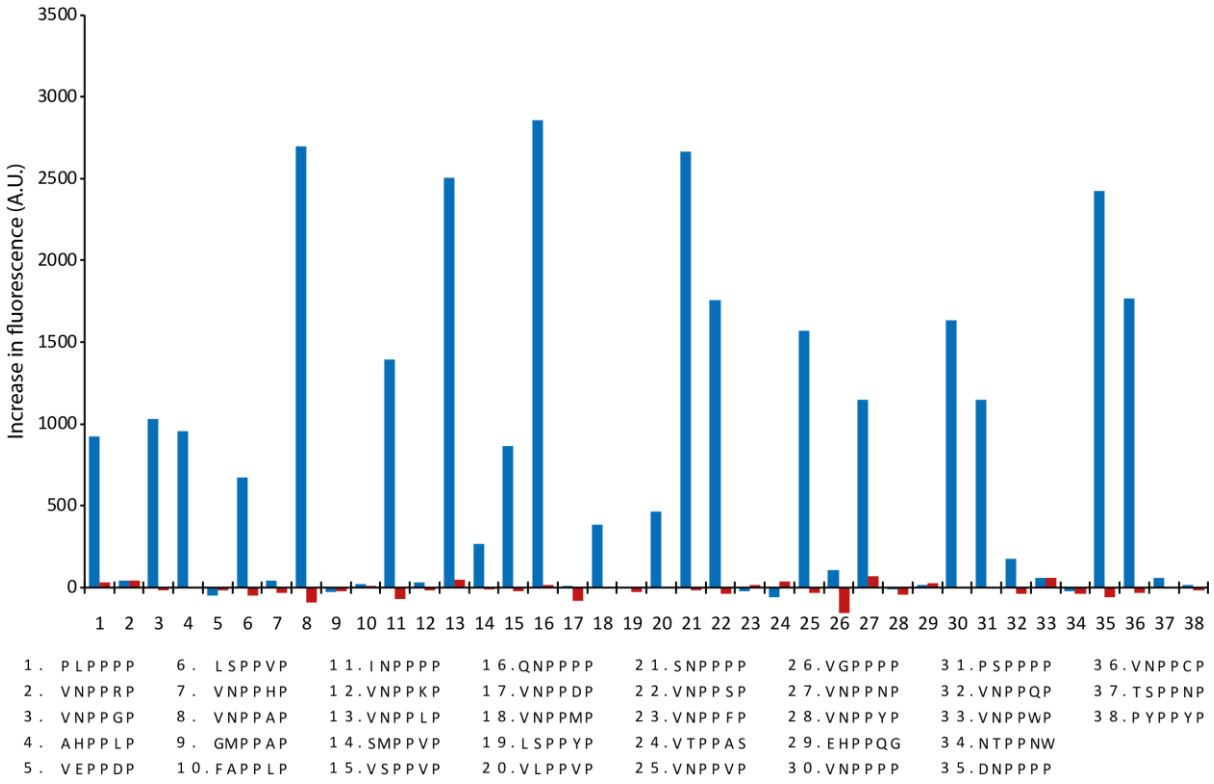
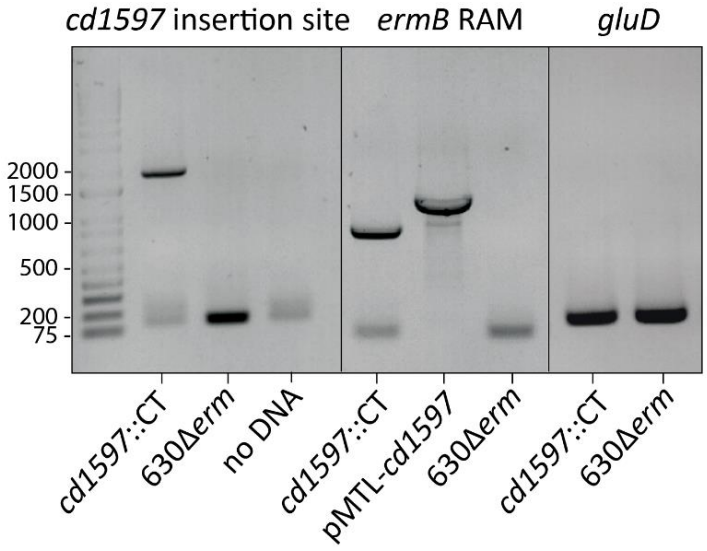


Figure 4. Cleavage of FRET-quenched peptides by CD1597 and PPEP-1. Increase in fluorescent signal after incubation for 1 h with either CD1597 (red) or PPEP-1 (blue) and the synthetic FRET-quenched peptides Lys(Dabcyl)-EXXPPXXD-Glu(EDANS), in which the residues at the X positions vary. For each peptide, the P3-P3' sequence, containing the fixed Pro-Pro at the P1-P1' is shown in the legend.

Generation of a *cd1597*::ClosTron mutant

Given the absence of detectable proteolytic activity, which could offer insights into potential substrates and hence the functionality of CD1597, alternative approaches involving a CD1597 mutant strain were employed to characterize the protein. For this purpose, a *C. difficile* 630Δ*erm* strain that was deficient in the production of CD1597 was generated by insertion of a group II intron in the *cd1597* gene using the ClosTron system [209]. The group II intron was designed to insert between bases 402 and 403 of the *cd1597* gene, positioned upstream of the predicted catalytic domain. The correct genotype was confirmed by PCR (**Figure 5**) and Sanger sequencing. In addition, whole-genome sequencing verified the insertion of the group II intron at a single locus (data not shown).



Product	Size (bp)
cd1597 + insert	1936
cd1597 - insert	136
Spliced intron	±900
Unspliced intron	±1300
gluD	158

Figure 5. Confirmation of the group II intron insertion in cd1597 by PCR. A PCR was performed to amplify the region spanning the predicted insertion site (left panel). The faint lower bands for the PCRs with cd1597::CT and the no DNA control are most likely primer dimers. A second PCR to amplify the ermB RAM intron (middle panel) was performed to discriminate between the spliced and unspliced, i.e., plasmid-based, intron. A third PCR confirmed the strain to be *C. difficile* due to the presence of the gluD gene (right panel).

Deletion of *cd1597* does not affect growth rates

To investigate the effect of the disruption of the *cd1597* gene on the growth rate of *C. difficile*, both the newly generated *cd1597*::CT mutant and the Wild-Type (WT) 630Δ*erm* strain were grown in BHIY medium (**Figure 6A**). The growth curves in BHIY of both strains were nearly identical, with the exception of the OD_{600} at the 24 h time point, where the OD_{600} was consistently lower for the *cd1597*::CT strain (mean OD_{600} 1.67 vs. 1.83). To assess potential medium-dependent effects on growth rates, the strains were also grown in YT (Yeast extract Tryptone, which does not contain glucose or cysteine) medium and CDMM (*C. difficile* minimal medium [210]) (**Figures 6B,C**). This time, a *tcdC*::CT strain was included as a control for the ClosTron mutagenesis, although no growth defects were previously observed in YT medium for this strain [59]. For the growth in YT medium, no differences were observed between the three strains (**Figure 6B**). In CDMM, however, the strains generated by ClosTron mutagenesis did show a reduced growth rate compared to the WT strain (**Figure 6C**), although no difference was observed between the *cd1597*::CT and *tcdC*::CT strains during the exponential phase. However, after 24 h of growth, the *cd1597*::CT strain had grown to a similar OD_{600} as the WT strain and therefore differed from the *tcdC*::CT strain. Based on the results in **Figure 6C**, it is hard to discern if there is an effect on the growth rate due to the absence of CD1597 or the ClosTron mutagenesis.

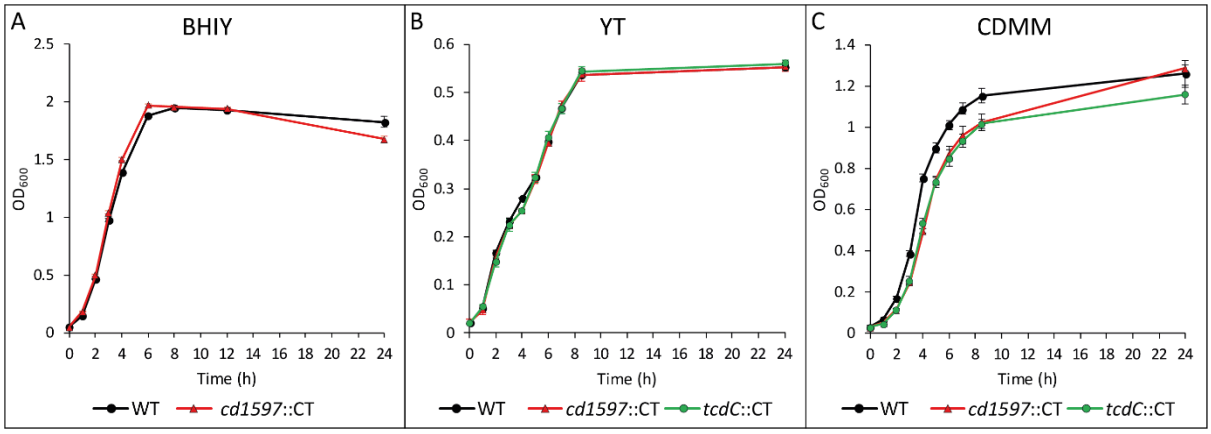


Figure 6. The effect of the disruption of *cd1597* on growth rates in different media. A) Growth curves for the WT (630Δ*erm*) and the *cd1597*::CT strains in BHIY. **B) and C)** Growth curves for the WT, *cd1597*::CT, and *tcdC*::CT strains in YT (**B**) and CDMM (**C**). The points in the growth curves are an average of three replicates. Error bars indicate the standard deviation of the OD_{600}.

Collectively, there is no clear indication that the disruption of *cd1597* influences the growth rate of *C. difficile*. Possibly, the larger drop in OD_{600} of the mutant in BHIY at 24 h might represent a true growth defect, caused by either an earlier onset of the decline phase or a more rapid decline during this phase. Certain events are most likely to be observed in the most nutrient-rich BHIY, since this allows for the most rapid growth of the bacteria and therefore the earliest onset of the decline phase. Yet, additional experiments were needed to characterize the effect of the absence of CD1597.

Expression of *cd1597* in different culture media

Analysis of *C. difficile* strain 630 transcriptome data published by Fuchs et al. (2021) suggested that (1) the expression of *cd1597* is low (e.g., compared to PPEP-1), (2) that *cd1597* expression is higher in YT medium than in BHI, and (3) that expression is higher in the late-exponential growth phase (OD_{600}=0.9) than stationary phase (3 h post entry) [211].

To gain more insight into *cd1597* expression, we generated a plasmid containing the *cd1597* promoter (P*cd1597*) upstream of a codon-optimized gene encoding a secreted luciferase reporter molecule ($sLuc^{opt}$) [212]. We monitored expression of *cd1597* by measuring luciferase activity during growth and in different culture media, while using *ppep-1* expression as a control (**Figure 7**). Luciferase activity, measured by luminescence, was corrected for the OD_{600} at the time of sample collection.

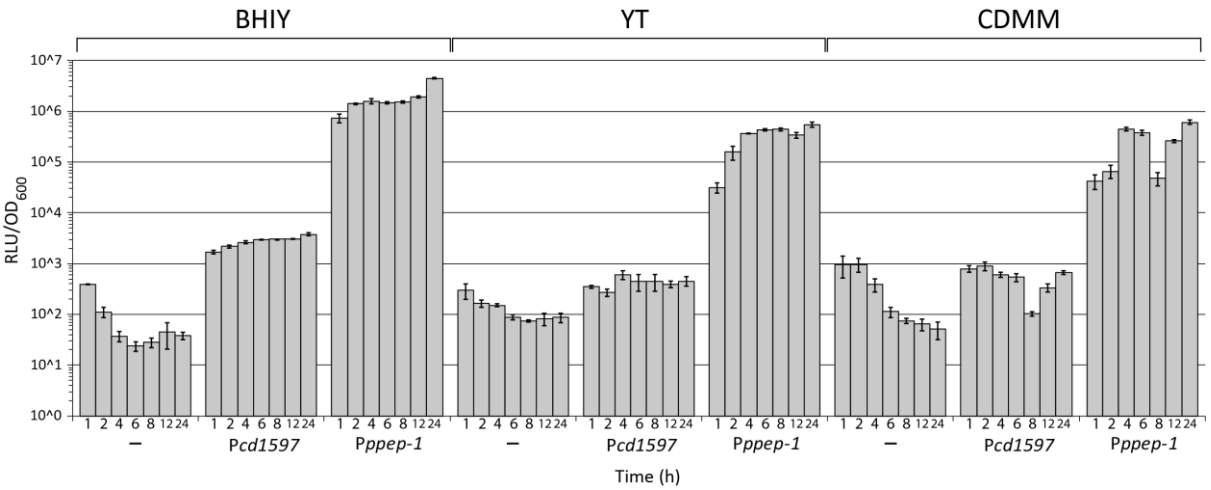


Figure 7. Promoter activity of *cd1597* during growth in different culture media. Strains containing either no plasmid (-), a P*cd1597-sLuc*^{opt} construct, or a P*ppep-1-sLuc*^{opt} construct were grown in BHIY, YT, and CDMM for 24 h. At different time points, OD_{600} was measured and samples were taken to measure luciferase activity. The relative fluorescence units (RFU) were corrected for the OD_{600} at every time point.

Consistent with the transcriptome data [211], P*cd1597* activity was low compared to P*ppep-1* activity (approximately three orders of magnitude lower) and often only marginally exceeded the negative control. In contrast to the transcriptome data, P*cd1597* activity was higher in BHIY than YT medium. However, differences in media composition and experimental setup between our study and that of Fuchs et al. (2021) may account for this discrepancy. In CDMM, luciferase activity was comparable to the negative control during the first two hours of growth, but Pcd1597-driven expression of luciferase was observed, as evidenced by the sustained luminescence levels after correction for OD_{600} (as was the case for the negative control). Overall, while PPEP-1 expression levels seemed to increase over time, in line with the model for the regulation of PPEP-1 [115], a similar trend was less evident for CD1597. In BHIY, a slight increase in promoter activity was apparent over time, but this was not mirrored in the other media.

For the negative control lacking a plasmid, luciferase activity appeared to decrease over time due to the constant background signal being divided by increasing OD_{600}. For the CDMM cultures, a sudden drop is observed at 8 h due to low luciferase activity, yet the reason for this is unknown. Nevertheless, since this is observed for both the *cd1597* and *ppep-1* promoters, this is unlikely to represent a biological phenomenon.

Collectively, these findings indicate a consistently low expression of CD1597 that is not influenced by the growth phase but was highest in the most nutrient-rich BHIY medium.

Effects of the mutation of *cd1597* on protein levels in *C. difficile* 630Δ*erm*

Since there was no evidence of proteolytic activity, a growth defect, or a growth phase-dependent expression of *cd1597*, a mass spectrometry (MS) based quantitative proteomics approach using TMTpro 16plex labeling was taken to investigate differences in protein levels between the WT and the *cd1597*::CT strains. Again, as a control for the potential effects of ClosTron mutagenesis, the *tcdC*::CT strain was included. Bacterial cultures were grown in BHIY and harvested during both the mid-logarithmic phase (OD_{600}=0.8) and stationary phase (22 h). Volcano plots were generated to display the differences in protein expression across these cultures (**Figure 8**).

We identified 2267 proteins with high confidence and at least two peptides (from a total of 2521 proteins) (**Supplemental Table S2**), which, to the best of our knowledge, represents one of the most comprehensive proteomic analyses of *C. difficile* [213–215]. Among the identified proteins was TcdC, which can only originate from the WT and *cd1597*::CT strains. However, the ratio *tcdC*/WT for this protein was 1.01 and 0.45 for the mid-logarithmic and stationary phases, respectively. The fact that the ratio did not approach zero was indicative of the phenomenon called ratio expression [216], which is caused by the co-fragmentation of other peptides in MS2, thereby resulting in an underestimation of the true differences in protein levels. Therefore, we decided to include proteins that showed >1.5-fold increase or decrease in the *cd1597* mutant in our analysis.

During the mid-logarithmic phase, few proteins were differently expressed due to the mutation of *cd1597* (**Figures 8A,B**). Of note, the protein with the lowest ratio *cd1597/tcdC* in **Figure 8B** (upper left blue dot) was ErmB (the antimicrobial resistance gene introduced by ClosTron mutagenesis), which is for reasons unknown more highly expressed in the *tcdC*::CT strain. However, more differences in protein expression were observed during the stationary phase (**Figures 8C,D**), indicating that the mutation of *cd1597* primarily impacts the bacteria during this growth phase.

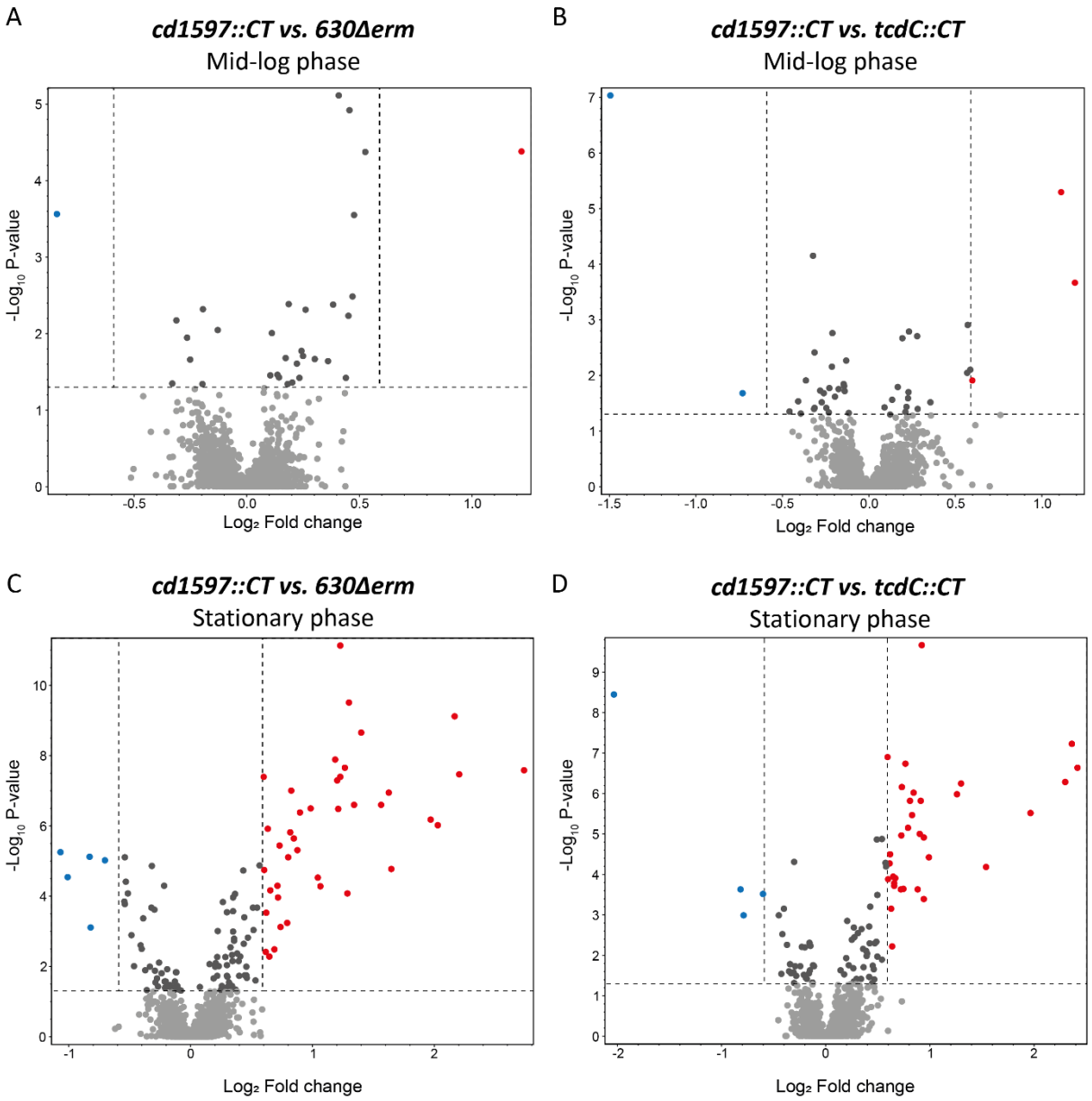


Figure 8. Differences in protein expression in the *cd1597*::CT strain. An overall comparative proteomic experiment using TMTpro 16plex labeling was performed with the WT, *cd1597*::CT, and *tcdC*::CT strains. Differences in protein levels were displayed in Volcano plots. Only proteins identified with high confidence and ≥2 PSMs were included in the analysis. Proteins that are >1.5-fold decreased in the *cd1597*::CT strain are shown in blue and those >1.5 increased in red. **A)** Differently expressed proteins between the *cd1597*::CT and WT strains during the mid-logarithmic phase. **B)** Differently expressed proteins between the *cd1597*::CT and *tcdC*::CT strains during the mid-logarithmic phase. **C)** Differently expressed proteins between the *cd1597*::CT and WT strains during the stationary phase. **D)** Differently expressed proteins between the *cd1597*::CT and *tcdC*::CT strains during the stationary phase.

For the stationary phase, we found 55 proteins to be differently expressed (ratio <0.66 or >1.5) when comparing the *cd1597* mutant to the WT, 45 proteins when comparing the *cd1597* mutant to the *tcdC* mutant, and 39 of these proteins were differently expressed compared to both (**Table 1**). Several proteins were thus only differently expressed when comparing the *cd1597* mutant to the WT. In addition, of the 39 proteins that were differently expressed in both comparisons, several had a larger difference when comparing the *cd1597* mutant to the WT than to the *tcdC* mutant. This indicated an effect due to the ClosTron mutagenesis rather than the mutation of *cd1597*. Indeed, when comparing our data to similar data of an unrelated strain that was similarly generated using ClosTron mutagenesis, we saw a large overlap in differently expressed proteins (data not shown), corroborating the idea that ClosTron mutagenesis affects protein expression. Therefore, we only included proteins in **Table 1** that were differently expressed in the *cd1597* mutant compared to both the WT and the *tcdC* mutant, and in addition added a remark if we believed a protein to be differently expressed due to the ClosTron mutagenesis. This remark was based on an integrated analysis that considered whether (1) the proteins showed a similar difference in expression as in the comparative proteomics data of the unrelated ClosTron mutant, (2) the ratio *cd1597*/WT was higher than the ratio *cd1597*/*tcdC*, and (3) whether the genes were part of an operon.

Table 1. Differently expressed proteins in the *cd1597*::CT strain during mid-logarithmic and stationary phase.

Increased expression in *cd1597* ::CT during mid-logarithmic phase					
Locus tag	**Gene**	**Description**	***cd1597*/WT**	***cd1597*/*tcdC***	**Remark**
CD3275		Putative phosphosugar isomerase	2.33	2.28	
Decreased expression in *cd1597* ::CT during mid-logarithmic phase					
Locus tag	**Gene**	**Description**	***cd1597*/WT**	***cd1597*/*tcdC***	**Remark**
-	-	-	-	-	
Increased expression in *cd1597* ::CT during stationary phase					
Locus tag	**Gene**	**Description**	***cd1597*/WT**	***cd1597*/*tcdC***	**Remark**
CD1199	*spoIIIAH*	Stage III sporulation protein AH	6.67	5.35	
CD2688	*sspA*	Small, acid-soluble spore protein alpha	4.61	5.15	
CD3275		Putative phosphosugar isomerase	4.09	3.92	
CD3249	*sspB*	Small, acid-soluble spore protein beta	3.92	4.93	
CD1065	*cotL*	Morphogenetic spore coat protein	3.53	4.39	
CD3567	*sipL*	Spore coat protein	3.14	2.91	
CD2960	*atpI*	V-type ATP synthase subunit I	2.96	1.99	Likely ClosTron effect
CD2961		Uncharacterized protein	2.64	1.80	Likely ClosTron effect
CD2629	*spoIVA*	Stage IV sporulation protein A	2.54	2.40	
CD2656	*spoVD*	Stage V sporulation protein D (Sporulation-specific penicillin-binding protein)	2.48	2.37	
CD2958	*atpE*	V-type ATP synthase subunit E	2.47	1.70	Likely ClosTron effect
CD2955	*atpB*	V-type ATP synthase beta chain	2.41	1.78	Likely ClosTron effect
CD2959	*atpK*	V-type ATP synthase subunit K	2.34	1.66	Likely ClosTron effect
CD1657	*gcvTPA*	Multifunctional fusion protein	2.34	1.90	Likely ClosTron effect
CD1658	*gcvPB*	Probable glycine dehydrogenase (decarboxylating) subunit 2	2.32	1.93	Likely ClosTron effect
CD2954	*atpD*	V-type ATP synthase subunit D	2.30	1.73	Likely ClosTron effect
CD2956	*atpA*	V-type ATP synthase alpha chain	2.28	1.76	Likely ClosTron effect
CD0770	*spoIIAA*	Anti-sigma F factor antagonist	2.09	1.92	
CD2373		Putative CstA-like carbon starvation protein	2.06	1.85	Likely ClosTron effect
CD1935	*spoVS*	Stage V sporulation protein S	1.98	1.89	
CD2957	*atpC*	V-type ATP synthase subunit C	1.86	1.53	Likely ClosTron effect
CD0780		Uncharacterized protein	1.84	1.58	Likely ClosTron effect
CD0725		Bifunctional carbon monoxide dehydrogenase/acetyl-CoA synthase, subunit delta	1.80	1.52	Likely ClosTron effect
CD2737		Putative nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	1.78	1.66	Possibly ClosTron effect
CD0723		Dihydrolipoyl dehydrogenase	1.75	1.57	Likely ClosTron effect
CD0777		Putative membrane protein	1.74	1.56	
CD0772	*sigF*	RNA polymerase sigma factor	1.70	1.68	
CD0663	*tcdA*	Toxin A	1.66	1.54	Possibly ClosTron effect
CD2738		Putative cytosine permease	1.65	1.65	Likely ClosTron effect
CD0721		MTHFR_C domain-containing protein	1.64	1.59	Likely ClosTron effect
CD3637		Putative NADPH-dependent FMN reductase	1.64	1.53	
CD0438		Uncharacterized protein	1.58	1.58	
CD0855	*oppA*	ABC-type transport system, oligopeptide-family extracellular solute-binding protein	1.54	1.55	Likely ClosTron effect
CD0722	*metF*	Methylenetetrahydrofolate reductase	1.52	1.50	Likely ClosTron effect
CD0779		Peptidase M20 domain-containing protein 2	1.52	1.51	Likely ClosTron effect
Decreased expression in *cd1597* ::CT during stationary phase					
Locus tag	**Gene**	**Description**	***cd1597*/WT**	***cd1597*/*tcdC***	**Remark**
CD0268	*flgG1*	Flagellar basal body protein	0.56	0.66	
CD3192	*cwp21*	Putative cell surface peptidase, M4 family-cwp20	0.50	0.58	Possibly ClosTron effect
CD0226		Putative lytic transglycosylase	0.48	0.57	

During the mid-logarithmic phase, only a single protein, a putative phosphosugar isomerase, was differently expressed in the *cd1597*::CT strain (**Table 1**). Interestingly, this protein was also more highly expressed during the stationary phase and therefore the only protein differently expressed during both growth phases.

In the stationary phase, more proteins were differently expressed in the *cd1597*::CT strain, of which many are thought to be the result of the ClosTron mutagenesis itself. However, when considering the proteins that an increased expression due to the lack of CD1597, a common theme emerges. Most of these proteins are directly linked to sporulation, either by regulating sporulation genes (SigF and SpoIIAA), the development of spores (SpoIIIAH, SpoIVA, SpoVD, and SpoVS), or being part of the finished spore (SspA, SspB, CotL, SipL) (**Table 1**). Of note, the anti-Sigma factor F protein, SpoIIAB, was also higher expressed (ratio *cd1597*/WT=1.55 and *cd1597*/*tcdC*=1.46, **Supplemental Table S2**). SpoIIAA and SpoIIAB, located in an operon, regulate SigF-directed transcription of forespore-specific genes involved in the early stages of sporulation [217,218].

Increased protein levels were also observed for the main virulence factors Toxin A (**Table 1**) and, to a lesser extent, Toxin B (ratio *cd1597*/WT=1.43 and *cd1597*/*tcdC*=1.34, **Supplemental Table S2**). Since these toxins are secreted from the cells, no distinction could be made between increased expression or reduced secretion of the proteins based on our data.

The most notable downregulated proteins are CD0226 and FlgG1, which are both part of the flagellar gene cluster (*cd0226-cd0272*, [219]). CD0226 is the first gene of an operon, yet a similar downregulation is not observed for the downstream genes (**Supplemental Table S2**).

Although we identified a large set of proteins, CD1597 was not identified. This further demonstrated the low expression of this protein which was also observed in **Figure 7**.

Collectively, the most notable difference in protein expression was the increase in sporulation-related proteins during the stationary growth phase.

Sporulation frequency is not affected in the *cd1597*::CT strains

Due to the elevated levels of several sporulation proteins in the *cd1597*::CT strain, we hypothesized that the sporulation frequency could affected due to the mutation in *cd1597*. To test this, we conducted microscopy-based assays to evaluate the sporulation frequencies of the WT, *cd1597*::CT, and *spo0A*::CT (negative control) strains (**Figure 9A,B**). We observed no difference in sporulation frequency when comparing the *cd1597*::CT strain to the WT. Furthermore, no difference was observed in average cell length when comparing the WT to the *cd1597*::CT strain (**Figure 9C**) and no other phenotypic changes were observed as a result of the mutation of *cd1597*.

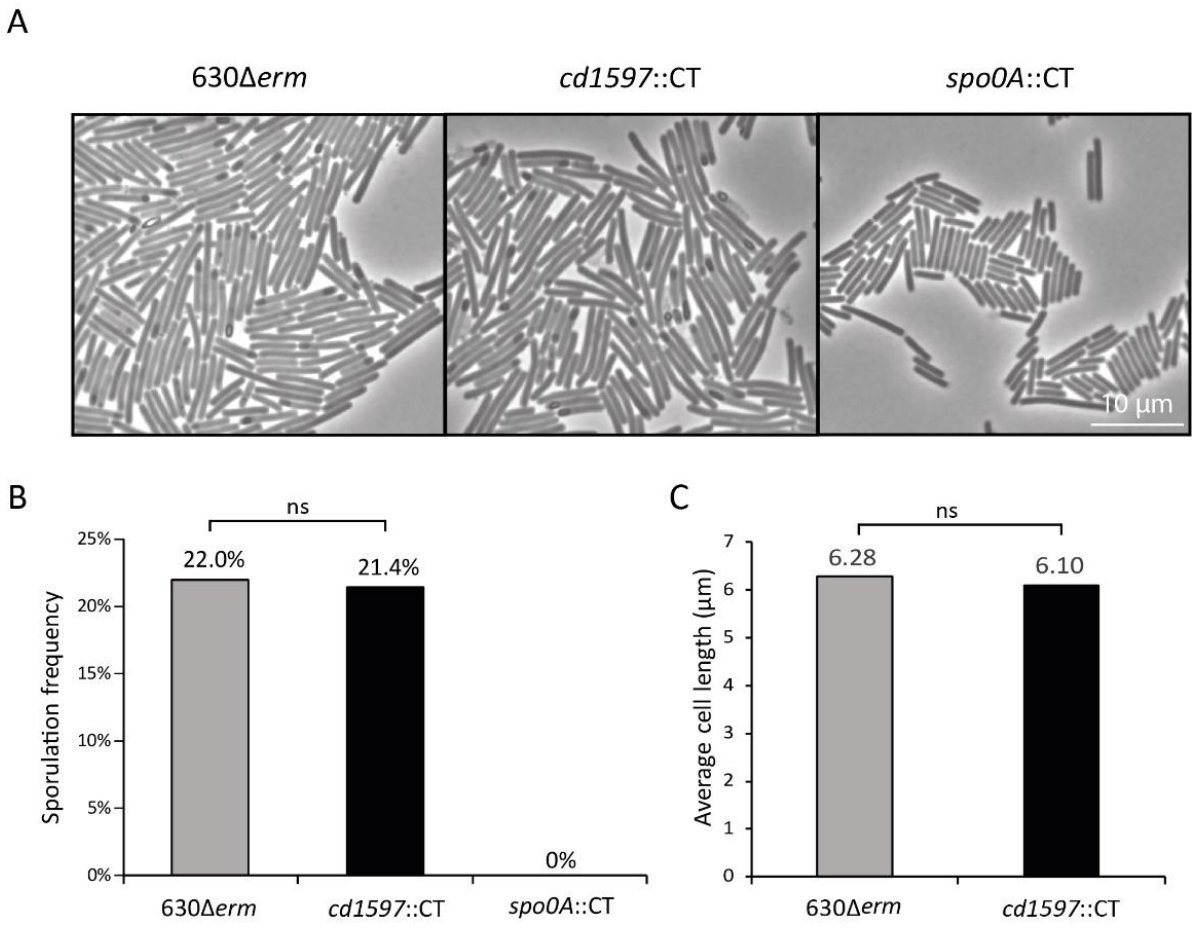


Figure 9. Sporulation frequency and cell length are not affected by the mutation of *cd1597*. A) Phase-contrast micrographs of the WT (630Δ*erm*), *cd1597*::CT, and *spo0A*::CT strains. The *spo0A*::CT strain was included as a non-sporulating control. **B)** Sporulation frequencies were determined by counting the total amount of completed spores and developing spores and dividing this by the total amount of cells (spores + vegetative cells). The difference between the sporulation frequencies of the WT and *cd1597*::CT strains was not significant as determined by a Chi-squared test (X^2 (1, N = 13148) = 0.641, p = 0.423). **C)** Analysis of cell length of strains 630Δ*erm* and *cd1597*::CT. The length of cells was determined by analyzing micrographs using Fiji (ImageJ). There was no significant difference in cell length between the 630Δ*erm* (M = 6.28, SD = 2.02) and *cd1597*::CT (M = 6.10, SD = 2.03) strains as determined by an independent samples t-test; t(638) = 1.0996, p = 0.2719.

Comparative proteomics analysis of spores of the *cd1597*::CT strain

Although no differences were observed in sporulation frequencies, we investigated the spore proteome of the *cd1597*::CT strain in a comparative proteomics analysis of purified spores of *C. difficile* to identify any differences in spore composition (**Figure 10**). Again, we included both the WT and the *tcdC*::CT strains as controls.

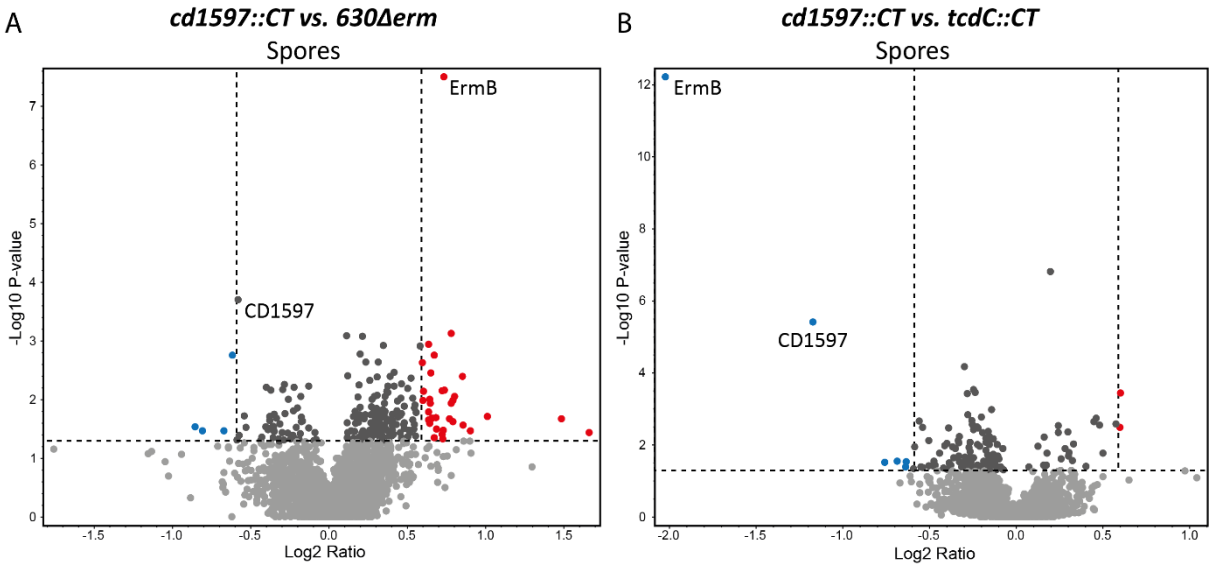


Figure 10. Differences in the spore proteome due to the mutation of *cd1597*. An overall comparative proteomics experiment using TMTpro 16plex labeling was performed on spores of the WT (630Δ*erm*), *cd1597*::CT, and *tcdC*::CT strains. Differences in protein levels were displayed in Volcano plots. Only proteins identified with high confidence and ≥2 PSMs were included in the analysis. Proteins that are >1.5-fold decreased in the *cd1597*::CT strain are shown in blue and those >1.5 increased in red. **A)** Differences in protein levels between the *cd1597*::CT and WT strains. **B)** Differences in protein levels between the *cd1597*::CT and *tcdC*::CT strains.

We identified 2401 proteins with high confidence and at least two peptides (from a total of 2688 proteins) (**Supplemental Table S3**). This high number of proteins likely indicated an incomplete separation of the spores from the vegetative cells, but the data also indicated a clear enrichment of spore proteins (e.g., the spore-coat proteins). Similar to the proteomics analyses of the whole cell cultures (**Figure 8**), we found a greater number of proteins to be differently expressed when comparing the *cd1597*::CT strain to the WT strain than when comparing the *tcdC*::CT ClosTron control (**Figure 10**). Therefore, we again focused on differently expressed proteins in both comparisons.

First of all, we identified six different peptides of CD1597 in the analysis of *C. difficile* spores (**Figure 10**). In addition, other proteomics analyses of spores also identified CD1597 (data not shown). The fact that CD1597 is exclusively identified in spores

indicates that this protein is part of the spore proteome rather than the vegetative cell proteome. Although CD1597 could only originate from the WT and *tcdC*::CT samples, we again observed ratio compression since the ratios for CD1597 do not approach zero (ratio *cd1597*/WT=0.669 and *cd1597*/*tcdC*=0.444).

We looked at proteins that showed >1.5-fold increase or decrease in the *cd1597*::CT strain compared to both the WT and the *tcdC*::CT strains. Three proteins, namely UxaA', CD3003, and CD3391, demonstrated higher levels in spores of the cd1597 mutant. However, due to various reasons that included limited peptides/PSMs identification and high variance, these differences in these protein levels were not considered statistically significant.

Although the mutation of *cd1597* does not greatly influence the spore proteome (apart from the effects of the ClosTron mutagenesis), CD1597 is identified in spores on multiple occasions, suggesting a role for CD1597 in sporogenesis, spore integrity/resistance, or germination.

Localization and overexpression of CD1597

Since CD1597 was identified in spores, we investigated whether CD1597 localizes to spores during bacterial growth. For this, we constructed an expression vector containing an inducible *cd1597-cfp^{opt}* gene. The localization of the CD1597-CFP^{opt} product was analyzed using fluorescence microscopy at different time points and concentrations of the inducer anhydrotetracycline (ATc) (**Figure 11**). As a control, we included a vector that expresses CFP^{opt}.

The localization of CFP^{opt} was mostly cytosolic, but at 200 ng/ml ATc, CFP^{opt} also localized to the spores (**Figure 11**). For the CD1597-CFP^{opt}, however, we observed a different pattern of localization. After 4 h, CD1597-CFP^{opt} localized to the poles of the cells at 25 ng/ml ATc, but the higher expression at 200 ng/ml also showed localization to midcell, and sometimes additional CFP signals were observed between the poles and midcell. At 24 h and 48 h, induction of the fusion protein resulted in dark spots in the phase contrast images, which in most cases did not coincide with a signal in the fluorescent images. Especially at 200 ng/ml, we also observed the formation of longer cells, suggesting a defect in cell division. Together, these results indicate the formation of inclusion bodies that contain the protein aggregates of insoluble CD1597-CFP^{opt}. Bacterial inclusion bodies are formed at the cell poles and cause abnormal cell division [220], resulting in a phenotype that is in line with the observations in **Figure 11**. The inclusion bodies, appearing as dark black spots in the cell, might render CFP non-fluorescent since in most cases these black spots do not produce a fluorescent signal.

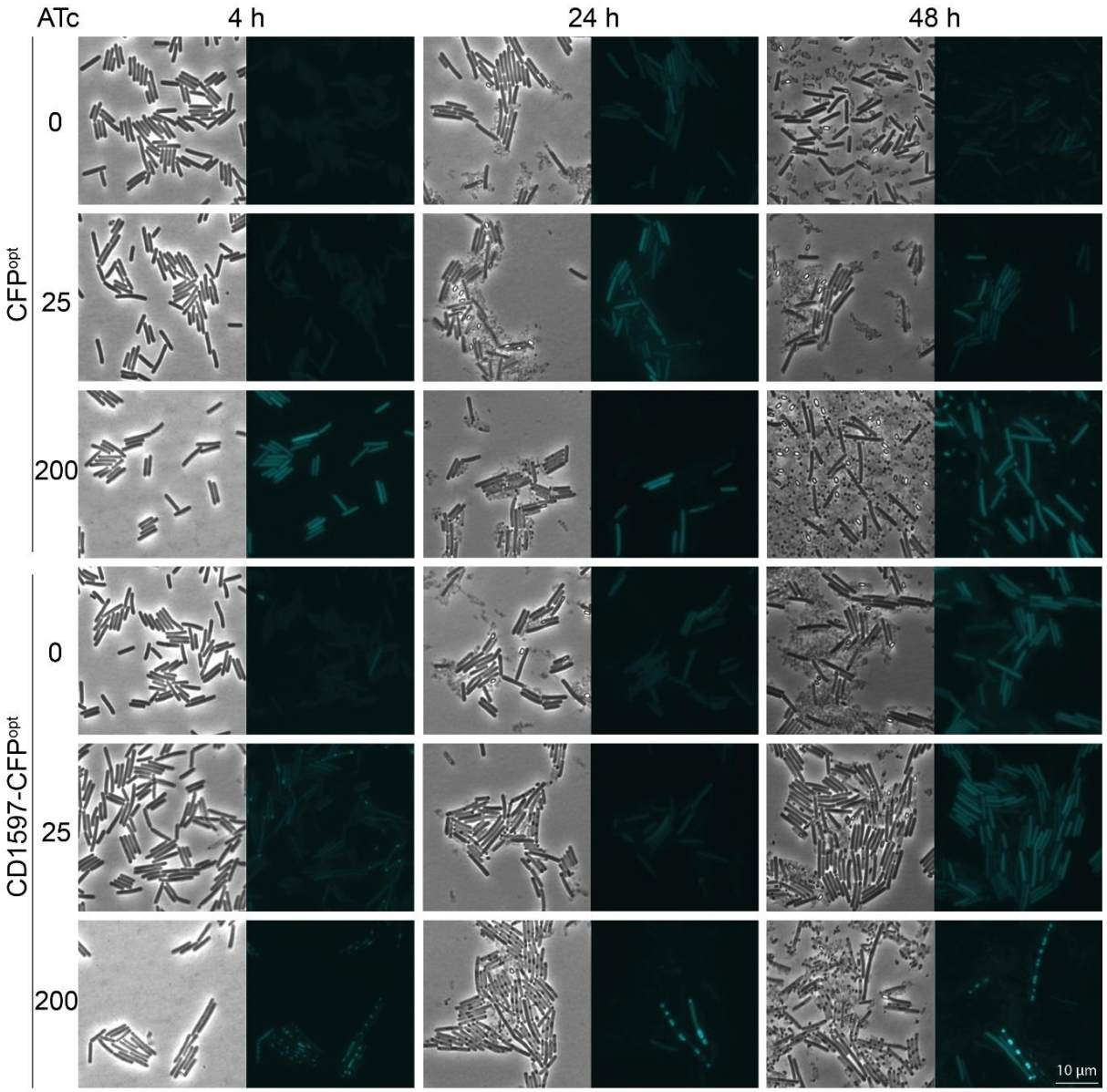


Figure 11. Localization of CD1597-CFP^{opt}. *C. difficile* strains containing vectors for the inducible expression of CFP^{opt} or CD1597-CFP^{opt} were grown in BHIY and analyzed by phase contrast and fluorescence microscopy. Expression was induced by adding 25 ng/ml or 200 ng/ml ATc. Samples were taken from the cultures after 4 h, 24 h, and 48 h.

To confirm that the observed phenotype resulting from the overexpression of the CD1597-CFP^{opt} fusion protein is due to the formation of inclusion bodies and not due to an effect of CD1597, we analyzed cells that overexpressed CD1597 by microscopy (**Figure 12**). The induction of CD1597 did not result in a similar phenotype as observed in **Figure 11**. In addition, no difference was observed in either cell length or sporulation between the cultures induced by ATc and the uninduced WT and *cd1597* mutant. Therefore, we conclude that it is not the presence of CD1597 that causes the phenotype observed in **Figure 11**, but rather the insolubility of the CD1597-CFP^{opt} fusion protein that leads to the formation of inclusion bodies.

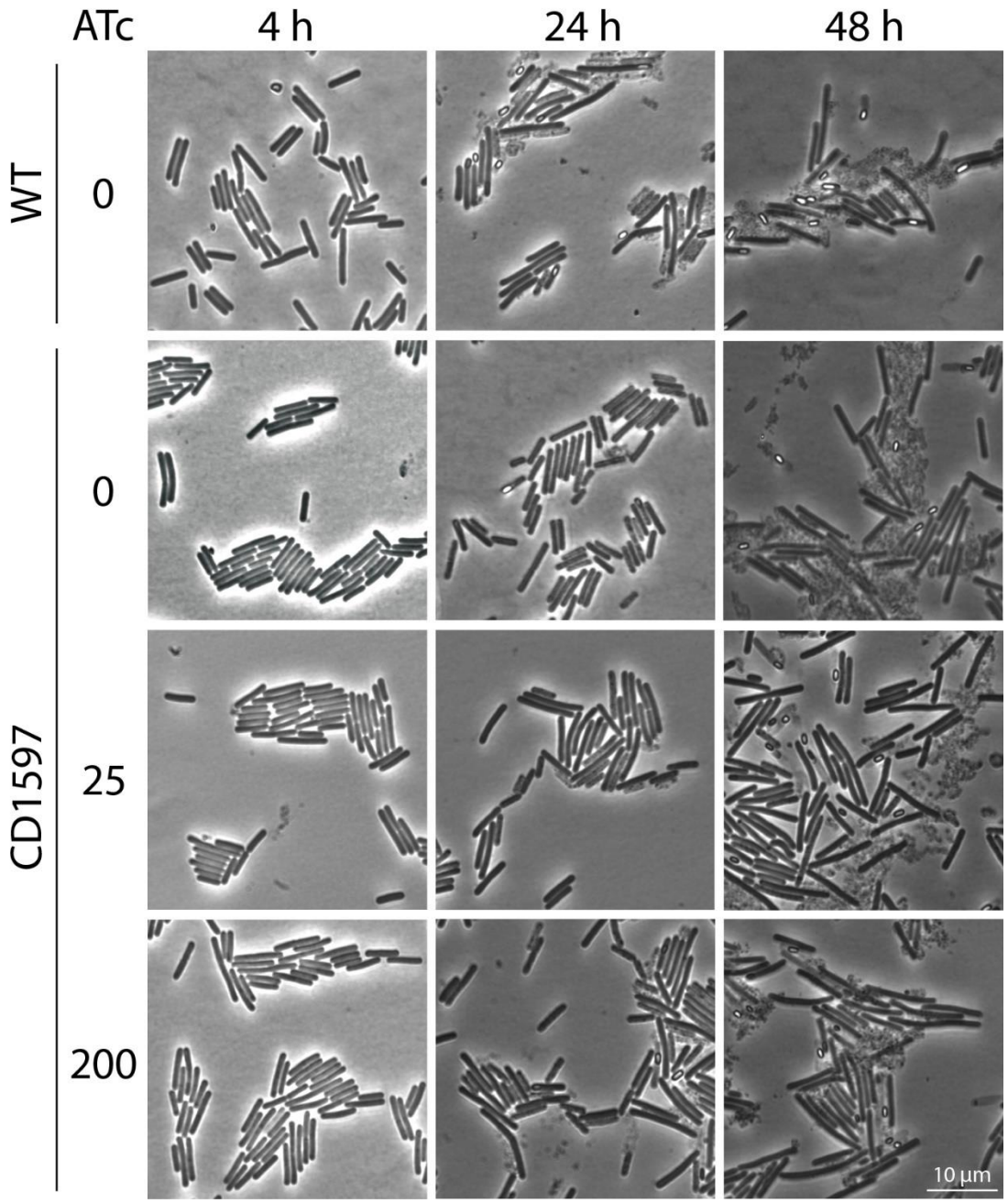


Figure 12. Overexpression of CD1597 does not produce a distinct phenotype. *C. difficile* strains with either no vector or a CD1597 expression construct were grown in BHIY and analyzed using phase contrast microscopy. Expression was induced using 25 ng/ml or 200 ng/ml ATc. Samples were taken from the cultures after 4 h, 24 h, and 48 h.

Discussion

Due to the structural resemblance to other PPEPs, we hypothesized that CD1597 displays PPEP-like proteolytic activity. However, no proteolytic activity was observed in two separate assays. The inability of CD1597 to cleave casein might be unsurprising, as PPEP-1 also lacks this capability. However, also a diverse collection of FRET-quenched peptides containing two consecutive prolines in their core, of which several are cleaved by PPEP-1, were not cleaved by CD1597. Of course, these results do not exclude the possibility that CD1597 is proteolytically active, since we might not have supplied the protein with the correct substrate(s). Assays using peptide libraries that offer a large range of potential substrates might overcome this limitation. However, previous investigations into the activity of CD1597 using a synthetic combinatorial peptide library specifically designed to profile PPEP specificity did not show any proteolytic activity toward Pro-Pro-containing peptides [206]. Alternatively, the substitution and insertion of residues compared to PPEP-1 could render the protein an inactive pseudoprotease, although CD1597 possesses an intact HEXXH domain that suggests metalloprotease activity [221,222].

Although we observed no obvious growth defect for the *cd1597*::CT strain, the OD_{600} was lower after 24 h than the WT in BHIY medium, suggesting a larger decline in cell numbers. Around this moment of the growth phase, we harvested our cells for the comparative proteomics experiment and saw an effect of the absence of CD1597. Possibly, the differences in protein levels caused by the mutation of *cd1597*, either due to the ClosTron mutagenesis or the lack of CD1597, results in a faster decline of OD_{600} in BHIY after 24 h.

We observed compression of the ratios (*cd1597*::CT/control) in both proteomics analyses using the OrbiTrap Exploris 480 mass spectrometer and TMTpro labels for quantification, which led to an underestimation of the differences in protein levels. In our experimental setup, this is likely the result of co-eluting ions with similar m/z values. We extensively fractionated our samples on an HPLC system to reduce the co-isolation of ions and also increase protein/peptide identifications. Another method to reduce ratio compression is by performing MS3 fragmentation [223], but this requires alternative instruments. We performed MS3 analysis on an Orbitrap Fusion Lumos instrument using the same spore proteome samples, which slightly improved the ratios for CD1597, but also reduced the number of proteins, peptides, and PSMs (data not shown).

ClosTron mutagenesis was used to produce the *cd1597*::CT insertion mutant. An initial comparison of the proteomes of the *cd1597*::CT and the WT strains showed many proteins to be differently expressed. Only after comparing this data to previous data

with an unrelated ClosTron mutant, we observed an overlap in differently expressed proteins, thereby indicating an effect of the ClosTron mutagenesis itself. Given the frequent use of the ClosTron system in *C. difficile* research, it is challenging to assess the implications of the secondary effects stemming from mutagenesis on past findings. For example, it has been shown that animals that are challenged with ClosTron mutants have a reduced survival time compared to those challenged with the WT strain [224]. To compare our results to that of others, we searched for quantitative proteomics data from other studies that analyzed the full proteome of vegetative cells of both the WT and ClosTron mutants. We identified a single study by Pettit et al. (2014, [225]) that investigated a *spo0A*::CT mutant. The absence of Spo0A, the master regulator of sporulation, has a large and pleiotropic effect on the bacteria and is therefore not suited to compare to our data for investigating the effects of the ClosTron mutagenesis. To study the secondary effects of ClosTron mutagenesis in more detail, studies using an insertional mutation in a non-coding region could identify these secondary effects more precisely and such a strain could provide a valuable control strain in experiments using ClosTron mutants.

We observed an increase of both the glucosylating exotoxins TcdA and TcdB from *C. difficile*. However, since these toxins are secreted from the cell, we could not discriminate between an increase in expression or a reduced secretion of these proteins. In our proteomic analysis, we did not identify TcdE, the holin-like protein involved in the secretion of the toxins [56] and can therefore not speculate on the amount of secretion based on TcdE levels. Also, it is unknown whether this increase in cytoplasmic toxins results from the absence of CD1597 or due to the ClosTron mutagenesis. The quantitative proteomics data from the unrelated ClosTron mutant also showed a similar increase in toxin expression, but other studies with ClosTron mutants that exhibited increased toxin expression showed that the toxin levels could be restored by complementation [188,226], indicating that ClosTron mutagenesis is not responsible for elevated toxin levels.

Vector-based complementation of a mutant gene is a powerful tool to prove that the mutated gene is responsible for an observed phenotype. However, in the case of proteomic analyses such as ours, the introduction of the vector for complementation and the selection of this vector using antibiotics had a profound effect on the proteome in our experiments (data not shown). The same was true for vector-based inducible (over)expression constructs, that necessitate an additional molecule for the induction of expression and thereby introduce more variation in the experimental setup. Because of the arguments presented above and since we did not observe any phenotypic changes in the *cd1597*::CT strain that could potentially be restored to a WT phenotype, we did not include a complemented strain in our quantitative proteomics experiments.

The only differentially expressed protein during both the exponential and stationary phases was CD3275, a putative phosphosugar isomerase. Phosphosugar isomerases bind phosphosugars and function as an isomerase. However, the substrate and product are unknown, but the *cd3275* gene is located in a predicted PTS operon for mannose, fructose, or sorbose transport and phosphorylation. The other proteins that were more abundant in the *cd1597*::CT strain during the stationary phase were involved in sporulation. Possibly, CD3275 functions as a link between CD1597 and the process of sporulation, and the result of this interplay is only observed during the stationary phase.

Quantitative proteomics analysis of whole cell cultures showed an increase in sporulation proteins during the stationary phase, but no changes were observed in the sporulation frequency of the *cd1597*::CT strain. And, although the effect of the absence of CD1597 was most profound in vegetative cells during the stationary phase, CD1597 was exclusively identified in spores. Furthermore, our proteomic analysis identified only minor, non-significant changes in the spore proteome of the mutant spores. Collectively, our results did not aid us in predicting a role for CD1597. Possibly, CD1597 functions exclusively when the bacteria encounter specific environments or stimuli that were absent in our experiments. Based on the presence of CD1597 in spores, we could expect a role in protecting these spores against environmental challenges or a role in germination. A role in germination is not unthinkable, since other (pseudo)proteases are involved in this process [227,228]. Alternatively, the presence of CD1597 in spores could be a remnant of the sporogenesis, meaning that CD1597 has already fulfilled its role upon completion of the spore. Further investigations into the possibility of proteolytic activity or spore resistance and germination might shed more light on the function of the enigmatic protein CD1597.

Experimental procedures

Growth of bacterial strains and culture media

The strains used in this study are listed in **Table 2**. *E. coli* strains were cultured in LB broth (Sigma) or LB agar plates supplemented with 25 µg/ml chloramphenicol, 50 µg/ml ampicillin, or 50 µg/ml kanamycin when required. *C. difficile* was grown in pre-reduced BHIY (37 g/L brain heart infusion [Oxoid]) supplemented with 5g/L yeast extract [Sigma]) or on pre-reduced BHIY agar plates (BHIY supplemented with 15 g/L agar [Alfa Aesar]). Alternatively, YT (8 g/L tryptone, 5 g/L yeast, 2.5 g/L NaCl), CDMM [210], or SMC [229] medium was used. *C. difficile* was cultured anaerobically in a Don Whitley VA1000 or A45 workstation (10% CO_2, 5% H_2 and 85% N_2 atmosphere) at 37 °C. Media were supplemented with 15 µg/ml thiamphenicol when required.

Purification of recombinant CD1597 and the CD1597 catalytic domain

PPEP-1 was expressed and purified as previously described [146,230]. For the expression of CD1597 and its catalytic domain, pET28a vectors containing *E. coli* codon-optimized 6xHis-CD1597 (pBC027) and 6xHis-CD1597(AA 211-416) (pBC029) constructs were ordered from Twist Bioscience. The expression vectors were transformed to *E. coli* strain C43 and protein expression was induced using 1 mM IPTG for 4 h at 37 °C. Lysates were prepared as described in the protocol for preparation of cleared *E. coli* lysates under native conditions as described in the fifth edition of the QIAexpressionist (Qiagen). The lysates were loaded onto a 1 ml HisTrap HP column (GE Healthcare) coupled to an ÄKTA Pure FPLC system (GE Healthcare). The column was washed using wash buffer (50 mM NaH2PO4, 300 mM NaCl, 20 mM imidazole) and 6xHis-CD1597 and 6xHis-CD1597(AA211-416) were eluted using a step gradient with elution buffer (50 mM NaH2PO4, 300 mM NaCl, 250 mM imidazole). Buffer was changed to PBS pH 7.4 using an Amicon 4ml 3k centrifugal filter. Glycerol was added to a final concentration of 10% before storage at -80 °C.

BODIPY TR-X casein cleavage assay

Red-fluorescent BODIPY TR-X casein (EnzChek™ Protease Assay Kit, Molecular Probes) was used to detect cleavage by CD1597, the CD1597 catalytic domain, PPEP-1, and Trypsin. A reaction mixture of 200 µl contained the provided reaction buffer (final concentration 1x), 1 µg enzyme, and 1 µg BODIPY TR-X casein. Fluorescence was measured every minute for 60 min in an Envision 2105 Multimode Plate Reader using excitation and emission wavelengths of 590 nm and 619 nm, respectively.

FRET-quenched peptide cleavage assays

Time course kinetic experiments with PPEPs were performed using fluorescent FRET-quenched peptides. FRET peptides consisted of Lys(Dabcyl)-EXXPPXXD-Glu(EDANS), in which X varied between the different peptides tested. Proteolysis of FRET peptides by PPEPs was tested in 150 µl PBS containing 50 mM FRET peptide and 2 µg of CD1597 or 500 ng for PPEP-1. Peptide cleavage was measured using the Envision 2105 Multimode Plate Reader. Fluorescence intensity was measured each minute for 1 h, with 10 flashes per measurement. The excitation and emission wavelengths were 350 nm and 510 nm, respectively.

Growth curves of *C. difficile cd1597*::CT

C. difficile 630Δ*erm*, *cd1597*::CT, and *tcdC*::CT were grown overnight in BHIY, YT or CDMM medium. For the BHIY growth curves, overnight cultures in BHIY with an OD_{600} <0.9 were used to inoculate fresh BHIY to an OD_{600} of 0.05. For the YT and CDMM growth curves, new pre-cultures were prepared by diluting overnight cultures in their respective media to an OD_{600} of 0.05. These fresh pre-cultures were grown for 2 h to ensure cells were in the exponential growth phase and used to inoculate fresh media to an OD_{600} of 0.05. OD_{600} was measured using an Implen™ OD600 DiluPhotometer™ at different time points.

Generation of a *cd1597*::ClosTron mutant

The ClosTron mutagenesis was performed as previously described [231]. In short, the re-targeting primers for ClosTron mutagenesis of CD1597 were designed using the intron design tool available at http://clostron.com/ and are shown in **Table 3**. The re-targeted intron was produced by PCR (with primers CD1597-402/403-IBS, CD1597-402/403-EBS1d, CD1597-402/403-EBS2, EBS universal primer, see Table 2 for sequences), purified from an agarose gel and inserted in pCR2.1 using a TOPO™ TA Cloning™ Kit (Invitrogen). The re-targeted intron was excised from pCR2.1 using HindIII and BsrGI and ligated into the pMTL007 vector backbone [209] that was digested using the same restriction enzymes. The resulting plasmid was transformed to *E. coli* CA434 and subsequently conjugated to *C. difficile* (see *Conjugation of plasmids to C. difficile*). Conjugants were directly streaked on BHIY plates supplemented with 20 µg/ml lincomycin. Resistant colonies were grown in BHIY and gDNA was isolated for PCR with primers oJC147 and oJC148 to test for the spliced *ermB* retrotransposition-activated marker (RAM) intron marker. In addition, to test for the correct insertion site, a PCR was performed with primers oBC084 and oBC085 (targeting the predicted insertion site).

Insertion of the *ermB* RAM was also confirmed by Sanger sequencing and whole genome sequencing (WGS). The resulting *cd1597*::CT strain (BC057) was stored at -80 °C.

Conjugation of plasmids to *C. difficile*

Conjugation procedures were as described previously [232]. In summary, the desired plasmid was introduced into the *E. coli* strain CA434 by transformation, and transformants were selected on LB plates supplemented with 20 µg/mL chloramphenicol. A single colony was grown overnight in LB medium with 20 µg/mL chloramphenicol. Subsequently, a 1 mL culture pellet from the transformed *E. coli* CA434 was transferred into the anaerobic chamber and combined with 200 µL of an overnight culture of *C. difficile*.

Droplets of this mixture were plated on a BHIY yeast plate and incubated for over 6 hours at 37 °C under anaerobic conditions. Following incubation, the bacteria were scraped from the plate with anaerobic PBS, and dilutions were plated on BHIY plates containing 15 µg/mL thiamphenicol and *C. difficile* selective supplement (Oxoid). Colonies were subjected to three consecutive passages on BHIY plates supplemented with thiamphenicol. After the last passage, the species and the presence of the plasmid were confirmed by PCR. Primers oWKS1070/oWKS0171 (targeting *C. difficile gluD*) and oWKS1387/oWKS1388 (targeting *traJ*, located on the plasmid), were used for this purpose.

All plasmids used in this study are mentioned in **Table 4**.

Promoter activity assay

Overnight cultures of the 630Δ*erm* strain carrying no plasmid, a P*cd1597*-sLuc^{opt} construct (BC040), or a P*ppep1*-sLuc^{opt} construct (JC178) were pelleted, washed once in CDMM, and resuspended in CDMM before inoculating BHIY, YT, or CDMM medium to an OD_{600} of 0.05. Samples were taken at different time points while measuring the OD_{600}. Samples were diluted 1:100 in BHIY and 90 µl of the diluted samples was transferred to white, flat bottom 96-wells plates. To each well, 20 µl reconstituted Nano-Glo substrate (50-fold diluted Nano-Glo substrate in kit buffer, Promega) was added and the plate was incubated for 10 min. Subsequently, relative light units (RLU) were measured in a GloMax® Explorer Multimode Microplate Reader (Promega) using standard settings. The RLUs were corrected for biomass by dividing the RLUs by the OD_{600} value of the bacterial culture at the time of sampling.

Sample preparation for overall comparative proteomics

For the overall comparative proteomic experiment using whole cell cultures of *C. difficile*, three biological replicates for *C. difficile* strain 630Δ*erm* and the *cd1597*::CT strain were included. For the *tcdC*::CT strain, two biological replicates were used. Since we looked at both the mid-logarithmic and stationary phases in a single experiment, this amounted to a total of 16 samples that were used in a TMTpro 16plex experiment. Sample preparation was performed as described previously [233]. Single colonies of *C. difficile* strain 630Δ*erm*, *cd1597*::CT, and *tcdC*::CT were picked and precultured overnight in BHIY. The precultures were used to inoculate fresh BHIY at a starting OD_{600} of 0.05, and cells were grown to an OD_{600} of 0.8 before harvesting half of the cells. The remaining cells were grown for 22 h in total. Cells were pelleted by centrifugation (6000 x g, 10 min, 4 °C). Pellets were resuspended in 10 mL of ice-cold PBS and washed twice (6000 x g, 10 min, 4 °C). After the last wash, pellets were resuspended in 5 mL urea lysis buffer (8 M urea, 50 mM Tris-HCl pH 7.5, 1x cOmplete protease inhibitor cocktail EDTA free). Resuspended cells were incubated for 20 min on ice prior to lysis by sonication, and cells were subsequently lysed by sonication for five bursts of 30 s with cooling on ice in between rounds. After lysis, tubes were centrifuged (15 min, 15000 x g, 4 °). Supernatants were transferred to new tubes and stored at −20 °C until further use.

For each strain, 100 µg of protein in 100 µL of ST buffer was used as the starting material. Proteins were reduced using 5 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) for 30 min, alkylated with 10 mM iodoacetamide for 30 min, and quenched with 10 mM Dithiothreitol (DTT) for 15 min, all at room temperature. Proteins were precipitated by chloroform-methanol precipitation. For this, 400 µL methanol, 100 µL chloroform, and 300 µL dH2O were added with vortexing in between each step. Following centrifugation (21130 x g, 2 min, RT), the pellet was washed two times with 500 µL methanol. The protein pellet was subsequently resuspended in 100 µL of 40 mM HEPES pH 8.4 containing 4 µg trypsin and incubated overnight at 37 °C. Then, another 4 µg of trypsin was added and incubated for 3 h at 37 °C.

For the overall comparative proteomic experiments using only spores, five biological replicates were used per strain in a TMTpro 15plex experiment. Single colonies of *C. difficile* strain 630Δ*erm*, *cd1597*::CT, and *tcdC*::CT were picked and cultured overnight in SMC medium. From each overnight culture, 400 µl was spread on two large SMC agar plates (⌀ 14.5 cm) for confluent growth. Spores were allowed to develop for 7 days. All cell material was resuspended in 3 ml sterile dH_2O by scraping and transferred to tubes and pelleted by centrifugation (3220 x g, 10 min, 4 °C). Pellets were washed three times with 10 ml dH_2O and ultimately resuspended in 10 ml dH_2O and stored for 4 days at 4 °C. Then, tubes were centrifuged (15000 x g, 10 min, 4 °C) and the supernatant was removed. Pellets were resuspended in 1.5 ml 20% Gastrografin (Bayer). The cell

suspension was carefully layered on top of 10 ml 50% Gastrografin and tubes were centrifuged (10000 x g, 30 min, RT) to separate the spores from the vegetative cells. The supernatant was removed and the spore pellets were resuspended in 500 µl dH_2O. Spores were washed three times in dH_2O before storage at -20 °C until further use. Spores were resuspended in 50 µl extraction buffer (4% SDS, 10% 2-mercaptoethanol, 1 mM DTT, 125 mM Tris-HCL pH 6.8, 10% glycerol) and proteins were extracted by incubation for 10 min in a heat block set to 108 °C. After centrifugation (15000 x g, 2 min, RT) supernatant was transferred to a new tube. Again, 50 µl extraction buffer was added to the pellet and the process was repeated, resulting in a total volume of 100 µl extracted spore proteins.

The total spore protein material per strain (<100 µg) was used as the starting material. A chloroform-methanol precipitation was performed and the resulting protein pellet was resuspended in urea lysis buffer. Proteins were reduced, alkylated, and quenched as described above. After chloroform-methanol precipitation, 2 µg LysC was added and the mixture was incubated for 2 h at 37 °C. Then, 4 µg trypsin was added and the tubes were incubated overnight at 37 °C before adding another 4 µg of trypsin and incubation for 3 h.

TMT labeling was performed as described previously [233] on 10 µg of tryptic peptides using TMTpro 16plex labeling (Thermo Fisher Scientific, lot no. UK292954 for whole cultures, WK334339 for spores) for 1 h at RT. Excess TMT label was quenched with 5% hydroxylamine for 15 min at RT. The labeled peptides from each sample were mixed and freeze-dried. The peptides were resuspended in 10 mM ammonium bicarbonate pH 8.4 and separated in 12 fractions on an Agilent Eclipse Plus C18 column (2.1 × 150 mm, 3.5 µM). Half of the labeled peptides (80 µg) were injected. Mobile phase A was 10 mM ammonium bicarbonate (pH 8.4). Mobile phase B was 10 mM ammonium bicarbonate in 80% acetonitrile (pH 8.4). The gradient was as follows: 2% B, 0–5 min; 2%–90% B, 5–35 min; 90% B, 35–40 min; 90%–2% B, 40–41 min; and 2% B, 41–65 min. The 12 collection vials were rotated every 30 s during sample collection. The 12 fractions were freeze-dried and stored at −20 °C prior to LC-MS/MS analysis.

LC-MS/MS analyses

The fractions were analyzed as described previously [230] with minor adjustments by online C18 nanoHPLC MS/MS with an Ultimate3000nano gradient HPLC system (Thermo, Bremen, Germany), and an Orbitrap Exploris 480 mass spectrometer (Thermo). Peptides were injected onto a pre-column (300 µm × 5 mm, C18 PepMap, 5 µm, 100 A), and eluted via a homemade analytical nano-HPLC column (30 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm, 120 A; Dr. Maisch, Ammerbuch, Germany). The gradient was run with a gradient of

2% to 40% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 142 min. The nano-HPLC column was drawn to a tip of ~10 µm and acted as the electrospray needle of the MS source. The mass spectrometer was operated in data-dependent MS/MS mode for a cycle time of 3 s, with HCD collision energies at 36V and recording of the MS2 spectrum in the Orbitrap, with a quadrupole isolation width of 1.2 m/z. In the master scan (MS1) the resolution was 120,000, the scan range 400-1500, at standard AGC target at a maximum fill time of 50 ms. A lock mass correction on the background ion m/z = 445.12003 was used. Precursors were dynamically excluded after n = 1 with an exclusion duration of 45 s and with a precursor range of 10 ppm. Charge states 2–5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 45,000 at an AGC target of 200% @maximum fill time of 60 ms.

LC-MS/MS data analysis

Data analysis was performed as described previously [233]. In the post-analysis process, raw data were converted to peak lists using Proteome Discoverer version 2.4.0.305 (for analysis of whole cell cultures) and 2.5.0.400 (for analysis of spores) (Thermo Electron) and submitted to the UniProt *C. difficile* 630Δ*erm* database (3752 entries) (Taxon ID: 272563) using Mascot v. 2.2.07 (www.matrixscience.com) for peptide identification. Mascot searches were performed with 10 ppm and 0.02 Da deviation for precursor and fragment mass, respectively, and trypsin was selected as enzyme specificity with a maximum of two missed cleavages. The variable modifications included Oxidation (M) and Acetyl (protein N-term). The static modifications included TMTpro (N-term, K) and Carbamidomethyl (C). Peptides with an FDR < 1% based on Percolator [196] were accepted.

Sporulation frequencies of strains *cd1597*::CT, 630Δ*erm* and *spo0A*::CT

For the strains *cd1597*::CT and 630Δ*erm,* three individual colonies were used to inoculate 10 ml BHIY supplemented with 0.1% taurocholate and 0.2% fructose. As a negative control, a single replicate for the *spo0A*::CT strain was included. From these cultures, a 10x dilution series was made until a dilution of 10^7 was reached. Cells were grown overnight. Exponential cultures (OD_{600} < 0.9) were diluted to OD_{600} = 0.5 in fresh BHIY. Next, 250 µl of this suspension was plated on 70:30 sporulation agar plates [234] and cells were grown for 24 h. Approximately $1/8^{th}$ of the cells were scraped off the plate and resuspended in 5 ml BHIY. A 5 µl sample was pipetted onto a 1% agarose slab for microscopy. Cells were analyzed using a Leica DMB6 phase-contrast microscope. Sporulation frequencies were determined by counting the presence of (developing)

spores in the *cd1597*::CT (n=7963) and 630Δ*erm* (n=5185) cells. For the *spo0A*::CT strain, no sporulation was observed.

Localization and overexpression of CD1597

A plasmid was constructed for the inducible expression of CD1597-CFP. For this, *cd1597* was amplified from *C. difficile 630Δerm* gDNA using primers oBC090 and oBC091. The PCR product was inserted in pCR2.1 using a TOPO™ TA Cloning™ Kit (Invitrogen). The resulting plasmid (pBC044) was digested using XhoI and SacI and the insert was ligated into the XhoI and SacI digested pRD2 backbone [235]. The resulting CD1597-CFP^{opt} plasmid (pBC053) was transformed to *E. coli* CA434 before conjugation to *C. difficile* 630Δ*erm*, producing strain BC103. As controls, a strain for the expression of CFP^{opt} (WKS1734, harboring pHEW91), only CD1597 (BC062, harboring pBC033, plasmid ordered from ATUM), and *C. difficile* 630Δ*erm* were used.

For the analysis using phase-contrast and fluorescent microscopy, the strains were precultured overnight in BHIY. Fresh BHIY was inoculated to an OD_{600} of 0.1 and the culture was grown for 4 h before inducing expression (only for BC103 and WKS1734) with 0, 25, or 200 ng/ml anhydrotetracycline (ATC). Cells were imaged 4, 24, and 48 h after induction. At each time point, 1 ml culture was taken and centrifuged for 2 min at 6000 x g. Supernatant was removed and cells were resuspended in 100 µl PBS. Five µl of each suspension was pipetted onto 1% agarose slabs on microscopy slides. Cells were imaged with a Leica DM6000 DM6B fluorescence microscope (Leica) equipped with a DFC9000 GT sCMOS camera using an HC PLAN APO 100x/1.4 OIL PH3 objective, using the LAS X software (Leica). A Leica filter set for CFP (11504163) was used.

Table 2. Strains used in this study

Strain	Organism	Genotype	Resistance	Reference
DH5α	*E. coli*	DH5α		
CA434	*E. coli*	CA434		
BC001	*C. difficile*	630Δ*erm*		
BC040	*C. difficile*	630Δ*erm*, P*cd1597-sLuc*^{opt} (pBC025)	Cam	This study
BC044	*E. coli*	DH5α, pBC062	Amp	This study
BC046	*E. coli*	CA434, pBC063	Cam	This study
BC057	*C. difficile*	630Δ*erm*, *cd1597*::CT	Erm	This study
BC062	*C. difficile*	630Δ*erm*, P*tet-cd1597* (pBC033)	Cam	This study
BC103	*C. difficile*	630Δ*erm*, P*tet-cd1597-cfp*^{opt} (pBC053)	Cam	This study
JC178	*C. difficile*	630Δ*erm*, P*ppep1-sLuc*^{opt} (pJC084)	Cam	This study
JC284	*C. difficile*	630Δ*erm*, *tcdC*::CT	Erm	[59]
JC336	*C. difficile*	630Δ*erm*, *spo0A*::CT	Erm	[182]
WKS1734	*C. difficile*	630Δ*erm*, P*tet-cfp*^{opt} (pHEW91)	Cam	[235]

Table 3. Primers used in this study

Name	Sequence (5’ to 3’)
CD1597-402/403-IBS	AAAAAAGCTTATAATTATCCTTAAACATCGAAGACGTGCGCCCAGATAGGGTG
CD1597-402/403-EBS1d	CAGATTGTACAAATGTGGTGATAACAGATAAGTCGAAGACTCTAACTTACCT TTCTTTGT
CD1597-402/403-EBS2	TGAACGCAAGTTTCTAATTTCGGTTATGTTCCGATAGAGGAAAGTGTCT
EBS universal primer	CGAAATTAGAAACTTGCGTTCAGTAAAC
oJC147	ACGCGTTATATTGATAAAAATAATAATAGTGGG
oJC148	ACGCGTGCGACTCATAGAATTATTTCCTCCCG
oBC084	GTGGATTTTCTTTTGCTTTTATATCATTGC
oBC085	GATGAGATTTATATAGACTTAAAACAAGCG
oBC090	AAAGAGCTCATTTGAATTTTTTAGGGGGAAAATACCATGGAAAACAATTTAA ATACAGCT
oBC091	AAACTCGAGACTTCCTGAACCAGATCCTGAATAGTTTAGTTCAAGTTTTTCAA GAAAATC
oWKS1070	GTCTTGGATGGTTGATGAGTAC
oWKS1071	TTCCTAATTTAGCAGCAGCTTC
oWKS1387	CAGATGAGGGCAAGCGGATG
oWKS1388	CGTCGGTGAGCCAGAGTTTC

Table 4. Plasmids used in this study

Name	Backbone	Insert	Resistance	Purpose	Reference
pCR2.1	-	-	Amp	TOPO TA cloning of re-targeted intron *cd1597* and *cd1597* for CFP fusion	-
pMTL007	-	-	Cam	Provides backbone for *cd1597* intron	[209]
pBC025	pAP24 [212]	P*cd1597-sLuc*^{opt}	Cam	Promoter activity assay	This study
pBC027	pET28a	*cd1597* (*E. coli* codon-optimized)	Kan	Expression of CD1597 for purification	This study
pBC029	pET28a	*cd1597*(AA 211-416) (*E. coli* codon-optimized)	Kan	Expression of the CD1597 catalytic domain for purification	This study
pBC033	pRPF185	P*tet-cd1597*	Cam	Expression of CD1597	This study
pBC044	pCR2.1	*cd1597*	Amp	Amplification of *cd1597* for CFP fusion	This study
pBC053	pRPF185	P*tet-cd1597-cfp*^{opt}	Cam	Expression of CD1597-CFP^{opt}	This study
pBC062	pCR2.1	Re-targeted intron *cd1597*	Amp	Amplification of *cd1597* intron for further cloning	This study
pBC063	pMTL007	Re-targeted intron *cd1597*	Cam	Mutagenesis of *cd1597*	This study
pJC084	pAP24 [212]	P*ppep1-sLuc*^{opt}	Cam	Promoter activity assay	This study
pRD2	pRPF185	P*tet-hupA-cfp*^{opt}	Cam	Backbone for CD1597-CFP^{opt}	[235]
pHEW91	pRPF185	P*tet-cfp*^{opt}	Cam	Expression of CFP^{opt}	[235]

Acknowledgements

We thank Ulrich Baumann and Fabian Wojtalla for supplying the BODIPY TR-X casein substrate kit.

Data summary

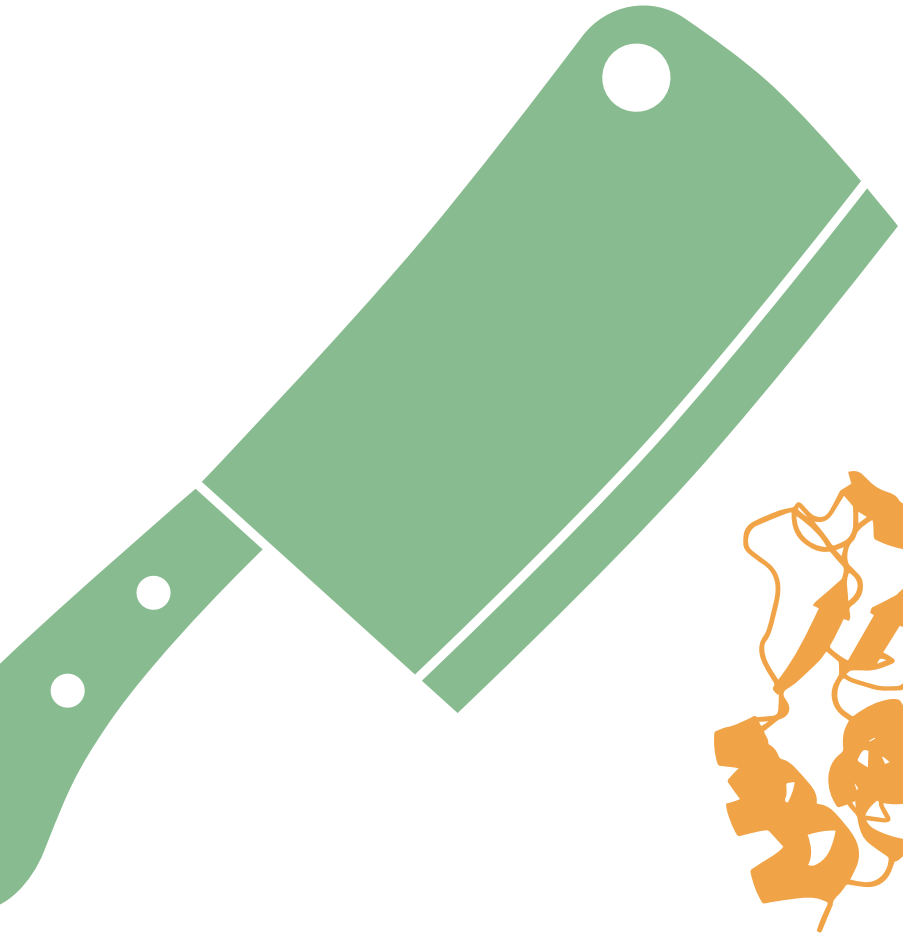
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [201] partner repository with the dataset identifier PXD052347.

Funding Information

This research was supported by an ENW-M Grant (OCENW.KLEIN.103) from the Dutch Research Council (NWO) and by the research program Investment Grant NWO Medium with project number 91116004, which is (partially) financed by ZonMw.

Supporting information

The supplemental Tables S1-S3 can be found online: https://www.microbiologyresearch.org/content/journal/acmi/10.1099/acmi.0.000855.v1#supplementary_data



4

In-depth specificity profiling of endopeptidases using dedicated mix-and-split synthetic peptide libraries and mass spectrometry

Bart Claushuis[1], Robert A. Cordfunke[2], Arnoud H. de Ru[1], Annemarie Otte[1], Hans C. van Leeuwen[3], Oleg I. Klychnikov[4], Peter A. van Veelen[1], Jeroen Corver[5], Jan W. Drijfhout[2], Paul J. Hensbergen[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands
[2] Department of Immunology, Leiden University Medical Center, Leiden, The Netherlands
[3] Department of CBRN Protection, Netherlands Organization for Applied Scientific Research TNO, Rijswijk, The Netherlands
[4] Department of Biochemistry, Moscow State University, Moscow, Russian Federation
[5] Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands

Published in Analytical Chemistry, 2023, 95, 31, 11621–11631

DOI: 10.1021/acs.analchem.3c01215

Abstract

Proteases comprise the class of enzymes that catalyze the hydrolysis of peptide bonds, thereby playing a pivotal role in many aspects of life. The amino acids surrounding the scissile bond determine the susceptibility towards protease-mediated hydrolysis. A detailed understanding of the cleavage specificity of a protease can lead to the identification of its endogenous substrates, while it is also essential for the design of inhibitors. Although many methods for protease activity and specificity profiling exist, none of these combines the advantages of combinatorial synthetic libraries, i.e. high diversity, equimolar concentration, custom design regarding peptide length and randomization, with the sensitivity and detection power of mass spectrometry. Here, we developed such a method and applied it to study a group of bacterial metalloproteases that have the unique specificity to cleave between two prolines, i.e. Pro-Pro endopeptidases (PPEPs). We not only confirmed the prime-side specificity of PPEP-1 and PPEP-2, but also revealed some new unexpected peptide substrates. Moreover, we have characterized a new PPEP (PPEP-3) which has a prime-side specificity that is very different from that of the other two PPEPs. Importantly, the approach that we present in this study is generic and can be extended to investigate the specificity of other proteases.

Introduction

Proteases comprise the class of enzymes that catalyze the hydrolysis of peptide bonds between amino acids in a polypeptide chain. Through cleavage of their substrates, proteases play a pivotal role in many aspects of life, ranging from viral polyprotein processing [236] to a wide range of human physiological and cellular processes, e.g. hemostasis, apoptosis and immune responses [237–239] Uncovering the endogenous substrate(s) is usually a key step towards dissecting the biological role of a protease. However, it is not straightforward to identify protease substrates without prior knowledge, e.g. without a clear phenotype in a protease knockout or lack of information from homologs in other species. Information about the cleavage specificity of a protease can aid in the identification of endogenous substrates. Moreover, such information is pivotal for inhibitor design or the development of diagnostic biomarker assays [240–242].

We study a group of bacterial proteases that have the unique specificity to cleave a peptide bond between two prolines, i.e. Pro-Pro endopeptidases (PPEPs). The first two members, PPEP-1 from the human pathogen *Clostridioides difficile* [160,243] and PPEP-2 from *Paenibacillus alvei* [157], are secreted enzymes which cleave cell surface proteins involved in bacterial adhesion. Initially, the specificity of PPEP-1 was determined based on a small synthetic peptide library that was designed based on the identification of a sub-optimal cleavage site in a human protein [146]. Following the elucidation of the endogenous PPEP-1 substrates, in which a total of 13 cleavage sites were found, a cleavage motif could be determined (**Figure 1A**). For PPEP-2, the endogenous cleavage site (**Figure 1A**) was experimentally determined following an *in silico* prediction of the substrate. This prediction was based on a similar genomic organization of the PPEP gene and its substrate in both *C. difficile* and *P. alvei*, i.e. they are adjacent genes (**Figure 1B**). Based on a bioinformatic analysis, we recently observed PPEP homologs in a wide variety of species [148], for example in *Geobacillus thermodenitrificans* (PPEP-3**, Figure 1**). The modeled structure of PPEP-3 shows a high degree of similarity with the crystal structures of PPEP-1 and PPEP-2 (**Figure 1C**). However, none of the genes adjacent to *ppep-3* encode a protein which contains a PPEP consensus cleavage motif (XXPPXP, **Figure 1A&B**), hampering the formulation of a testable hypothesis about its substrate(s). Hence, to gain insight in the activity and specificity of hitherto uncharacterized putative PPEPs, a general method to profile their specificity is needed.

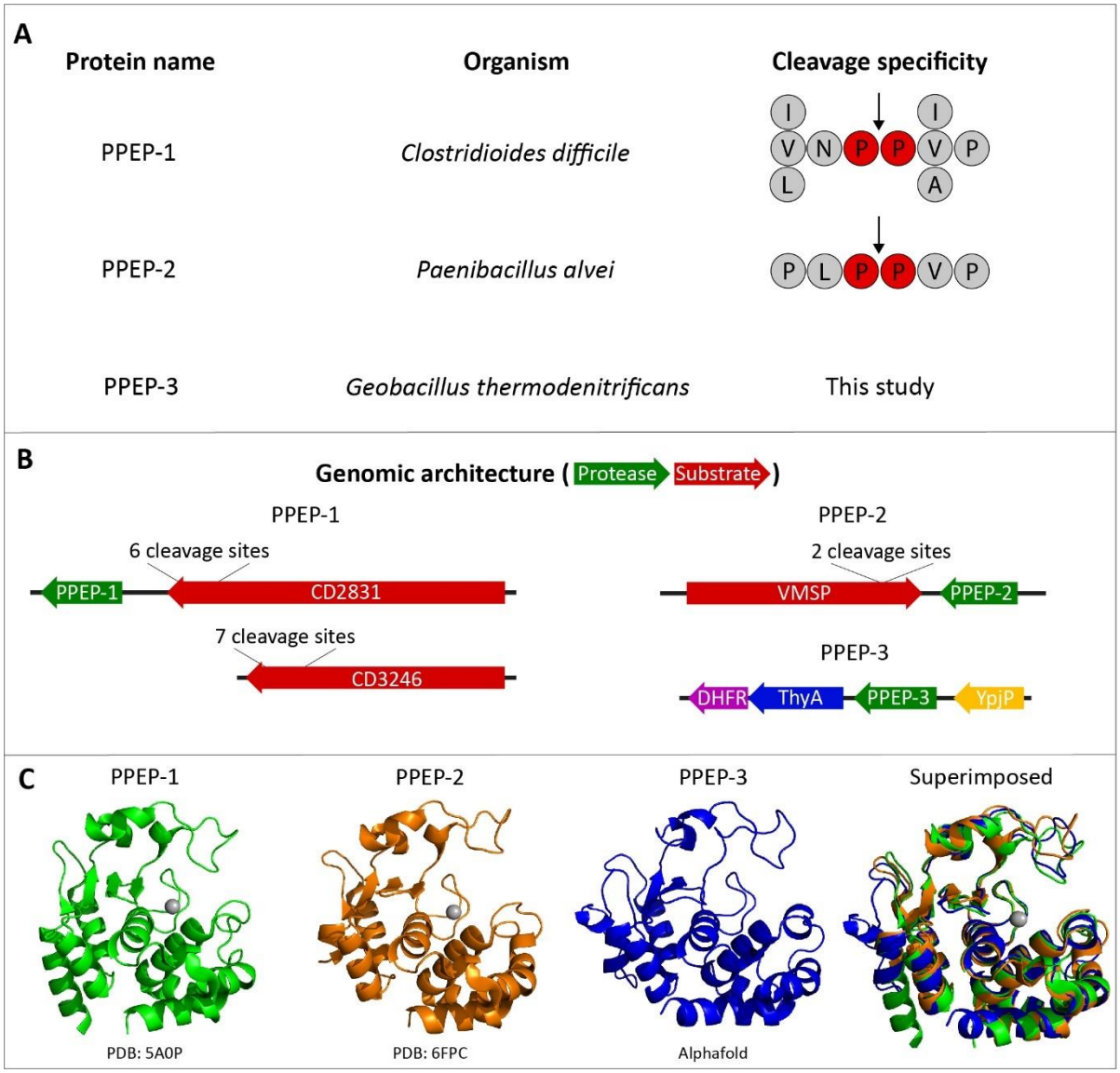


Figure 1. Overview of the PPEPs used in this study. A) The three PPEPs that are used in this study and their respective origins and substrate specificity. For PPEP-1 and 2 the cleavage specificity is based on the endogenous substrates. For PPEP-3, no substrates have been described yet. **B)** The genomic architecture of the PPEPs and their substrates. For PPEP-1, the gene encoding the substrate CD2831 is adjacent to PPEP-1. The gene encoding the second substrate (CD3246) is positioned elsewhere on the genome. The genes for PPEP-2 and its substrate VMSP are also located adjacent to each other. For PPEP-3, no adjacent genes contain the consensus PPEP cleavage motif (i.e. PPXP). **C)** Crystal (PPEP-1 and PPEP-2) and predicted (PPEP-3) structures [157,160]. PPEP-3 structure was predicted using the Alphafold algorithm [199].

A wide variety of methods for protease activity and specificity profiling has been developed [241,244,245]. Several strategies rely on the identification of protease-generated protein neo-*N*-termini in cells expressing the protease of interest as compared to controls. For this purpose, positive and negative selection procedures for the enrichment of *N*-terminal peptides in combination with quantitative mass spectrometry based proteomics methods, collectively known as *N*-terminomics, have

been developed [166–168,246–248]. However, for an optimal experimental setup for such an experiment, a protease knockout cell line or strain is necessary.

Other strategies seek to identify the protease substrate specificity by making use of peptide libraries, either by phage display technologies [169,170] or as a collection of (synthetic) peptides. For the latter, mass spectrometry analysis is an attractive readout because it determines the signature proteolytic event in a highly specific manner, i.e. information on the amino acid(s) surrounding the scissile bond is obtained. For example, MALDI-based approaches using synthetic peptide arrays have been used to profile protease activity and specificity, but for such approaches, each peptide requires individual synthesis, treatment, and analysis [249,250]. In addition, proteome-derived peptide libraries have been shown to be a rich source of peptides for these types of analyses [251–253]. Although with this method a wide variety of potential substrates is tested in a single reaction, the concentration range of the peptides present may easily span a few orders of magnitude. This may complicate the assessment of whether a product peptide is derived from a very good substrate present at a low concentration or a poor substrate at a high concentration instead. Another method uses a small set of synthetic peptides in which amino acid pairs are cleverly positioned in order to contain a wide variety of potential cleavage sites [254]. However, this design was based on the assumption that for a protease only the correct positioning of two amino acids is necessary for a protease to cleave its substrate. Based on the inspection of the list of 228 peptides [255], we predict that none of these would be cleaved by one of the PPEPs, making MSP-MS not suitable for specificity profiling of PPEPs, and probably other proteases as well.

The combination of equimolar peptide concentrations with a high diversity would be the ideal scenario for the design of a peptide library. This can be achieved by constructing a synthetic combinatorial peptide library, for example using the one-bead-one-compound approach [256], and such libraries have been used to profile protease specificity [171,172]. As a read-out for cleavage of peptides, both fluorescence detection [171,240] and Edman degradation [257,258] have so far been used.

Given the beneficial characteristics of mass spectrometry mentioned above, we reasoned that it would be highly advantageous if this could be applied to analyze the product peptides following incubation of a combinatorial synthetic peptide library with a protease of interest in a single reaction, but this has hitherto not been done. Obviously, the complexity of combinatorial libraries tend to increase dramatically when multiple positions are randomized, thereby impeding MS analysis. Therefore, two aspects are pivotal to make such an approach suitable. First of all, in the design of the library, any prior knowledge or hypothesis about the protease specificity should be

utilized. Secondly, a strategy to enrich, analyze and identify the product peptides has to be implemented.

Therefore, the aim of the current study was to develop a novel method to study the activity and specificity of a protease, which combines the advantages of a combinatorial synthetic peptide library, i.e. high diversity and equimolar peptide concentrations, with the sensitivity and specificity of MS detection. Testing the method with PPEP-1 and PPEP-2 showed results that were in good agreement with previous data, while also some unexpected peptide substrates were observed. Importantly, the new method clearly established PPEP-3 as a genuine PPEP, but also showed that it has a markedly different prime-side specificity compared to PPEP-1 and PPEP-2.

Results and Discussion

Combinatorial peptide library design and experimental setup

Since PPEPs are defined by their ability to hydrolyze Pro-Pro bonds, and substrate specificity is further determined by positions P3-P3' surrounding the scissile bond [146,147,160], we constructed a combinatorial peptide library containing a XXPPXX motif. In this motif, the X positions represent any amino acid residue (with the exception of cysteine), while the core proline (P) residues (corresponding to the P1-P1' positions) are fixed (**Figure 2**).

In order to analyze product peptides after incubation of the library with a PPEP, the core sequence (XXPPXX) was modified in two ways. First, a six amino acid tail consisting of Gly-Gly-Leu-Glu-Glu-Phe (GGLEEF) was added at the C-terminus (**Figure 2**)**.** This sequence was chosen because PPEP cleavage between the two prolines would then provide retention of the C-terminal product peptides (PXXGGLEEF) on a C18-column. Moreover, the fragmentation pattern of such a peptide (PYVGGLEEF) that we observed in a previous study provided good sequence coverage of the N-terminal region (**Figure S1**). Second, a biotin was attached to the N-terminus of each peptide, connected to the rest of the peptide by a small linker (Ahx-Glu, Ahx=1-aminohexanoic acid, **Figure 2**). This allows for the enrichment of C-terminal product peptides by removal of biotinylated peptide molecules, i.e. non-cleaved peptides and N-terminal product peptides, using streptavidin beads. This is similar to a previously approach which used Edman degradation instead of mass spectrometry to sequence the protease generated product peptides [259]. In addition to the lower sensitivity of this method, several amino acids could not be accurately detected and information on subsite cooperativity [260] is lost.

Synthesis of the library was performed using the one-bead one-compound (OBOC) method [256] in order to achieve equimolar amounts of each unique peptide. Initially, we synthesized 19 sub-libraries for which the amino acid at the X corresponding to the P3 position (the first X in the sequence XXPPXX) was known. Each of these sub-libraries contains 6859 peptides (19x19x19). Since the process of linking biotin to the N-terminus is not 100% efficient, non-biotinylated peptides were also present. To remove these unwanted peptides prior to incubation with a PPEP, the library was pre-cleaned on an avidin column (**Figure 2**). The biotinylated peptide library that was obtained after elution from the avidin column was then incubated with a PPEP and subsequently depleted for biotinylated peptides using streptavidin. C-terminal, non-biotinylated, product peptides (PXXGGLEEF) were collected in the flow-through and analyzed by mass spectrometry. Peptide identification was accomplished using standard database searching (see Experimental section for details). Following this, the amino acids at the P2' and P3' positions were determined (**Figure 2**).

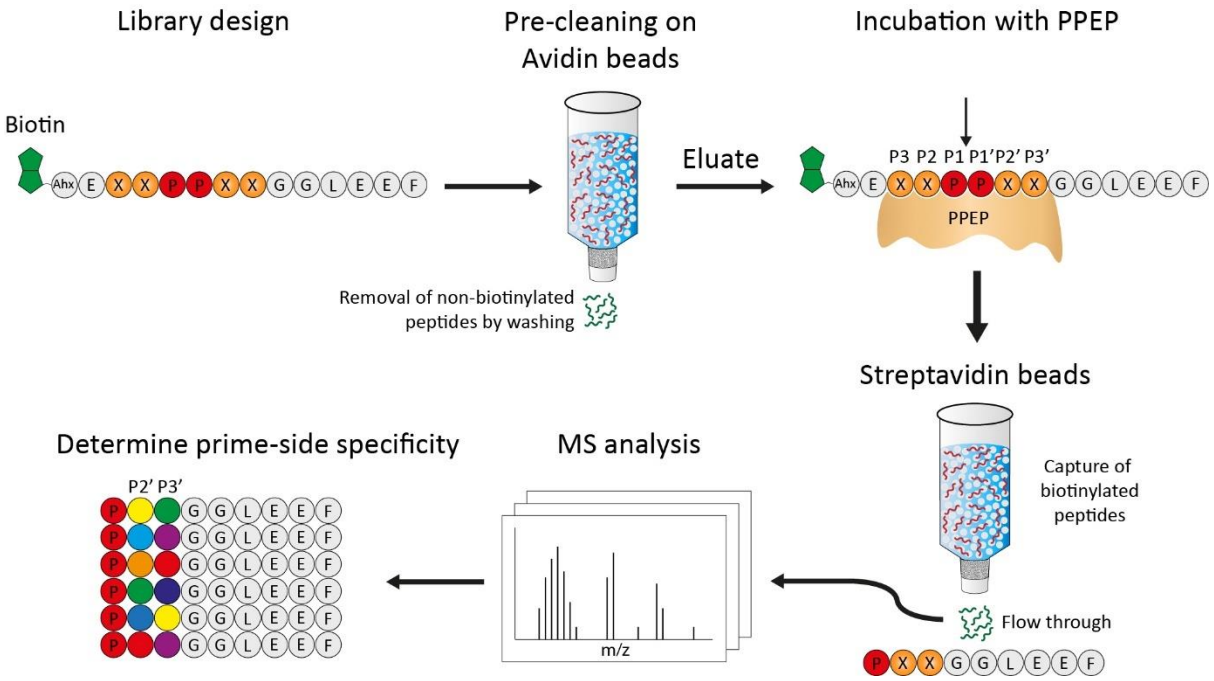


Figure 2. Design of the synthetic combinatorial peptide library and workflow to determine the activity and prime-side specificity of a Pro-Pro endopeptidase (PPEP). The library was designed to contain an XXPPXX motif, X representing any residue (X≠Cys). At the N-terminus, peptides were modified with a biotin, allowing removal of uncleaved peptides and N-terminal product peptides after incubation of the library with a protease, i.e. PPEP. At the C-terminus, a peptide tail (GGLEEF) was added in order for the C-terminal cleavage products to be compatible with LC-MS/MS analysis. This stretch of amino acids was also chosen based on a previously recorded MS/MS spectrum, showing favorable fragmentation characteristics (**Figure S1**). First, the library was pre-cleaned on avidin beads to remove non-biotinylated peptides. Then, the library was incubated with a PPEP. The scissile bond is indicated by the arrow. Following this, biotinylated peptides (non-cleaved peptides and N-terminal product peptides) were captured on a streptavidin column. The flow-through, containing non-biotinylated C-terminal product peptides (PXXGGLEEF) were then analyzed by LC-MS/MS, after which the prime-side specificity could be determined. Ahx: 1-aminohexanoic acid.

Incubation of PPEP-1 with two sub-libraries confirms the preference of PPEP-1 for valine over lysine at the P3 position

In our previous studies, we showed a preference of PPEP-1 for a Val as compared to a Lys at the P3 position [147]. Hence, to test the feasibility of our approach, two sub-libraries with either a Val or Lys at this position were incubated with PPEP-1. The formation of products due to proteolysis of substrate peptides present in the library was assessed using MALDI-FT-ICR MS (**Figure 3)**. As expected, product peptides were clearly visible when using the P3=Val library (**Figure 3**, upper panel), while these were not observed when the P3=Lys library was used instead (**Figure 3**, lower panel).

Although no fragmentation was performed, we could assign several product peptides when using the P3=Val library based on the accurate mass and our current

understanding of the specificity of PPEP-1 (**Figure 1**) [146,147], i.e. we were expecting PXPGGLEEF peptides. The highest signal was observed for the PPPGGLEEF peptide (*m/z* 942.459, $[M+H]^+$). Although three prolines at P1'-P3' are not found in the endogenous substrates (**Figure 1**), it had been demonstrated that PPEP-1 prefers all prolines at these positions [146]. In addition, a peptide matching with the product peptide PIPGGLEEF was observed, although based on the MALDI-FT-ICR MS analysis alone we cannot exclude the possibility that it corresponds to PPIGGLEEF, nor that it might contain a leucine instead of an isoleucine at the site corresponding to the P2'/P3' position. We also observed a peptide corresponding to PVPGGLEEF (or PPVGGLEEF). Even though the signal for this peptide partially overlapped with the second isotope peak of the PPPGGLEEF peptide (theoretical *m/z* value: 944.462, $[M+H]^+$), a separate peak for the signal at *m/z* 944.474 ($[M+H]^+$) was clearly visible. Lastly, a peptide was observed corresponding to either PHPGGLEEF or PPHGGLEEF even though it was hitherto unknown that PPEP-1 allows for a histidine at the P2' or P3' position.

Overall, the above results with the two combinatorial sub-libraries demonstrated the applicability of our approach to detect PPEP activity and study its preference for amino acids surrounding the scissile Pro-Pro bond.

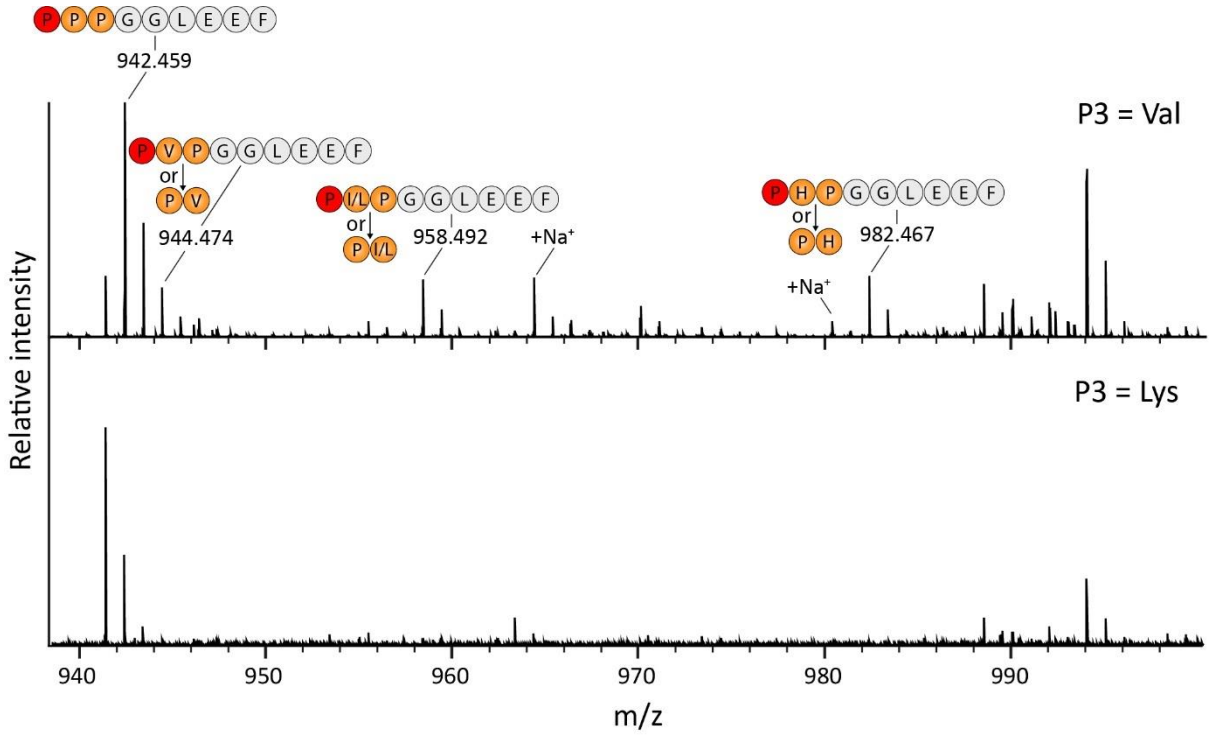


Figure 3. MALDI-FT-ICR MS analysis of PPEP-1 product peptides using two different combinatorial sub-libraries. The P3=Val and P3=Lys sub-libraries were incubated with PPEP-1 for 3 h. Following depletion of biotinylated peptides, non-biotinylated product peptides (PXXGGLEEF) were analyzed using MALDI-FT-ICR MS. The two indicated sodiated species are from the PPPGGLEEG and P(I/L)PGGLEEF/(PP(I/L)GGLEEF peptides, respectively.

PPEP-1, PPEP-2 and PPEP-3 display distinct substrate specificity after incubation with the full combinatorial peptide library

Following the successful tests of the method with the two sub-libraries and PPEP-1, we applied our method with the full combinatorial peptide library (a mix of all 19 sub-libraries, containing 130,321 peptides) to determine the prime-side substrate specificity of PPEP-1, PPEP-2 and PPEP-3. In order to increase the sensitivity and include fragmentation of the product peptides, samples were analyzed with LC-MS/MS. A non-treated sample was included as a control.

Initially, we analyzed the results by standard database searching against an in-house generated database (see Experimental Procedures for details). For PPEP-1 and PPEP-2 treated samples, the peptides with the highest intensities represented the expected PXXGGLEEF product peptides (**Table S1**). Moreover, an enrichment for prolines at the P2' and/or P3'positions was observed **(Table S1)**, in line with what was expected based on the specificity of PPEP-1 and PPEP-2 (**Figure 1**). For the PPEP-3 treated sample, the most highly abundant peptide was PPPGGLEEF. Hence, this clearly demonstrated that also PPEP-3 is an authentic PPEP. In addition, other 9-mer PXXGGLEEF product peptides were present among the most abundant peptides in the PPEP-3 treated sample (**Table S1**).

The results from the database search showed ambiguity in the position of the proline at the P2'/P3' position as assigned by the search algorithm (i.e. PXPGGLEEF or PPXGGLEEF). Also, several MS/MS spectra were matched with sequences that did not match with the expected 9-mer PXXGGLEEF sequence. For example, some MS/MS spectra were assigned to the 8-mer sequence KYGGLEEF. However, we argue that these represent wrong annotations due to the fact that the mass and elution time of this peptide is exactly the same as the PPPGGLEEF peptide, (one of) the highest product peptides observed for all three PPEPs (**Table S1**). Furthermore, in all cases that an isoleucine or leucine was present at the P2' or P3' position, obviously no distinction could be made by the search algorithm.

To substantiate our results, we combined manual inspection of the MS/MS spectra with additional LC-MS/MS analyses of a set of synthetic peptides. First of all, KYGGLEEF/YKGGLEEF peptides elute much earlier than the PPPGGLEEF peptide, and the fragmentation of such peptides is very distinct from PXXGGLEEF peptides, PPPGGLEEF in particular (**Figure S2**). Secondly, fragmentation spectra of PXPGGLEEF and PPXGGLEEF peptides showed clear differences (**Figure S3**). Importantly, spectra of PXPGGLEEF peptides are dominated by the unique PGGLEEF (y_7) fragment ion (*m/z* 748.351, **Figure S3**). This was for example essential in distinguishing PIPGGLEEF from PPIGGLEEF. The other unique fragment ion of PXPGGLEEF peptides, i.e. the b_2 corresponding to PX,

appeared less informative because it could also represent non-discriminatory internal fragments. We believe that this was one of the reasons why the results from the database searches were often ambiguous. Possibly other search algorithms, or training thereof, and new developments for prediction of tandem MS spectra [261] could aid in the correct assignment of product peptides in terms of the amino acids at the second and third position in the protease-generated product peptides.

In addition to peptide fragmentation characteristics, separation of isomeric peptides using our reversed-phase chromatography system as part of the LC-MS/MS system was also essential. For example, we observed that peptides with an isoleucine elute earlier than the isomeric peptide having a leucine (**Figure S4B,C**), in line with what is known about the relative contribution of these two residues to the retention on a reversed phase column [262]. Another way to discriminate between these two options is using a stable isotope labeled leucine/isoleucine during the synthesis of the library. PXPGGLEEF and PPXGGLEEF peptide pairs with an identical X residue that we have tested were well separated, with the exception of PIP and PPI (**Figure 5 & S4)**. For example, histidine containing peptides were separated depending on the position of the histidine within the peptide, as also observed previously [263].

Based on these additional analyses, we could refine the results from the database search and accurately assign the identity and abundance of the individual product peptides. Because, as opposed to proteome-derived peptide libraries [264,265], peptides in our library are present in equimolar concentrations, the relative abundance of the individual product peptides enabled us to obtain an estimate of how well specific amino acids are tolerated at the prime-sides (**Figure 4**). However, the difference in intensities between the signals of the individual product peptides in the MS data also relate to how well these peptides are ionized, especially when extra basic amino acids are present, i.e. histidine, arginine and lysine [266]. We believe that this could explain the relative high contribution of these amino acids to the prime-side cleavage motifs that we have obtained (**Figure 4**). Because most of the total intensity of the 9-mer product peptides could be explained by the 10 most abundant ones, we focused on these. Of note, since the proline at the P1' was fixed (**Figure 2**), no variation is observed at this position in **Figure 4**. We also observed longer peptides (**Table S1**) but given the large number of isomeric peptides, and the extra efforts needed to correctly assign the amino acid sequence for the PXXGGLEEF peptides as described above, we decided to not include these in the further analysis of the prime side specificity. Notwithstanding, they could potentially also provide some information about the P1' specificity when looking at the 11-mer peptides.

The prime-side residues of the endogenous substrates of PPEP-1 (**Figure 1**) were all represented among the top 10 product peptides, again demonstrating the feasibility of

our method. In addition, the preference of PPEP-1 to hydrolyze substrates with three prolines at the P1'-P3' (**Figure 3**) [146] was also demonstrated using the full combinatorial library (**Figure 4A**). Interestingly, our approach revealed several previously unknown prime-side options that allow for cleavage by PPEP-1. The most striking findings included the cleavage of substrates that had either PPH, PPA, or PPY at their P1'-P3' positions (**Figure 4A**), since the presence of a Pro residue at P3' was thought to be a determinant for proteolytic activity [146,160]. The requirement for a Pro residue at P3' was explained by the presence of a diverting loop in the co-crystal structure of PPEP-1 with a substrate peptide [160]. The Pro at P3' aligns with Trp-103 of PPEP-1 due to a parallel aliphatic-aromatic interaction, thereby redirecting the remainder of the substrate (P4' and onwards) out of the binding pocket by inducing a kink at the P2' position. Therefore, it was initially hypothesized that the PHPGGLEEF/PPHGGLEEF product observed using MALDI-FT-ICR MS (**Figure 3**) would in fact be PHPGGLEEF. However, manual inspection of the MS/MS fragmentation spectra revealed that PPEP-1 does tolerate PPH but not PHP at the P1'-P3' sites. To corroborate this finding, we synthesized two FRET-quenched peptides (Lys_{Dabcyl}-EVNPPHPD-Glu_{Edans} and Lys_{Dabcyl}-EVNPPPHD-Glu_{Edans}) and tested these with PPEP-1. As expected, based on our library results, PPEP-1 is able to hydrolyze a VNP↓PPH, but not a VNP↓PHP peptide (**Figure S5**). Notwithstanding these exceptions, an overall preference of PPEP-1 for a Pro at the P3' was observed **(Figure 4A**). The ability of PPEP-1 to hydrolyze substrates with His, Phe, and Tyr at P3' might be the result of aromatic-aromatic interactions (π-π stacking) with the Trp-103 and these residues [267]. In this scenario, a Pro residue at the P2' position is probably necessary to redirect the substrate from the diverting loop.

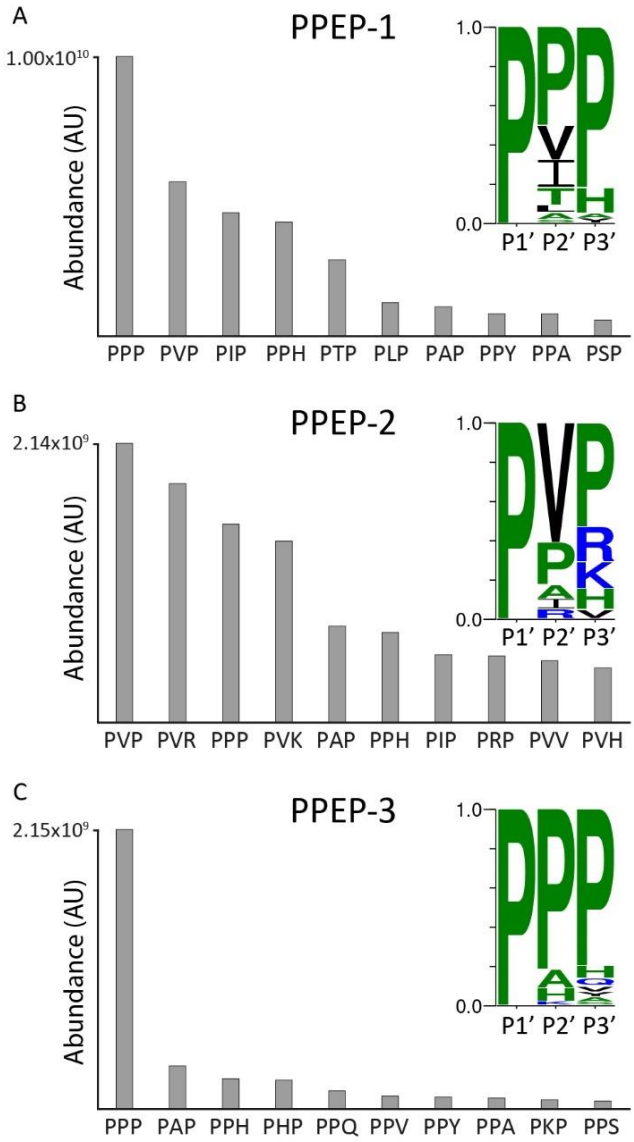


Figure 4. Top 10 most highly abundant 9-mer product peptides of PPEP-1, 2 and 3 reveal differences in prime-side specificity. The full combinatorial peptide library was incubated with recombinant PPEP-1, PPEP-2, or PPEP-3. Product peptides were analyzed using LC-MS/MS. Abundances were determined by summing the intensities of singly and doubly charged peptides. Discrimination between PXP and PPX peptides relied on both inspection of fragmentation spectra and C18 column separation (**Figures S3 & S4**). The 10 most highly abundant 9-mer product peptides formed by PPEP-1 (**A**), PPEP-2 (**B**) and PPEP-3 (**C**) and their abundances are represented as bars. A cleavage motif was constructed based on the relative intensities of the products peptides. The sequence on the X-axis represents the P1'-P3' residues of the PXXGGLEEF product peptides.

For PPEP-2, much less was known about the prime-side specificity because the initial identification of its cleavage site (PLPPVP) was based on the similarity in genomic organization of PPEP-1 and -2 and their endogenous substrates [157]. To a certain extent, PPEP-2 showed an overlapping specificity with PPEP-1 (**Figure 4B**). For example, a high level of the PPPGGLEEF peptide was found and PPEP-2 also allows PPH at the P1'-

P3’ positions. However, in line with the endogenous substrate (**Figure 1**), PPEP-2 prefers a valine at the P2’ (**Figure 4B**). Moreover, in contrast to PPEP-1, not all optimal substrates for PPEP-2 had at least two prolines at their P1’-P3’ positions. Of note, all peptides without prolines at the P2’ and P3’ positions had a Val at the P2’ position (**Figure 4B**), again indicating that this is a strong determinant for PPEP-2 susceptibility (**Figure 1**).

As mentioned above, we demonstrated for the first time that PPEP-3 is a genuine PPEP that cleaves Pro-Pro bonds (**Figure 4C**). For PPEP-3, the most abundant product peptide corresponded to PPPGGLEEF (**Figure 4C**)**.** Since this peptide was relatively much more abundant than peptides with other amino acids at the P2’ and P3’ positions, this resulted in an overall motif that was dominated by proline at the P1’-P3’ positions. Still, PPEP-3 allowed several other residues at the P3’ that were not tolerated by the other two PPEPs. Furthermore, unlike the other PPEPs, PPEP-3 was able to cleave a PPHP motif (P1-P3’), as represented by the PHPGGLEEF product peptide (**Figure 4C**).

Collectively, the above results showed that all three PPEPs preferred at least one proline at the P2’ or P3’ position. To emphasize the differences in such product peptides, extracted ion chromatograms (EIC) of every possible PXPGGLEEF/PPXGGLEEF peptide were constructed (**Figure 5**). Not only does this clearly show the difference in product profiles, it also reveals the differences between PXP and PPX peptides such as PHP and PPH.

To test the reproducibility of our method, we performed three additional replicate experiments with all three PPEPs. The results from these experiments show excellent reproducibility (**Figure S6**). Moreover, the overall profiles of the PXPGGLEEF/PPXGGLEEF peptides look very similar to the ones presented in **Figure 5**.

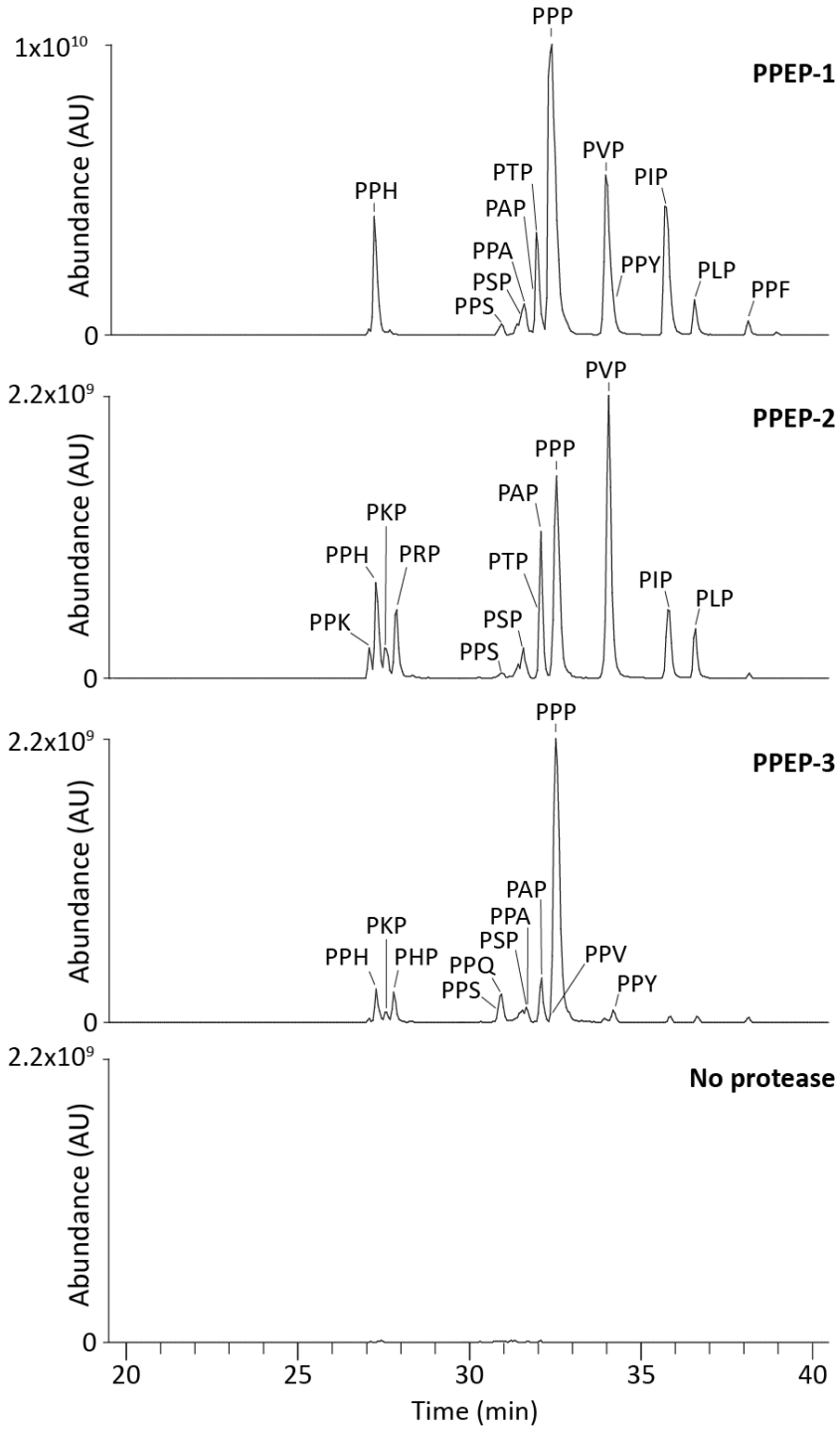


Figure 5. Extracted ion chromatograms of PXP(GGLEEF)/PPX(GGLEEF) product peptides after incubation with PPEPs reveal prime-side specificity profiles. The full combinatorial peptide library was incubated with each of the PPEPs for 3 h. A non-treated control was included to identify the amount of background peptides. After analysis of the product peptides using LC-MS/MS, EIC were constructed for all possible PXP/PPX product peptides (in total 19, both 1+ and 2+ *m/z* values were used). Discrimination between PXP and PPX peptides relied on both inspection of fragmentation spectra and separation on a C18 column **(Figure S3 & S4)**. If product peptides were not separated on the column, lines indicate the relative abundances of the non-separated peptides. Mass tolerance was set to 10 ppm.

Although in the current design, our library is primarily suitable to investigate PPEPs, other proteases that can cleave between the two "XX" sequences in the library peptides could also be tested, assuming that their activity is not compromised by the presence of the surrounding prolines. However, we anticipate that for other proteases a different library design would be beneficial, while still using the same central concept of our approach. For example, the addition of the GGLEEF tail as used in our library can be easily translated to other libraries as well. Although for the current experiments with the PPEPs we used a library with two fixed positions, we believe that a strategy using randomization at five sites, with only one fixed position, would still be possible and provide a broad understanding of the subsite specificity. However, due to the OBOC principle [256], not all individual peptides (2.4 million options when using 19 amino acids) will be present in such a library when starting with the same number of beads as used for our current synthesis (approx. 1.000.000). Although our experiments with PPEP-1 and the two P3-sublibraries showed that partial information about the non-prime side specificity can also be obtained with our method, we believe that a complementary XXPPXX library, in which the biotin is attached to the C-terminus of the peptides, is essential for a more comprehensive characterization of the non-prime-side specificity. Since the negative selection for substrates proceeds identically to that of the current library, both libraries can be mixed, allowing for the profiling of both the prime-side as well as the non-prime-side in a single experiment.

Incubation of PPEP-1 with a collection of FRET-quenched substrate peptides confirms its preference for different amino acids at the P2' position

Based on the endogenous substrates (**Figure 1**) and a small synthetic peptide library [146] PPEP-1 was expected to only tolerate V, I, A and P at the P2' position. To substantiate our results with the combinatorial peptide library, we synthesized twenty PPEP-1 FRET-quenched substrate peptides that only differed at the P2' position (Lys_{Dabcyl}-EVNP↓PXPD-Glu_{Edans}) and tested these with PPEP-1 in a time course kinetic assay. The results of these experiments are depicted in **Figure 6**, in which substrates are ranked (from top left to bottom right) based on their increase in fluorescence during the 1 h incubation. Overall, these data (**Figure 6**) correlated well with the results of the combinatorial library experiment (**Figure 5**). Although cysteines were not included in the combinatorial library design (**Figure 2**), the results with the VNPPCP FRET-peptide showed that it is not tolerated at the P2' position by PPEP-1.

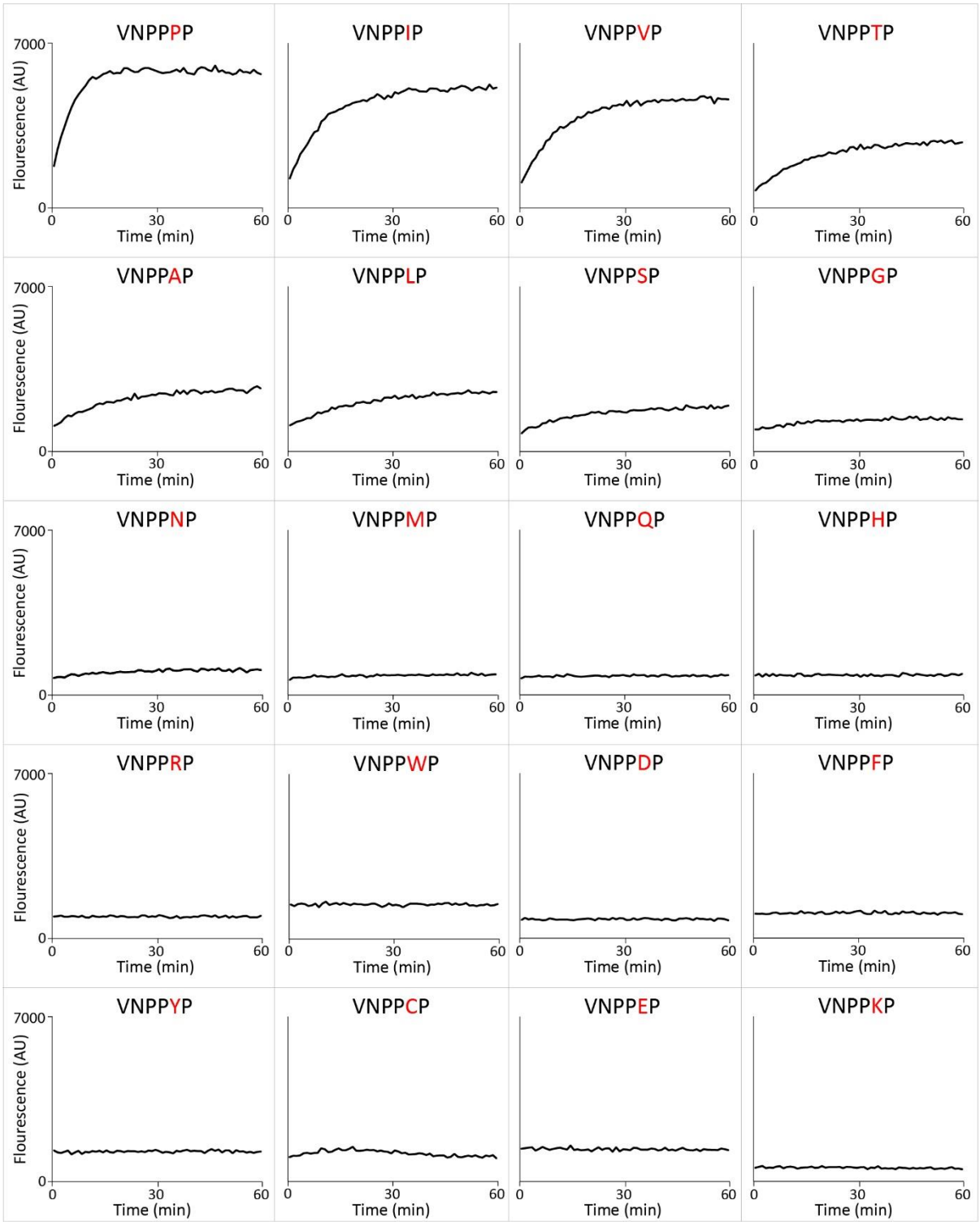


Figure 6. Time course of PPEP-1 mediated cleavage of synthetic FRET-quenched peptides with permutations at the P2' position. The PPEP-1 substrate peptide VNP↓PVP was permutated to generate FRET-quenched peptides (Lys_{Dabcyl}-EVNPPXPD-Glu_{Edans}) containing any of the standard 20 amino acids at the P2' position. These peptides were incubated with PPEP-1 and fluorescence was measured during 1 hr. Peptides are sorted from top left to bottom right based on their cleavage efficiency.

PPEP-3 is able to cleave endogenous PPEP-1 and PPEP-2 substrates when the valine at the P2' position is replaced by a proline

The endogenous substrates of PPEP-1 and PPEP-2 contain the PVP motif at P1'-P3' (**Figure 1**) and the corresponding product peptides (PVPGGLEEF) were clearly observed using the combinatorial library approach (**Figure 5**). However, this product peptide was not observed with PPEP-3 (**Figure 5**), indicating that the corresponding PPEP-1 and PPEP-2 substrate peptides are most likely not cleaved by PPEP-3. We tested this hypothesis using two synthetic FRET-quenched substrate peptides, i.e. Lys_{Dabcyl}-EVNPPVPD-Glu_{Edans} and Lys_{Dabcyl}-EPLPPVPD-Glu_{Edans}, representing substrates of PPEP-1 and PPEP-2, respectively (**Figure 1**). In line with our expectations, PPEP-3 did not hydrolyze either peptide (**Figure 7A**). However, when the P2' Val of both peptides was replaced by a Pro, cleavage by PPEP-3 did occur (**Figure 7A**). On the contrary, although PPEP-1 and PPEP-2 can cleave peptides with four prolines at the P1-P3'position (**Figure 4A,B & 5**), they can still not cleave each other's substrate when the Val at the P2' position is replaced by a proline (**Figure 7A**).

The high specificity of each of the PPEPs for amino acids surrounding the Pro-Pro motif remains obscure. Remarkably, based on the amino acid residue at position 103 (Trp-103) in PPEP-1, two groups were distinguished [148]. In addition to PPEP-1, the Trp-103 group also includes PPEP-2. The other group, to which PPEP-3 belongs, has a Tyr at this position (**Figure S7**). Interestingly, a PPEP-1 W103Y mutant showed very low activity towards a substrate peptide as compared to WT [162]. For PPEP-2, the importance of this residue is less explored. Nevertheless, our data with PPEP-3 show that a tyrosine at this position is compatible with PPEP activity. Whether the tyrosine in PPEP-3 that corresponds to the Trp-103 in PPEP-1 (Tyr-112, **Figure S7**) is responsible for the difference in prime-side specificity between PPEP-3 and the other two PPEP-s requires structural information, especially of a substrate-bound co-crystal.

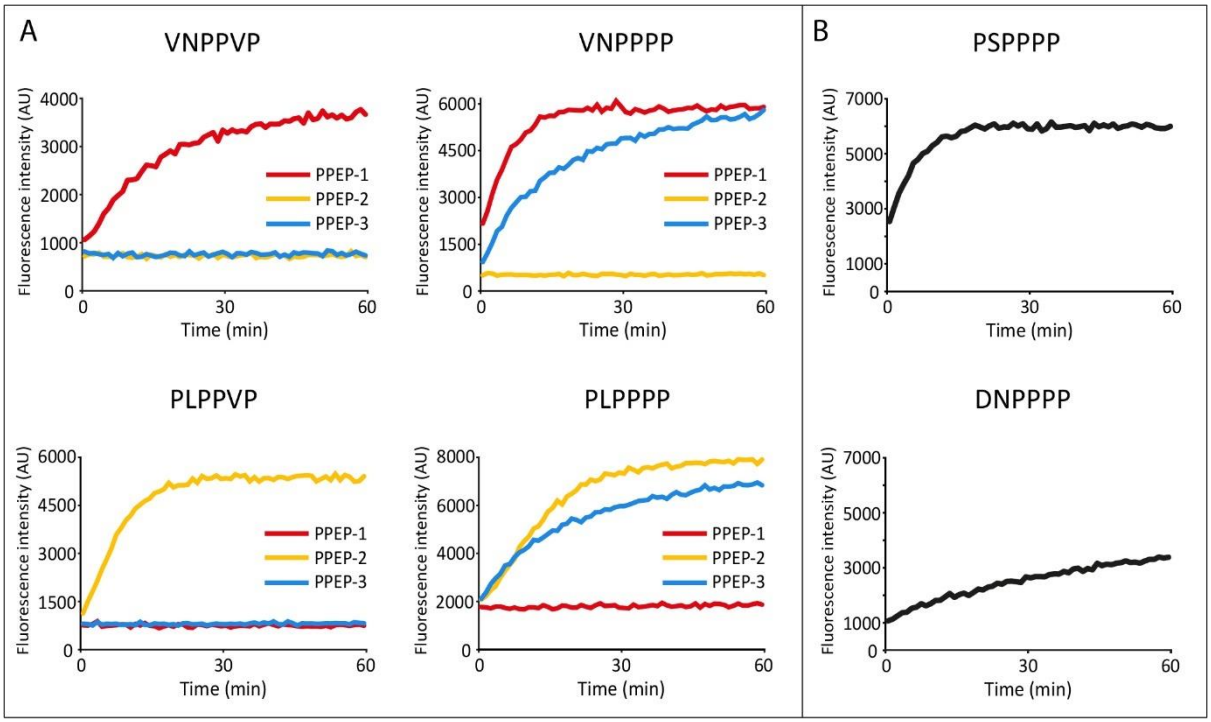


Figure 7. Time course of cleavage of synthetic FRET peptides by PPEP-1, PPEP-2 and PPEP-3. A) Cleavage of PPEP-1 (Lys_{Dabcyl}-EVNP↓PVPD-Glu_{Edans}) and PPEP-2 (Lys_{Dabcyl}-EPLP↓PVPD-Glu_{Edans}) substrate peptides, and their P2'=Pro variants, by PPEP-1, PPEP-2, and PPEP-3. **B)** Cleavage of peptides containing cleavage motifs from putative *G. thermodenitrificans* PPEP-3 substrates by PPEP-3. Only the core sequences (P3-P3') of the individual FRET-quenched peptides are indicated.

Peptides with an XXPPPP motif as observed in *Geobacillus thermodenitrificans* proteins are cleaved by PPEP-3

Next, we looked for possible endogenous substrates of PPEP-3. For PPEP-1 and PPEP-2, genes encoding their substrates are found adjacent to the protease gene (**Figure 1**). Next to PPEP-3, a gene encoding a protein (YpjP, **Figure 1**) with three XXPPXX sequences is found (VTPPAS, EHPPQD and NTPPNW). In line with the data from the combinatorial library, corresponding FRET-quenched peptides were not cleaved by PPEP-3 (data not shown). Overall, our data from the library experiment indicate a strong preference of PPEP-3 for all prolines at the P1-P3' positions (**Figure 4C, 5 & 7A)**. Based on this observation, we hypothesized that possible endogenous substrates containing an XXP↓PPP motif are present in *G. thermodenitrificans* strain NG80-2. Indeed, *G. thermodenitrificans* encodes for four proteins containing four consecutive prolines, two of which contain a signal peptide for secretion as determined by DeepTMHMM and SignalP 6.0 (**Figure S8**) [268,269]. This last feature is thought to be of importance, since PPEP-3 itself is predicted to be a secreted protein. One of the identified proteins, GTNG_0956, contains both a putative CAP-domain as well as an SCP-domain. Admittedly, signal peptide prediction by SignalP 6.0 is inconclusive for this protein, since the signal

peptide would be short in length and no cleavage site is predicted (**Figure S8B**). In contrast, DeepTMHMM predicts a signal peptide with higher confidence (**Figure S8C**). The other protein with an XXPPPP motif and a signal peptide is GTNG_3270. This protein is predicted with high confidence to possess a Sec/SPII signal sequence for integration in the lipid membrane. However, no functional domains were found for this protein. The putative PPEP-3 cleavage sites in GTNG_0956 and GTNG_3270 are PSP↓PPP and DNP↓PPP, respectively. We tested synthetic FRET-quenched peptides containing these motifs for cleavage by PPEP-3 (**Figure 7B**). Both FRET peptides were indeed cleaved by PPEP-3, with PSPPPP being the optimal substrate of the two. MALDI-ToF MS analysis confirmed cleavage between the two prolines within these peptides (**Figure S9**). Collectively, the above data show that the results from the library experiment resulted in testable hypotheses about possible endogenous PPEP-3 substrates in *G. thermodenitrificans* strain NG80-2. For PPEP-1 and PPEP-2, the endogenous substrates were identified based on synthetic peptides, bio-informatic predictions and MS-based secretome analyses [147,157]. Interestingly, none of the sites in the endogenous substrates of these two PPEPs has four consecutive prolines, even though for both proteases the PPPGGLEEF product peptide was (one of) the major product peptides. In order to identify the endogenous substrate of PPEP-3, additional experiments such as secretome analyses in combination with gene knockout studies are needed, although we cannot exclude the possibility that the substrate(s) originates from a different organism than *G. thermodenitrificans*.

Conclusion

In conclusion, we show for the first time a strategy to study the activity and specificity of a protease by combining a combinatorial synthetic peptide library with mass spectrometry. Our method takes each amino acid into account (with the exception of cysteine) and directly showed combinations of amino acids that were tolerated at the P2’ and P3’ positions. We believe that the strategy presented here is a generic one which can, with a tailored design of the library, also be used to explore substrate specificities of other proteases. Importantly, with the new method we have not only confirmed the prime-side specificity of PPEP-1 and PPEP-2, but also revealed some new unexpected peptide substrates. Moreover, we have characterized a new PPEP (PPEP-3 from *Geobacillus thermodenitrificans*) which has a prime-side specificity that is very different from that of the other two PPEPs.

Experimental procedures

Expression and purification of PPEPs

PPEP-1 and PPEP-2 were expressed and purified as previously described [146,157]. For the expression of PPEP-3, a pET28a vector containing an *E. coli* codon optimized 6xHis-PPEP-3 (lacking the signal peptide) construct was ordered from Twist Bioscience. The pET-28a 6xHis-PPEP-3 plasmid was transformed to *E. coli* strain Rosetta and PPEP-3 expression was induced using 1 mM IPTG. Lysates were prepared as described in the protocol for preparation of cleared *E. coli* lysates under native conditions as described in the fifth edition of the QIAexpressionist (Qiagen). The lysates were loaded onto a 1 ml HisTrap HP column (GE healthcare) coupled to an ÄKTA Pure FPLC system (GE healthcare). Column was washed using wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole) and 6xHis-PPEP-3 was eluted using a step gradient with elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole). Imidazole was removed by dialysis using 50 mM NaH_2PO_4, 300 mM NaCl.

Synthesis of the combinatorial peptide library

Combinatorial peptide libraries were synthesized basically as has been previously described [270]. In short, peptide libraries were synthesized by solid phase peptide synthesis on a Syro II peptide synthesizer (Multisyntech, Germany). Synthesis was performed in 19 reactors (2 ml) using about 1 g of Tentagel resin (Rapp-polymere, Germany) resin (total loading 190 µmol), applying Fmoc chemistry with HATU/NMM activation, 20 % piperidine in NMP for Fmoc removal and NMP as a solvent. For each fixed position in each reactor the same amino acid was coupled, for each random position (X) in each reactor a different amino acid was coupled, after which the resin beads were removed from each reactor, mixed thoroughly, and equally split over the 19 reactors again to allow for the subsequent stages of the synthesis. After the last random position, the resin beads were not mixed, leaving 19 sub-libraries. Biotin was introduced into the resin bound peptides by a two hour coupling with a sixfold equimolar preactivated mixture of biotin and PyBop. Cleavage using TFA/water/ethanethiol 18/1/1, 3h, RT, was used to isolate the peptides from the resin. Approx. 12 ml ether/pentane was added to each sub-library and sub-libraries were incubated at -20 °C for 10 min before centrifugation at 3300 rpm for 10 min at -9 °C. Pellets were washed with approx. 13 ml ether/pentane and air-dried. Dried pellets were resuspended in 2 ml MilliQ/acetonitrile and freeze-dried. Stocks of 10 nmol peptide/µl were prepared in DMSO.

MALDI-FT-ICR MS

To analyze samples using MALDI-FT-ICR MS, the vacuum dried product peptides were reconstituted in 20 µl 0.1% formic acid. Next, 1 µl sample was combined with 1 µl matrix (5 mg/ml α-Cyano-4-hydroxycinnamic acid) and 1 µl was spotted on an AnchorChip target (Bruker). Analysis was performed on a 15 T MALDI-FT-ICR MS (Bruker Daltonics).

Combinatorial peptide library assays

To remove non-biotinylated peptides, 50 nmol of peptides from the (sub)library (5 µl 10 nmol/µl stock in 1 ml PBS) was loaded onto a 3 ml filter column containing 1 ml Pierce Monomeric Avidin Agarose beads (Thermo) (binding capacity is >1.2 mg/ml biotinylated BSA or >18 nmol/ml). Prior to loading the libraries, the avidin column was washed five times with 1 ml 0.01% formic acid (pH 2.7) and subsequently washed five times with 1 ml PBS. After loading peptides, the flow-through was collected. Next, 1 ml PBS was loaded onto the column and flow-through was collected. Then, the collected flow-throughs were reapplied to the column to ensure saturation of the avidin beads. The column was washed five times with 1 ml PBS to remove non-biotinylated peptides. Next, 1 ml 0.1 M glycine (pH 2.7) was applied to the column and the flow-through was discarded because the pH of the last drop of this fraction was still neutral as checked with a pH indicator strip. Then, biotinylated peptides were eluted with 9 ml 0.1 M glycine (pH 2.7). Eluted peptides were desalted using reversed-phase solid phase extraction cartridges (Oasis HLB 1cc 30mg, Waters) and eluted with 400 µl 50% acetonitrile (v/v) in 0.1% formic acid. Samples were dried by vacuum concentration and stored at -20 °C until further use. If the binding efficiency of the avidin beads is the same for the peptide library as for biotinylated BSA, and no peptides are lost during the pre-wash steps, we expect approx. 20 nmol of peptide yield after the avidin pre-clearing step.

Pre-cleaned (sub)libraries (approx. 10 nmol) were incubated with a PPEP (200 ng) for 3 h at 37 °C in PBS. A non-treated control was included. After incubation, the samples were loaded onto an in-house constructed column consisting of a 200 µl pipette tip containing a filter and a packed column of 100 µl of Pierce High Capacity Streptavidin Agarose beads (Thermo, column was washed four times with 150 µl PBS prior to use), in order to remove the biotinylated peptides. The flow-through and 4 additional washes with 125 µL were collected. The resulting product peptides were desalted using reversed-phase solid phase extraction cartridges (Oasis HLB 1cc 30mg, Waters) and eluted with 400 µl 30% acetonitrile (v/v) in 0.1% formic acid. Samples were dried by vacuum concentration and stored at -20 °C until further use.

LC-MS/MS analyses

Product peptides were analyzed as described previously [271] by on-line C18 nanoHPLC MS/MS with a system consisting of an Ultimate3000nano gradient HPLC system (Thermo, Bremen, Germany), and an Exploris480 mass spectrometer (Thermo). Fractions were injected onto a cartridge precolumn (300 µm × 5 mm, C18 PepMap, 5 µm, 100 A, and eluted via a homemade analytical nano-HPLC column (50 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm, 120 A (Dr. Maisch, Ammerbuch, Germany). The gradient was run from 2% to 36% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 52 min. The nano-HPLC column was drawn to a tip of ~10 µm and acted as the electrospray needle of the MS source. The mass spectrometer was operated in data-dependent MS/MS mode for a cycle time of 3 seconds, with a HCD collision energy at 30 V and recording of the MS2 spectrum in the orbitrap, with a quadrupole isolation width of 1.2 Da. In the master scan (MS1) the resolution was 120,000, the scan range 350-1600, at standard AGC target @maximum fill time of 50 ms. A lock mass correction on the background ion m/z=445.12003 was used. Precursors were dynamically excluded after n=1 with an exclusion duration of 10 s, and with a precursor range of 10 ppm. Charge states 1-5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 30,000 at an AGC target of 100% @maximum fill time of 60 ms.

LC-MS/MS data analysis

We generated a database containing all 6859 peptides from the P3=Val sublibrary, i.e. Ahx-EVXPPXXGGLEEF. The Ahx in all peptide sequences was replaced by a Ile (they have an identical mass). Raw data were converted to peak lists using Proteome Discoverer version 2.4.0.305 (Thermo Electron), and submitted to the in-house created P3=Val sublibrary database using Mascot v. 2.2.7 (www.matrixscience.com) for peptide identification, using the Fixed Value PSM Validator. Mascot searches were with 5 ppm and 0.02 Da deviation for precursor and fragment mass, respectively, and no enzyme specificity was selected. Biotin on protein N-terminus was set as a variable modification. Raw data analysis was performed in Xcalibur Qual Browser (Thermo). The EIC displaying all PXPGGLEEF/PPXGGLEEF peptides was created by plotting the intensities of the signal corresponding to the monoisotopic *m/z* values of both 1+ and 2+ charged peptides. To assign individual peptides to their respective peaks, each individual peptide was plotted in an EIC and peptides were assigned to peaks based on retention time and abundance.

FRET peptide cleavage assays

Time course kinetic experiments with PPEPs were performed using fluorescent FRET-quenched peptides. FRET peptides consisted of Lys_{Dabcyl}-EXXPPXXD-Glu_{Edans}, in which X varied between the different peptides tested. To test cleavage of FRET peptides by PPEPs, 75 µL of FRET peptide (100 µM in PBS) was added to a well of a 96-well Cellstar black plate (Greiner). Immediately prior to the assay, 75 µl PBS containing a PPEP (0.2-1 µg) was added. Peptide cleavage was measured using the Envision 2105 Multimode Plate Reader. Fluorescence intensity was measured each minute for 1 h, with 10 flashes per measurement. The excitation and emission wavelengths were 350 nm and 510 nm, respectively. When comparing PPEP-1, PPEP-2, and PPEP-3 in a single experiment, the relative fluorescence was determined by regarding the highest signal as 100%.

Bioinformatic analyses

PPEP-3 structure prediction was carried out using Colabfold [199] with the following parameters: template mode=none, MSA mode=MMseqs2, pair mode=unpaired+paired, model type=auto, and number of recycles=3. Signal peptide predictions were performed using DeepTMHMM [268] and SignalP 6.0 [269]. For sequence alignments, the Clustal Omega Multiple Sequence Alignment tool was used [272].

The cleavage motifs were created using Weblogo 3 [273] with the units set to probability. The sequences logos were generated based on the relative intensities of the 10 most abundant product peptides for each PPEP. A list with these 10 product peptides was created, in which each individual peptide occurred a number of times, according to its relative abundance to the other peptides. For example, if product 'A' was 100 times more abundant than product 'B', product 'A' was present 100 times more in this list than product 'B'.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [200] via the PRIDE [201] partner repository with the dataset identifier PXD038277.

Acknowledgements

This research was supported by an ENW-M grant (OCENW.KLEIN.103) from the Dutch Research Council (NWO). We thank prof. Ulrich Baumann and dr. S. Nicolardi for the critical reading of an earlier version of this manuscript. O.I.K. acknowledges support by the Interdisciplinary Scientific and Educational School of Moscow University “Molecular Technologies of the Living Systems and Synthetic Biology”.

Supporting information

Table S1: Database search of LC-MS/MS data (XLSX) can be found online

https://pubs.acs.org/doi/10.1021/acs.analchem.3c01215

Supporting information

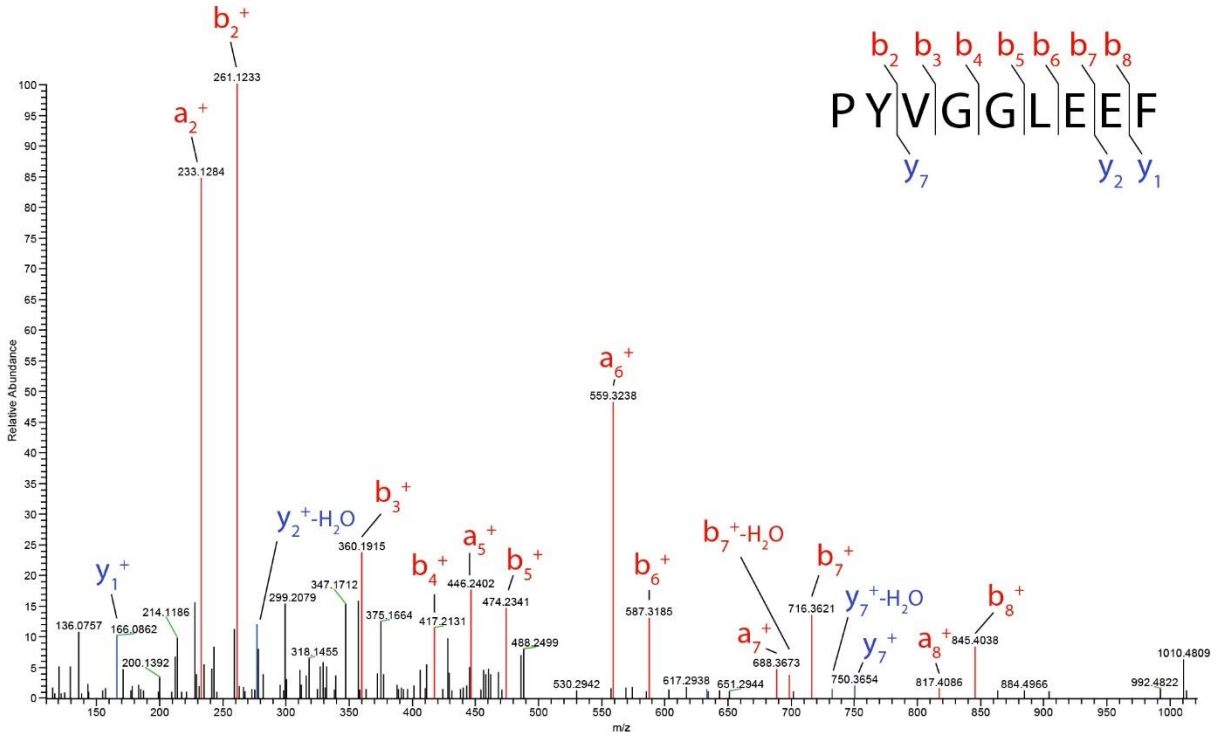


Figure S1. Fragmentation spectrum of PYVGGLEEF. MS/MS spectrum of the PYVGGLEEF peptide that was used in the design of the combinatorial peptide library.

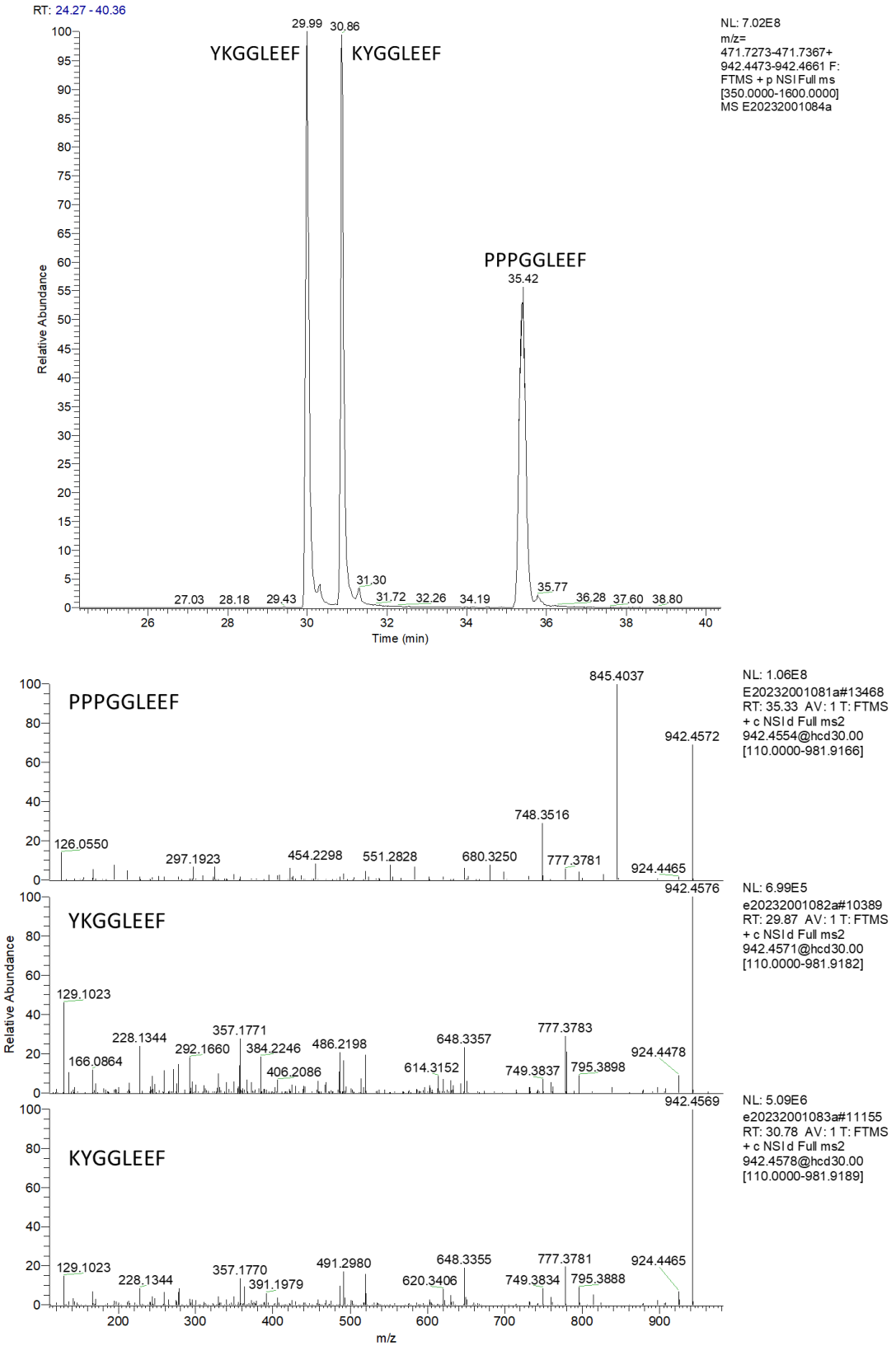


Figure S2. Chromatographic and MS/MS fragmentation characteristics of peptides KYGGLEEF, YKGGLEEF and PPPGGLEEF. Synthetic peptides KYGGLEEF, YKGGLEEF and PPPGGLEEF (1 uL of a 1:1:1 (v/v/v) mix of 200 fmol/uL of each peptide) were analyzed using LC-MS/MS. Upper panel: chromatographic behavior. Peaks are assigned based on injections of individual peptides. Lower panel: MS/MS spectra.

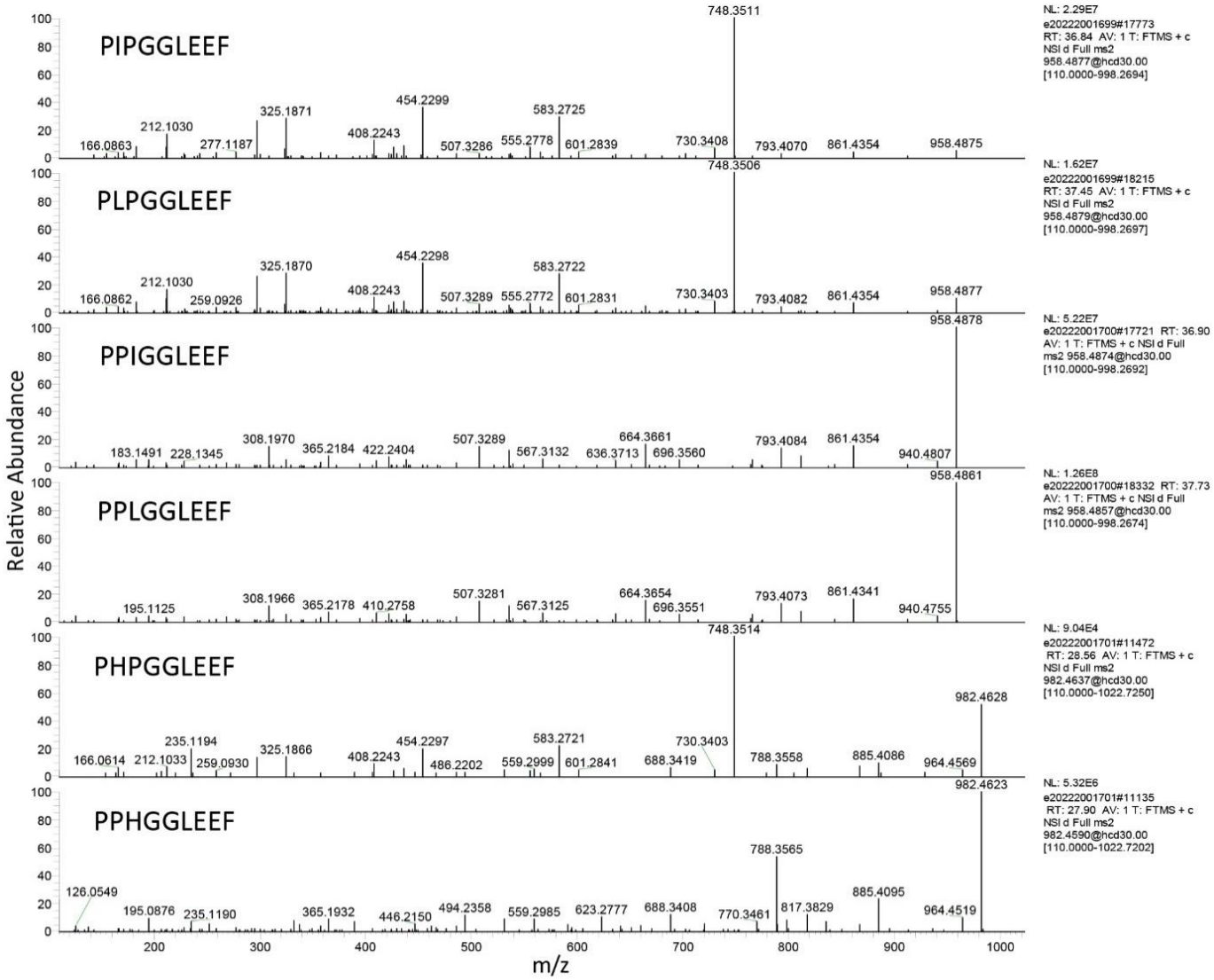


Figure S3. Synthetic PXP/PPX product peptides display distinct fragmentation spectra. Synthetic peptides with either a PXP or PPX motif were analyzed using LC-MS/MS.

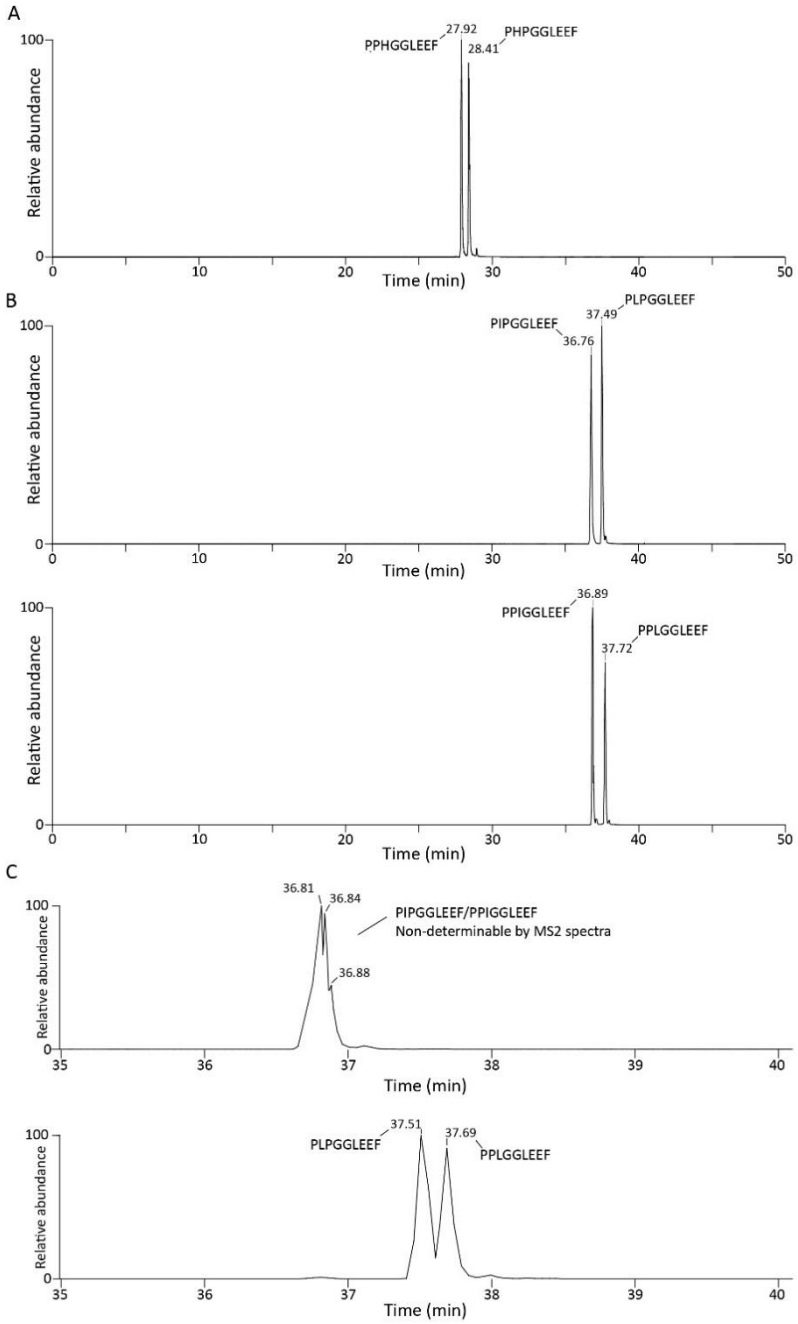


Figure S4. Separation of PXP/PPX peptides on C18 column. Shown are extracted ion chromatograms of the corresponding peptides with a mass tolerance of 10 ppm. **A)** Synthetic peptides PHPGGLEEF and PPHGGLEEF were mixed at an equimolar concentration and analyzed using LC-MS/MS. Assignment of the peaks is based on LC-MS/MS analyses of the two peptides separately (data not shown). **B)** PXP/PPX peptides that contain either a Leu or Ile residue are separated on a C18 column. Assignment of the peaks is based on LC-MS/MS analyses of the two peptides separately (data not shown). **C)** PIPGGLEEF and PPIGGLEEF are not fully separated on a C18 column. Although some separation is observed, no conclusions could be made about the order of elution. However, panel B shows a lower retention time for PIPGGLEEF. PLPGGLEEF and PPLGGLEEF are, although not completely, separated on the column.

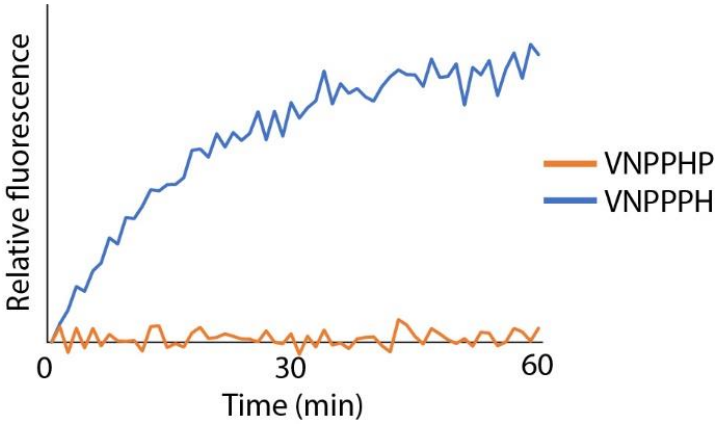


Figure S5. PPEP-1 can cleave a VNPPPH peptide but not a VNPPHP peptide. FRET-quenched peptides containing either VNPPHP or VNPPPH were incubated with PPEP-1 for 1 h. Fluorescence was measured in a fluorescence microplate reader.

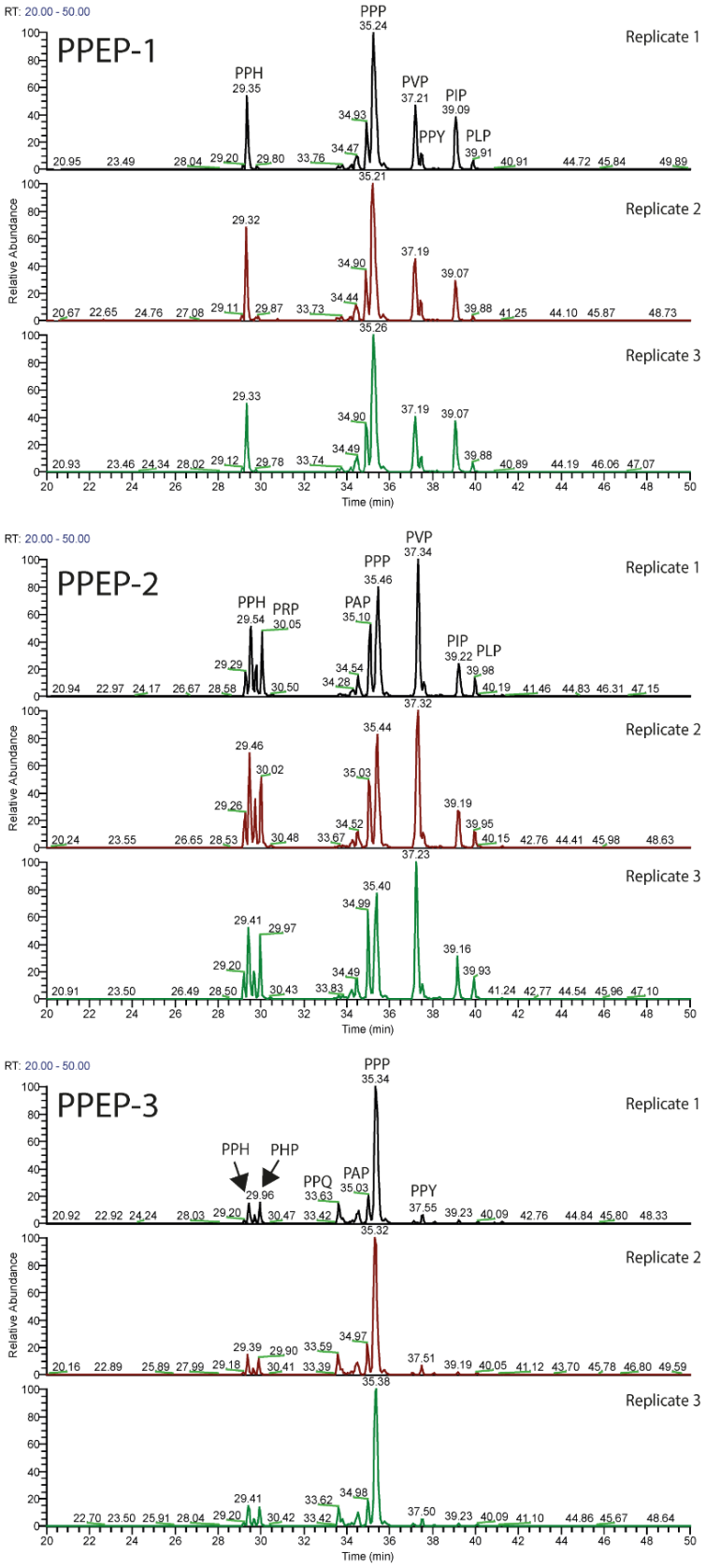


Figure S6. Extracted ion chromatograms of PXP/PPX product peptides from three independent incubations of the full peptide library with PPEP-1, -2 and -3. See also Figure 5.

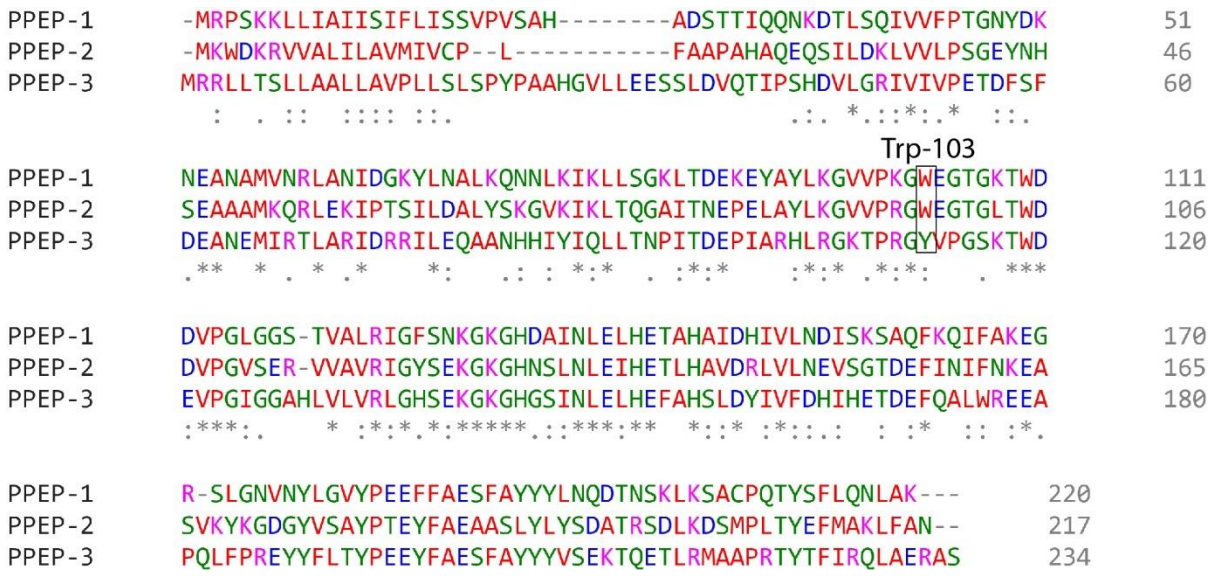


Figure S7. Alignment of PPEP-1, PPEP-2 and PPEP-3. Alignment was created using the Clustal Omega multiple sequence alignment tool.

A

Protein	Gene name	Predicted PPEP-3 cleavage motif
Uncharacterized protein	GNTG_0956	PSP↓PPP
Uncharacterized protein	GTNG_1059	TAP↓PPP
Uncharacterized protein	GTNG_1937	VDP↓PPP
Uncharacterized protein	GTNG_3270	DNP↓PPP

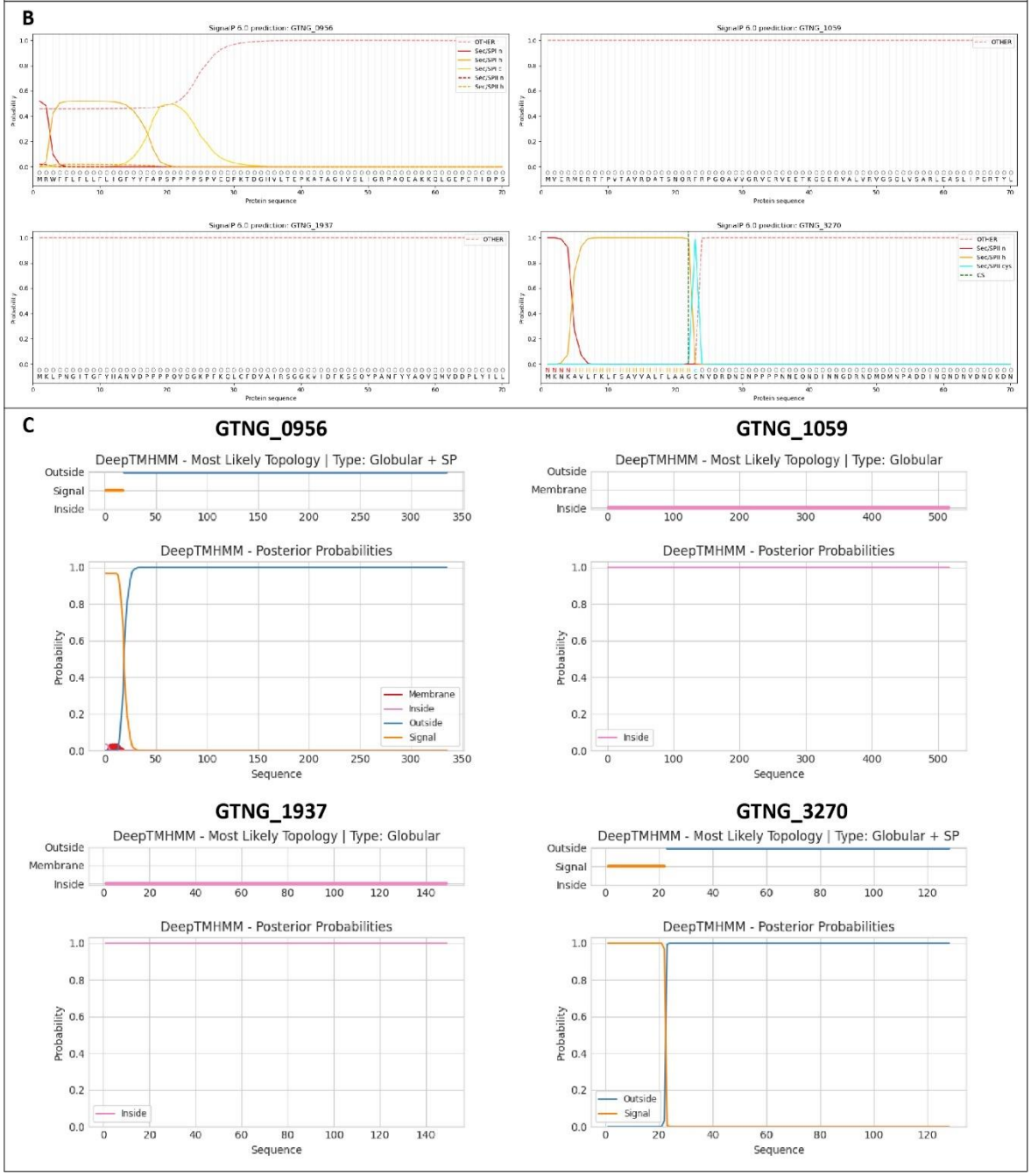


Figure S8. Signal peptide prediction of putative PPEP-3 substrates in *G. thermodenitrificans*. A) Overview of the proteins and their predicted PPEP-3 cleavage motifs. **B)** Signal peptide prediction by SignalP 6.0. **C)** Signal peptide prediction by DeepTMHMM.

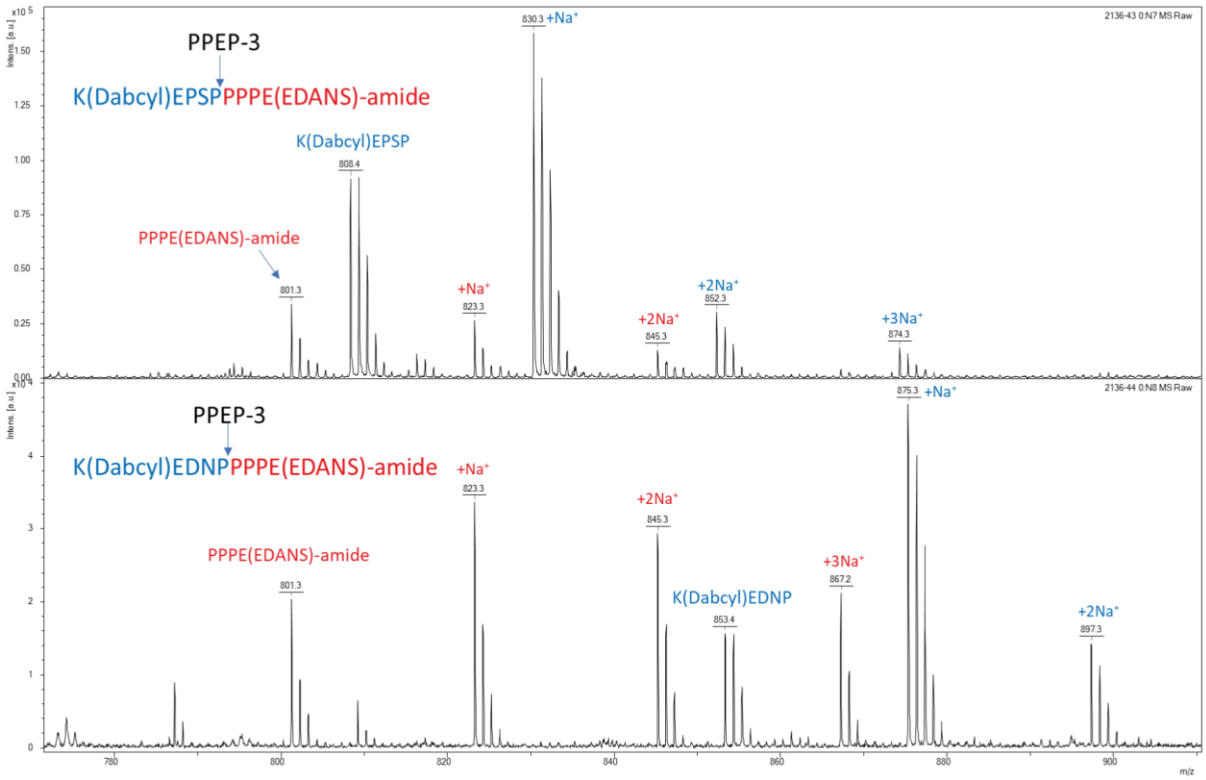


Figure S9. MALDI-ToF MS analysis of the product peptides from the incubations of PPEP-3 with the two FRET-peptides as presented in Figure 7B.

5

Non-prime- and Prime-side Profiling of Pro-Pro Endopeptidase Specificity Using Synthetic Combinatorial Peptide Libraries and Mass Spectrometry

Bart Claushuis[1], Robert A. Cordfunke[2], Arnoud H. de Ru[1], Jordy van Angeren[1], Ulrich Baumann[3], Peter A. van Veelen[1], Manfred Wuhrer[1], Jeroen Corver[4], Jan W. Drijfhout[2], Paul J. Hensbergen[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] Department of Immunology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[3] Department of Chemistry, Institute of Biochemistry, University of Cologne, Cologne, 50674, Germany
[4] Leiden University Center of Infectious Diseases, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

Published in FEBS journal, 2024, 17160

DOI: 10.1111/febs.17160

Abstract

A group of bacterial proteases, the Pro-Pro endopeptidases (PPEPs), possess the unique ability to hydrolyze proline-proline bonds in proteins. Since a protease's function is largely determined by its substrate specificity, methods that can extensively characterize substrate specificity are valuable tools for protease research. Previously, we achieved an in-depth characterization of PPEP prime-side specificity. However, PPEP specificity is also determined by the non-prime-side residues in the substrate.

To gain a more complete insight into the determinants of PPEP specificity, we characterized the non-prime- and prime-side specificity of various PPEPs using a combination of synthetic combinatorial peptide libraries and mass spectrometry. With this approach, we deepened our understanding of the P3-P3' specificities of PPEP-1 and PPEP-2, while identifying the endogenous substrate of PPEP-2 as the most optimal substrate in our library data. Furthermore, by employing the library approach, we investigated the altered specificity of mutants of PPEP-1 and PPEP-2.

Additionally, we characterized a novel PPEP from *Anoxybacillus tepidamans*, which we termed PPEP-4. Based on structural comparisons, we hypothesized that PPEP-4 displays a PPEP-1-like prime-side specificity, which was substantiated by the experimental data. Intriguingly, another putative PPEP from *Clostridioides difficile*, CD1597, did not display Pro-Pro endoproteolytic activity.

Collectively, we characterized PPEP specificity in detail using our robust peptide library method and, together with additional structural information, provide more insight into the intricate mechanisms that govern protease specificity.

Introduction

Proteases represent a diverse and indispensable class of enzymes that play pivotal roles in cellular homeostasis, protein turnover, and the regulation of various biological pathways. Their ability to hydrolyze peptide bonds is vital for the activation, maturation, or degradation of proteins. Among the diverse array of proteases found across different organisms, bacterial proteases are particularly intriguing due to their significance in bacterial physiology [274,275], pathogenesis [276,277], antimicrobial drug targets [278] and biotechnological applications [279].

Protease function is largely determined by substrate specificity, i.e., which residues are tolerated surrounding the cleavage site. Proline residues, for instance, are generally excluded as part of the cleavage site due to their cyclic structure, imposing conformational constraints that hinder proteolytic cleavage [152,280]. However, several proteases have been described that selectively cleave N- or C-terminally of proline residues [281–284]. A notable group of bacterial proteases, the Pro-Pro endopeptidases (PPEPs), possess the unique ability to specifically hydrolyze proline-proline bonds.

PPEPs are predicted to be present in many bacterial species [148] and several PPEPs have been characterized [147,157,230]. Although these enzymes appear very similar based on their protein sequence, small structural differences result in distinct substrate specificities [230]. PPEP specificity is at the minimum dependent on the six residues flanking the cleavage site (P3-P2-P1↓P1'-P2'-P3', Schechter and Berger nomenclature [17]), with the permissible residues dictated by interactions within the active site. A comprehensive understanding of PPEP specificity holds promise for predicting endogenous substrates, facilitating industrial applications, developing inhibitors, and their use as potential biomarkers.

Several methods are available to profile protease specificity, such as gel- and fluorescence-based methods [285], N-terminomics approaches [20,23], and library methods. For the latter, approaches using phage display [169,170], positional scanning [286], proteome-derived libraries [24] and synthetic combinatorial peptide libraries exist [240,258]. Synthetic combinatorial peptide libraries consist of systematically synthesized peptides that cover all possible amino acid combinations around a cleavage site. Compared to proteome-derived peptide libraries, synthetic combinatorial peptide libraries contain potential substrates in equimolar amounts, which allows for a more quantitative approach.

Previously, we reported a novel method to profile PPEP prime-side specificity by combining the use of a synthetic combinatorial peptide library with LC-MS/MS analysis [230]. In this approach, protease-generated product peptides are enriched by negative selection and subsequently analyzed by LC-MS/MS. This approach allowed for an in-

depth characterization of the prime-side specificity of PPEP-1, PPEP-2, and PPEP-3 and revealed the differences between the three PPEPs. However, PPEP specificity is also determined by the substrate's non-prime-side residues. To obtain a thorough understanding of PPEP specificity, a method that allows for the characterization of the non-prime-side specificity is needed as well.

In order to achieve an integrated analysis of both non-prime- and prime-side specificities, we expanded our combinatorial peptide library method by synthesizing a complementary library that allowed us to profile the non-prime-side specificity of PPEPs. In addition, profiling of the complete specificity of PPEPs was achieved by combining the non-prime- and prime-side libraries. We not only used our method with known PPEPs but also applied it to determine the specificity of two uncharacterized PPEP homologs from both *Clostridioides difficile* (formerly known as *Clostridium difficile*) and *Anoxybacillus tepidamans*. By combining the specificity profiles of PPEPs with structural information, we elaborate on the structure-function relationship of PPEPs.

Results and Discussion

Design and testing of a synthetic combinatorial peptide library to determine PPEP non-prime-side specificity

To determine the non-prime-side specificity of PPEPs, we constructed a new synthetic combinatorial peptide library according to the sequence PTEDAVXX**PP**XXEZZO motif (X=any residue except Cys, Z=6-aminohexanoic acid, O=Lys(biotin)-amide) (**Figure 1A**). In analogy with the previous library that was used to determine the prime-side specificity of PPEPs (**Figure 1A**) [230], the two core Pro residues (P1-P1') were fixed while surrounding positions (P3-P2, P2'-P3') could contain any amino acid residue (except Cys, omitted to prevent disulfide bridges). In contrast to the prime-side library, the biotin in the new library was added at the C-terminus, while the peptide tail was added to the N-terminus. The sequence of this tail (PTEDAV) showed good chromatographic behavior and fragmentation characteristics (data not shown) and was based on product peptides from the endogenous substrates of PPEP-1 [146]. Collectively, the new library was designed to allow for the identification of PTEDAVXXP product peptides following a similar approach as previously described (**Figure 1B**) [230].

To test the specificity profiling potential of the newly synthesized peptide library, we incubated the library with PPEP-1 and PPEP-2 and created extracted ion chromatograms (EICs) of the product peptides after LC-MS/MS analysis (**Figure 2**). Previously, we reported that PPEP-1 and PPEP-2 have a markedly different non-prime-side specificity since the proteases are unable to cleave each other's substrates [157]. In addition, PPEP-1 is known to tolerate multiple amino acids at the P2 and P3 positions [146], while for PPEP-2 the non-prime-side specificity has been less explored. The data in **Figure 2** corroborated the difference in non-prime specificity between PPEP-1 and PPEP-2, since most of the product peptides were not shared between the two PPEPs. In addition, it was readily apparent that the non-prime-side specificity was less stringent for PPEP-1 than for PPEP-2, i.e., more different product peptides were needed to account for >90% of the total intensity.

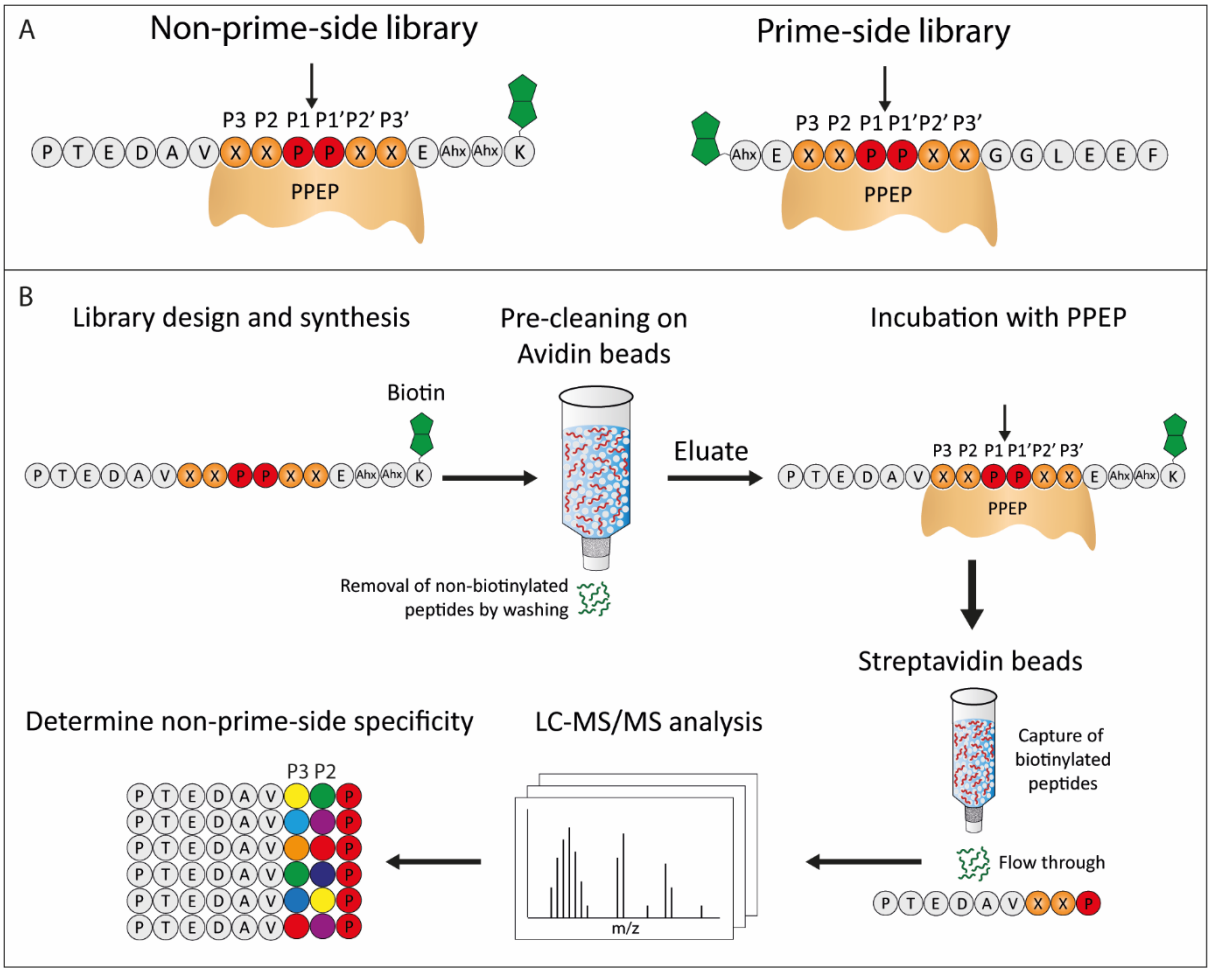


Figure 1. Design of the non-prime-side synthetic combinatorial peptide library. A) Design of the non-prime-side library (left) and the previously described prime-side library (right) [230]. The expected cleavage site is indicated with an arrow and biotin is represented in green. Ahx=6-aminohexanoic acid. **B)** Strategy for determining the non-prime-side specificity of PPEPs. Nonbiotinylated peptides are removed by washing the peptide library on an avidin column. Then, the eluted peptide library is incubated with a PPEP and subsequently loaded onto a streptavidin column. The biotinylated peptides are captured, while the PTEDAVXXP product peptides pass through the column. The product peptides are analyzed by LC-MS/MS, after which the non-prime-side specificity can be determined. Figure was adapted from Claushuis *et al.* (2023) [230], which is available under Creative Commons Attribution 4.0 International License.

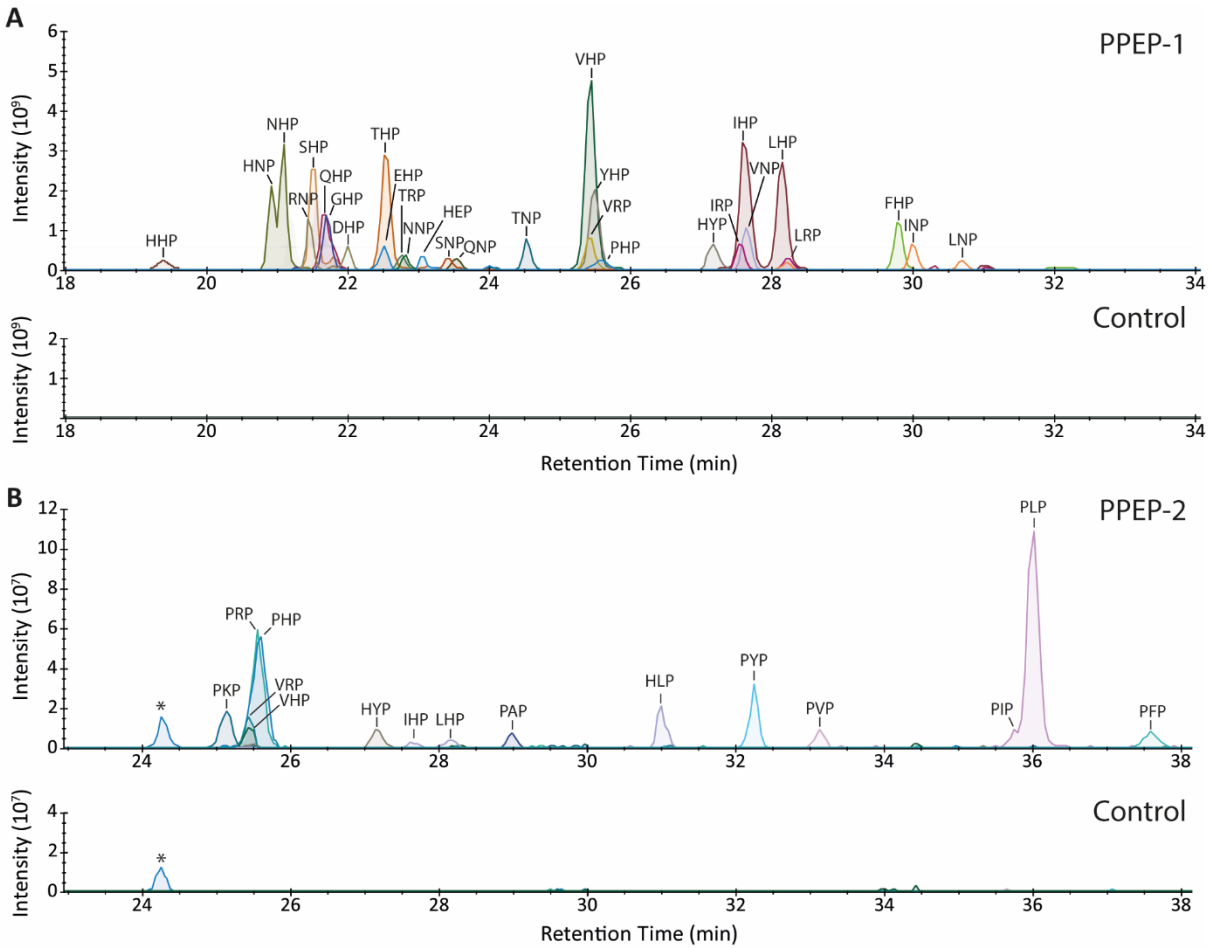


Figure 2. Incubated of the non-prime side combinatorial library with PPEP-1 and PPEP-2. Peptides were incubated with either **A)** PPEP-1 or **B)** PPEP-2. After product peptide enrichment and analysis with LC-MS/MS, a database search was performed using an in-house created database containing the 130.321 possible 16-mer peptides. The results were filtered for the 9-mer product peptides corresponding to the cleavage between the two fixed prolines (PTEDAVXXP). From these, the most abundant product peptides that together accounted for >90% of the total intensity were used to create an EIC. To distinguish between isomeric product peptides, manual inspection of MS/MS spectra, combined with LC-MS/MS of additional synthetic peptides, was performed. Mass tolerance was set to 5 ppm. An untreated control sample was included. *A molecule corresponding to the mass of PTEDAVPHP (481.7325, $[M+2H]^{2+}$) was observed but MS/MS spectra indicate no PTEDAVXXP product peptide.

Also with the new library, the assignment of specific product peptides was based on manual inspection of MS/MS spectra and additional LC-MS/MS analyses of candidate product peptides. For example, this allowed us to unambiguously assign the major product peptide of PPEP-2 as PTEDAVPLP and not PTEDAVPIP (**Figures 2 and 3A**). Moreover, a cleavage assay with PPEP-2 and FRET-quenched peptides showed that PLPPVP is cleaved much more efficiently than PIPPVP by PPEP-2 (**Figure 3C**). Similar analyses were used to correctly assign other isomeric peptides, e.g. PTEDAVHIP, PTEDAVHLP, PTEDAVIHP and PTEDAVLHP (**Figures 3B,D**).

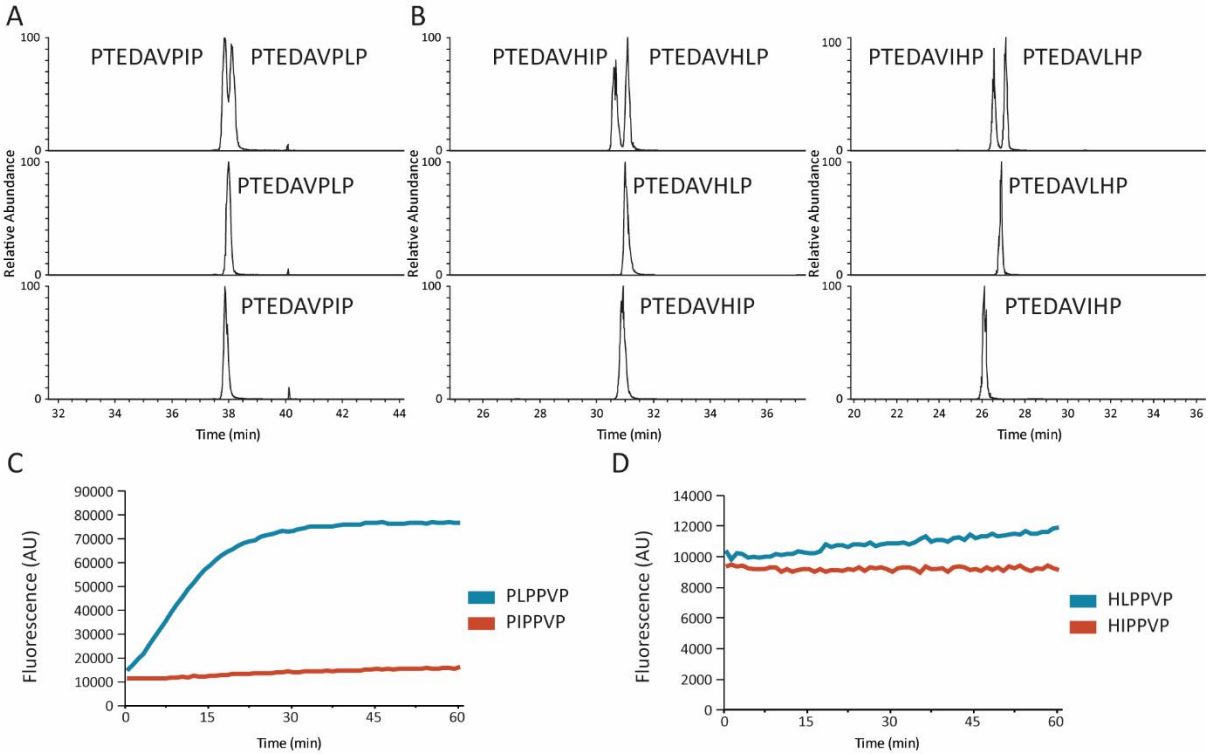


Figure 3. Additional analyses using synthetic peptides to assign product peptides. To correctly assign product peptides with equal masses, several peptides were synthesized to assess the separation during LC or the proteolysis by PPEP-2. **A)** Separation of PTEDAVPLP and PTEDAVPIP. **B)** Separation of peptides from the non-prime-side library containing HIP/HLP/IHP/HLP (P3-P1). **C)** Time course of PPEP-2 mediated cleavage of the synthetic FRET-quenched peptides Lys(Dabcyl)-EP(I/L)PPVPD-Glu(EDANS). **D)** Time course of PPEP-2 mediated cleavage of the synthetic FRET-quenched peptides Lys(Dabcyl)-EH(I/L)PPVPD-Glu(EDANS).

We also removed two 9-mer product peptides from our analyses, PTEDAVGGP and PTEDAVAGP, because manual inspection of the data demonstrated that they corresponded to the isomeric 8-mer peptides PTEDAVNP and PTEDAVQP, resulting from cleavage before the first fixed Pro at P1 in our design (PTEDAVXX↓PPXX).

Although the product peptide signals in the PPEP-2 treated sample and the untreated control greatly differ due to the proteolysis by PPEP-2, the signals in the PPEP-2 treated sample are approximately two orders of magnitude lower compared to the signals we observe in the PPEP-1 treated sample (**Figure 2**). Previously, we observed roughly five times lower signals for PPEP-2 prime-side product peptides compared to those of PPEP-1 [230]. This is a markedly smaller difference than we have observed in our non-prime-side results. A likely explanation is that, either for PPEP-1 or PPEP-2, the P4 or the P4' position also determines substrate specificity, since these differ in the designs of both peptide libraries while all other factors remained constant (**Figure 1**).

Overall, the results with PPEP-1 and PPEP-2 as described above were in line with our expectations and showed that profiling of PPEP specificity could be achieved with the non-prime-side peptide library.

Profiling the non-prime- and prime-side specificity of PPEPs in a single experiment

Having tested the new library, we next sought to profile the P3-P3' specificity of PPEPs in a single experiment. To this end, we mixed the previously described prime-side peptide library [230] with the newly synthesized non-prime-side library (1:1) and incubated the mixture with either PPEP-1 or PPEP-2. The total amount of peptides was left unaltered, meaning that in comparison to the non-prime-side library experiment, only half the amount of each peptide was used. The results of the LC-MS/MS analyses were used to create EICs of the product peptides (**Figures 4A,B**). Based on the intensities of the product peptides, we constructed logos depicting the relative occurrence of a residue at a position surrounding the cleavage site (**Figures 4C,D**). The logos show how strongly the specificity at a position surrounding the cleavage site is determined by certain residues. Although the logos in **Figures 4C,D** do not take subsite cooperativity into account (they only show the relative occurrence of a residue at a certain position in the product peptides), this is the case for the EICs (**Figures 4A,B**). Therefore, the EICs and logos are complementary to each other. Since we make use of two different peptide libraries, we cannot draw conclusions about the subsite cooperativity spanning the cleavage site, e.g., the influence of a residue at the P2 position on the tolerance by a protease for a residue at the P2' position.

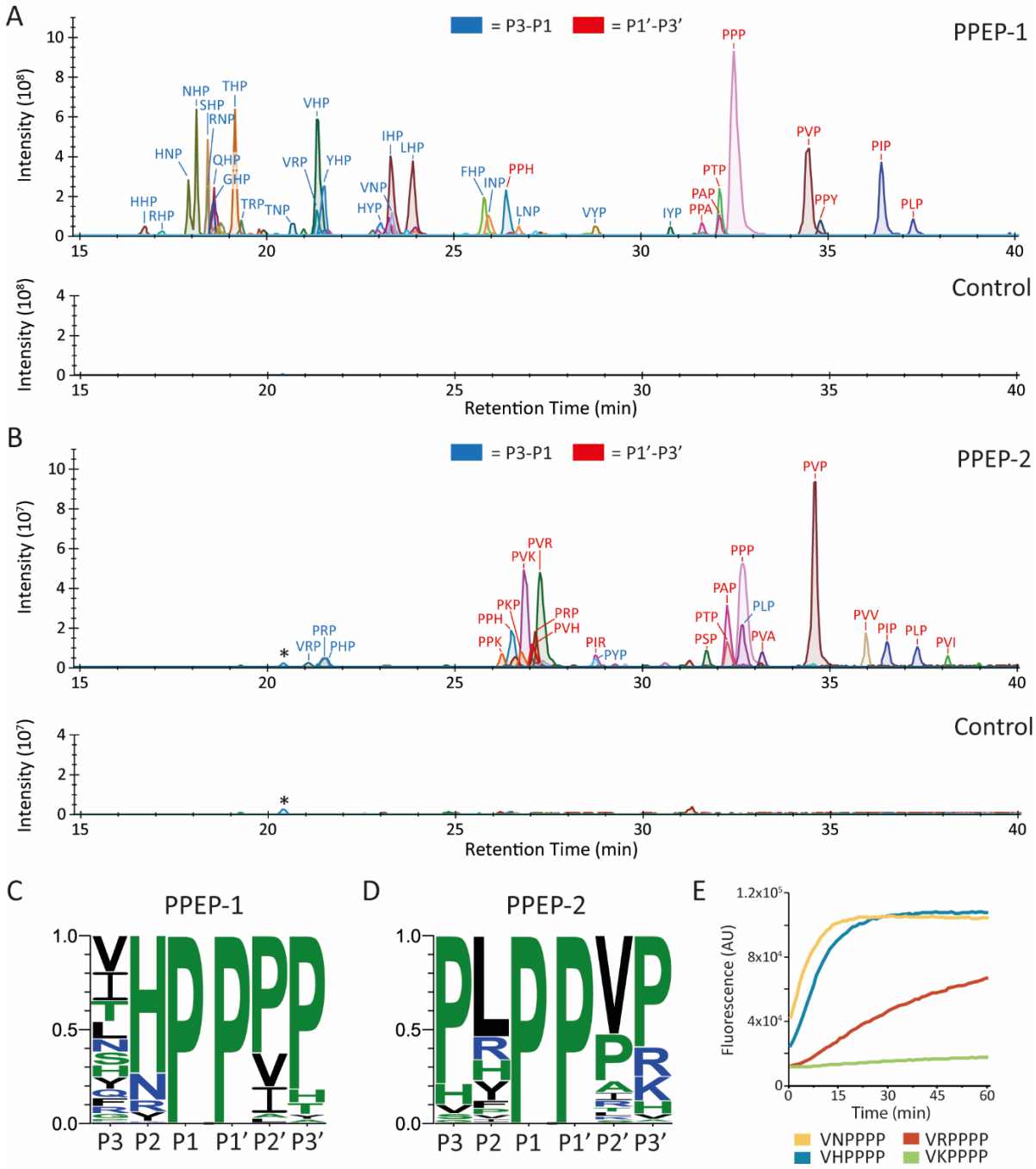


Figure 4. Incubation of the mixed non-prime- and prime-side libraries with PPEP-1 and PPEP-2. The non-prime- and prime-side libraries were mixed and the peptides were incubated with either **A)** PPEP-1 or **B)** PPEP-2. The product peptides were analyzed using LC-MS/MS. Results were searched against a database containing all 722 9-mer product peptides based on cleavage between the two fixed prolines, i.e., PTEDAVXXP and PXXGGLEEF (X≠Cys). The most abundant products that together account for >90% of the total abundance per library were used to create the EICs. Mass tolerance was set to 5 ppm. An untreated control sample was included. *A molecule corresponding to the mass of PTEDAVPHP (481.7325, $[M+2H]^{2+}$) was observed but MS/MS spectra indicate no PTEDAVXXP product peptide. **C) and D)** The results from the EICs were used to create a logo that displays the observed frequency of a residue at positions P3-P3' for PPEP-1 **(C)** and PPEP-2 **(D)**. **E)** Time course of PPEP-1 mediated cleavage of the synthetic FRET-quenched peptides Lys(Dabcyl)-EV(N/H/R/K)PPPPD-Glu(EDANS).

First of all, the new data for the prime-side specificity of PPEP-1 and PPEP-2 are consistent with our previously reported data [230], thereby demonstrating the excellent reproducibility of the method, even with the inclusion of the new combinatorial peptide library in the experiment (**Figures 4C,D)**.

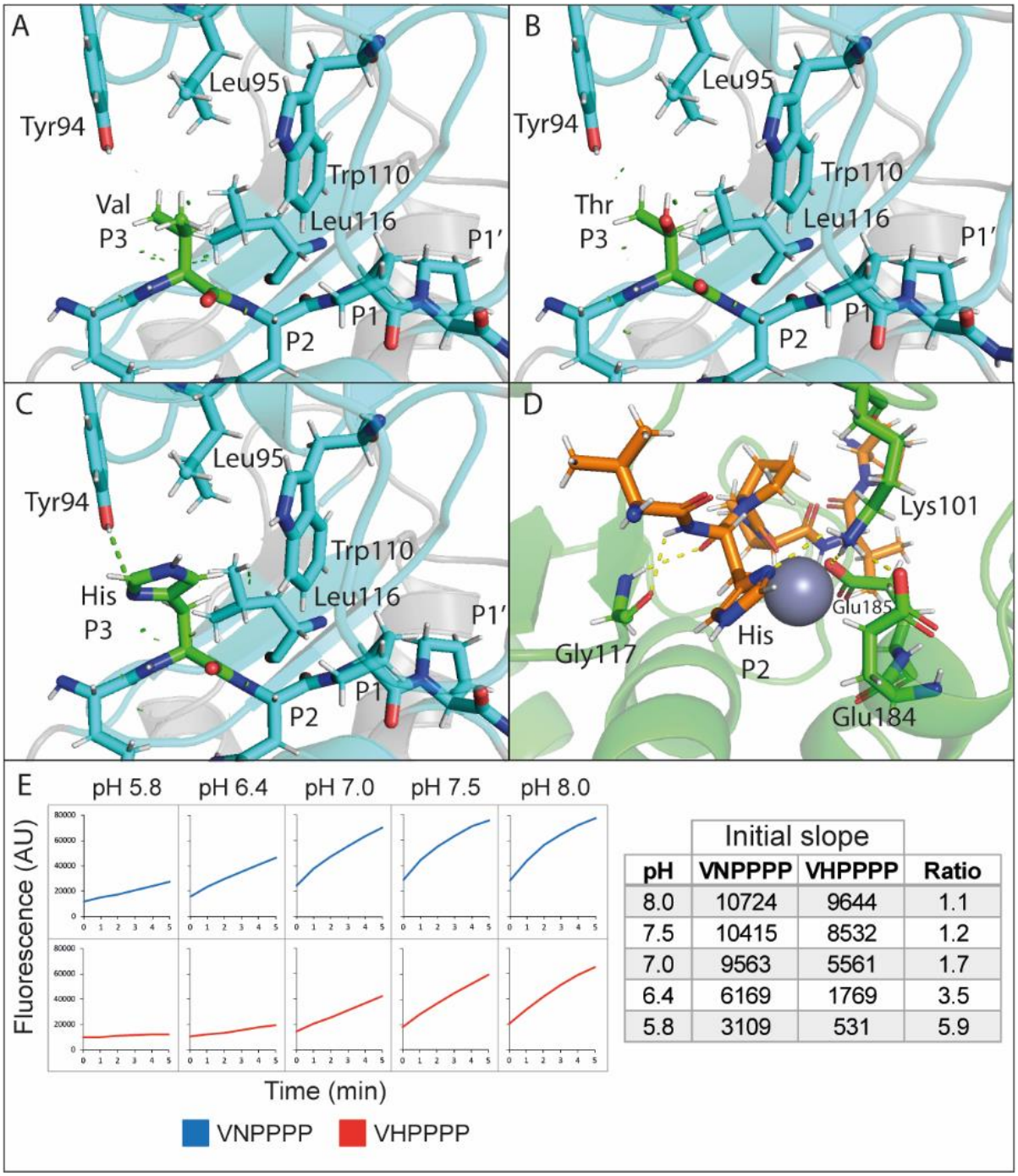
Notably, for PPEP-1, the logo highlights the variability of permissible residues at the P3 position. This observation aligns with biological expectations, since the endogenous substrates (CD2831 and CD3246) have a Val, Ile, or Leu at the P3 position, suggesting a lower stringency at this position for PPEP-1 activity in *C. difficile* [147]. In the context of an Asn at the P2 position, as is the case in the endogenous substrates of PPEP-1, product peptides containing TNP and HNP (although the intensity of this signal is influenced by the ESI response factor [230,266]) are among the highest signals aside from VNP (**Figures 2A and 4A**). In the PPEP-1 cocrystal with substrate VNPPVP (P3-P3'), the Tyr94, Leu95, Trp110, and Leu116 are found in close proximity to the Val (P3) and likely influence the P3 specificity (**Figure 5A**) [162]. Substitution of the Val (P3) with a Thr in the PPEP-1 cocrystal had no consequences, because of their similar sizes and the absence of attractive or repulsive polar interactions (**Figure 5B**). However, substituting the Val (P3) with His could produce polar interactions with the Tyr94 in PPEP-1 (**Figure 5C**), potentially strengthening the interaction between protease and substrate. In this configuration, the His side chain extended away from the P3 contacting residues, which might be a common mechanism to mitigate steric clashes and could explain the high variability of residues at the P3 position.

In line with the data presented above **(Figure 2)**, the PPEP-1 logo shows a high preference for His at the P2 position (**Figure 4C**). This is surprising since this residue is not observed at the P2 position in the endogenous substrates. The high abundance of His and the other basic amino acids (Arg and Lys) in the logos could relate to their ionization efficiency [266]. To assess how well His is tolerated at the P2 position compared to Asn and the other basic amino acids, a cleavage assay using FRET-quenched peptides was performed (**Figure 4E**). Although the His at the P2 position is well tolerated by PPEP-1, an Asn at that position produces the optimal substrate. In line with **Figure 4A**, VRP (P3-P1) is cleaved less efficiently than VHP, while VKP represents a very poor substrate.

The preference of PPEP-1 for Asn at the P2 stems from the interactions of the side chains of Lys101 with the side chains of Glu184, Glu185, and the Asn at the P2 position, collectively termed the KEEN interface [162]. Modeling of a PPEP-1 cocrystal with a substrate in which the Asn has been substituted for a His residue reveals that the histidine side chain can interact with Lys101 via hydrogen bonding as well (**Figure 5D**). In addition, the backbone atoms of the His interact with Gly117, similar to Asn. In the interaction as shown in **Figure 5D**, the electronegative nitrogen (N1) of His interacts with

the protonated Lys101. However, in the case of His protonation, this interaction might be lost. To test this, we incubated PPEP-1 with the substrate VHPPPP at various pH values (**Figure 5E**). At pH 5.8, we expected most of the His to be protonated since the pKa of the His side chain is 6.04. To see the effect of His protonation rather than the effect of a lower pH on the PPEP activity in general, we also included the substrate VNPPPP as a control. When comparing the initial slope of the reactions, the reactions of both substrates were similar at pH 7.5 and 8.0. However, by lowering the pH, VHPPPP cleavage became increasingly less efficient compared to VNPPPP, and at pH 5.8, the initial slope of the VHPPPP reaction was around 6x lower than that of VNPPPP. This indicated that protonation of the His at P2 inhibits cleavage by PPEP-1, likely due to the loss of the interaction with the Lys101.

Importantly, the results of our synthetic combinatorial peptide library approach demonstrate the ability to identify endogenous substrates using this method as **Figures 4B,D** clearly show the preference of PPEP-2 for PLPPVP (P3-P3'). When forming a hypothesis about the endogenous substrates without any other prior knowledge, a search for this motif in the proteome of *P. alvei* directly leads to the identification of the endogenous substrate VMSP of PPEP-2 [14]. Previous modeling of PPEP-2 with the endogenous substrate PLP↓PVP (P3-P3') predicted the Pro at the P3 to produce a kink in the polypeptide, thereby redirecting the upstream polypeptide away from the salt bridge formed by Glu113 and Arg145 [157]. The need for this diversion was supported by the data in **Figure 4D**, since the presence of a Pro at the P3 position was a strong determinant for proteolytic activity.



	Initial slope		
pH	VNPPPP	VHPPPP	Ratio
8.0	10724	9644	1.1
7.5	10415	8532	1.2
7.0	9563	5561	1.7
6.4	6169	1769	3.5
5.8	3109	531	5.9

Figure 5. Structural basis for the P3 and P2 specificity of PPEP-1. A) The cocrystal of PPEP-1 with substrate VNPPVP (cyan, PDB: 6R5C). The Val at the P3 position of the substrate is shown in green. The P3 contacting residues Tyr94, Leu95, Trp110, and Leu116 are shown as sticks. **B)** Substitution of the Val at the P3 position with Thr. No steric clashes or polar interactions were formed. **C)** Substitution of the Val at the P3 position with His. Polar interactions are shown as a green dotted line. Of note, a second rotamer in which the imidazole ring is rotated 180° is possible, but this produces a weaker interaction (hydrogen bonding distance=3.5 Å). **D)** Cocrystal structure of PPEP-1 (green, PDB: 5A0X) with a substrate peptide (orange) in which the Asn is substituted by a His. Yellow dotted lines indicate the interactions between residues. **E)** Time course of the PPEP-1 mediated cleavage at different pH with FRET-quenched peptides Lys(Dabcyl)-EV(H/N)PPPPD-Glu(EDANS) (left). The initial slope represents the increase in fluorescence in the first 5 min. The ratios of the initial slopes were calculated to compare both reactions (right).

Exchanging the β3/β4 loop of PPEP-1 and PPEP-2 shifts specificity towards one another

While PPEP-1 and PPEP-2 share a close relationship, their non-prime-side specificity exhibits marked differences. Among other structural elements, the β3/β4 loop plays a role in non-prime-side specificity [157]. Notably, this loop varies largely between the two proteases. Previous studies demonstrated that replacing the β3/β4 loop of PPEP-2 (^{112}SERV^{115}) with that of PPEP-1 (^{117}GGST^{120}) alters the enzyme's preference for the P3 position, shifting from Pro to Val and making it more PPEP-1-like [157].

Surprisingly, this effect was not mirrored in PPEP-1, as the mutant PPEP-1_{SERV} failed to cleave the tested peptides VNPPVP and PLPPVP. To gain a more comprehensive understanding of the significance of the β3/β4 loop for PPEP specificity, we conducted experiments using the combined non-prime- and prime side libraries incubated with PPEP mutants (PPEP-1_{SERV} and PPEP-2_{GGST}). EICs of the non-prime-side product peptides with VNP (PPEP-1 substrate), PLP (PPEP-2 substrate), and their combinations PNP and VLP, were generated (**Figure 6**).

As expected, for wild-type (WT) PPEP-1, VNP was the predominant product peptide. However, the mutant PPEP-1_{SERV} displayed a shift in specificity towards PPEP-2, evidenced by increased signals for PNP and PLP product peptides, whereas the specificity for VLP remained unchanged. Conversely, PPEP-2_{GGST} exhibited a similar shift towards WT PPEP-1 specificity, with a relative increase in signals for PNP, VNP, and VLP compared to PPEP-2. Notably, the original substrates continued to be favored, suggesting that the β3/β4 loop is not the primary determinant of non-prime-side specificity. Instead, residues Lys101 and Glu184 in WT PPEP-1 that are part of the KEEN interface [162] and interact with the Asn at the P2 position (aligning with residues Arg96 and T180 in PPEP-2), may play a more decisive role in P3 and P2 specificity. Additionally, the β3/β4 loop in PPEP-2 is stabilized by a salt bridge between residues Glu113 and Arg145 [157]. However, this stabilizing interaction might be absent in PPEP-1_{SERV} due to the substitution of Arg145 with His in WT PPEP-1. Consequently, the steric hindrance caused by this salt bridge might be absent in PPEP-1_{SERV}, potentially allowing a Val at the P3 position.

It's worth noting that PPEP-1_{SERV}, previously incapable of cleaving FRET-quenched peptides containing the VNPPVP and PLPPVP sequences [157], now demonstrated tolerance for both VNP and PLP at the non-prime-side. Prime-side specificity remained unaffected by the mutation, with PVP (P1'-P3') product peptides present post-incubation with PPEP-1_{SERV} (**Supplemental Table S1**). Discrepancies between current and previous results may be attributed to differences in the P4 position, where the FRET-quenched peptides featured a Glu, while the non-prime-side library had a Val at that position.

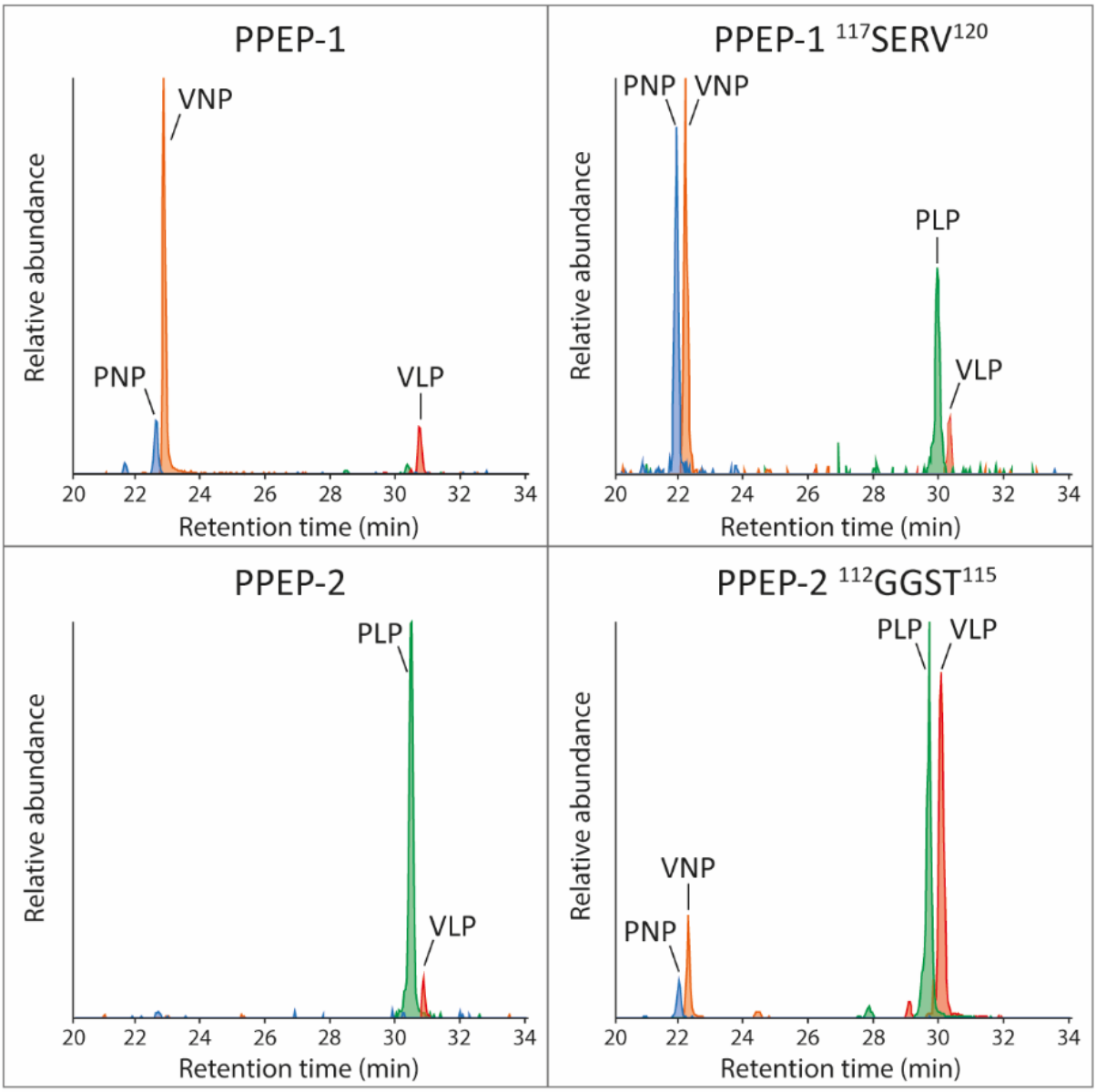


Figure 6. Altered non-prime-side specificity of PPEP-1$_{SERV}$ and PPEP-2$_{GGST}$. EICs were constructed for the product peptides containing the non-prime-side P3-P1 sequences VNP, PLP, VLP, and PNP after incubation of the mixed library with PPEP-1 (upper left panel), PPEP-2 (bottom left panel), PPEP-1$_{SERV}$ (upper right panel), or PPEP-2$_{GGST}$ (bottom right panel).

A second putative PPEP from *C. difficile* does not exhibit Pro-Pro endopeptidase activity

In *C. difficile*, a second PPEP-like protein, CD1597 (Gene: CD630_15970, UniProt ID: Q186F3), is present. Interestingly, CD1597 differs from the PPEPs that have hitherto been described in several ways. Apart from a PPEP-like domain, CD1597 contains an additional N-terminal domain that makes up about half of the protein. Moreover, in contrast to the other PPEPs, CD1597 is not predicted to contain a signal peptide for secretion. Furthermore, several amino acid insertions in CD1597 are observed in a sequence alignment with PPEP-1 and PPEP-2. However, these insertions are not found within the presumed active site of CD1597.

To assess the capability of CD1597 to hydrolyze Pro-Pro substrates, we conducted separate incubations of non-prime- and prime-side libraries with CD1597. Subsequent database searches aimed at identifying product peptides did not reveal the formation of any products. Therefore, to visualize the data, we constructed EICs for all the possible 9-mer product peptides (PTEDAVXXP and PXXGGLEEF) (**Figure 7**). In contrast to the previous experiments, the intensity of signals in the treated samples was comparable to those in the control samples, indicating a lack of proteolytic activity of CD1597. In some cases, however, a signal was exclusively observed in the treated sample, but manual inspection of the MS/MS spectra indicated no product peptides of any kind. To rule out the possibility of an inhibitory effect of the N-terminal domain that is absent in other PPEPs, experiments with only the predicted proteolytic domain were conducted but yielded similar results (data not shown). Additional investigations into CD1597, using both FRET-quenched and plain peptide cleavage assays, consistently revealed no cleavage (data not shown). In light of all these findings, we conclude that, despite its structural resemblance to a PPEP, CD1597 does not exhibit PPEP-like activity.

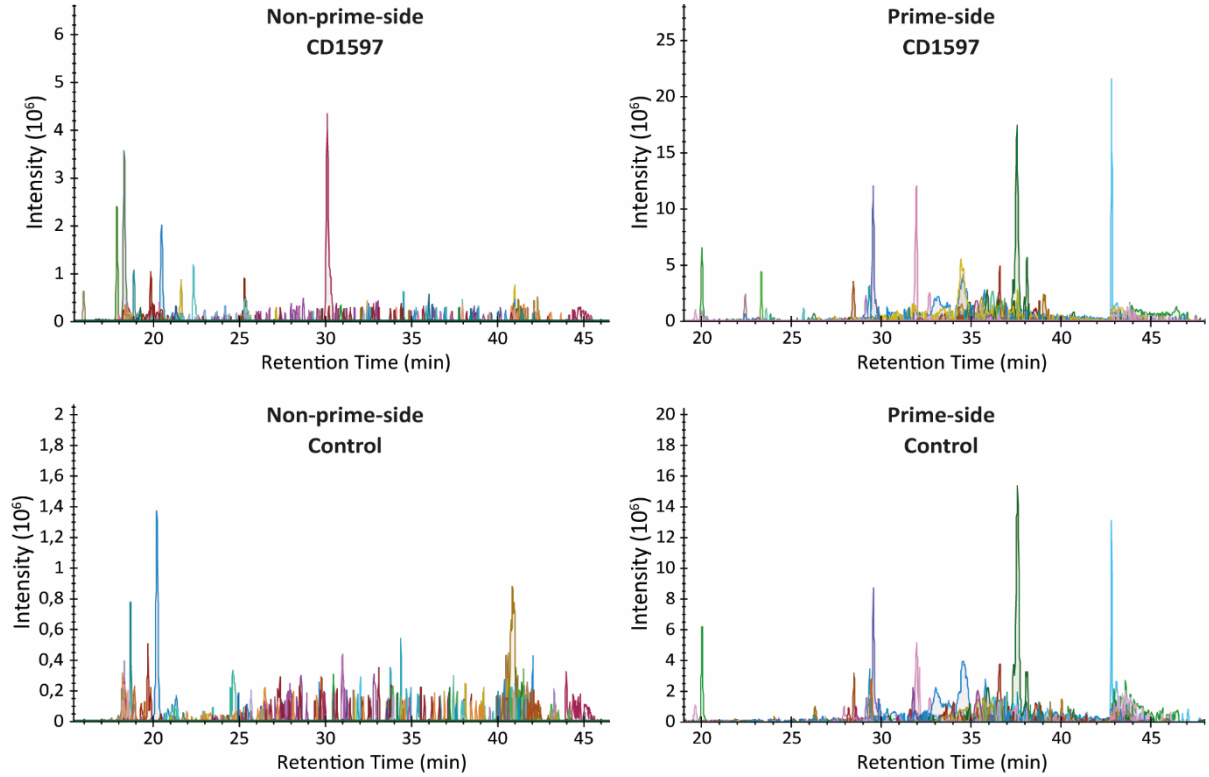


Figure 7. Incubation of the separate non-prime- and prime-side libraries with CD1597. The non-prime- and prime-side libraries were separately incubated with CD1597. EICs were constructed that include all possible 9-mer product peptides (PTEDAVXXP and PXXGLEEF).

The underlying cause of CD1597's inactivity remains elusive. Although we cannot rule out the possibility that we have purified an inactive recombinant protease, we find this unlikely given the previous work on other PPEPs for which we use similar purification routines. While the absence of activity could suggest CD1597 is a pseudoprotease, it's noteworthy that the protein harbors an intact catalytic HEXXH motif, which contradicts the typical characteristics of a pseudoprotease [221,222]. Interestingly, we have only identified a single peptide of CD1597 in two extensive proteome analyses of overnight cultures of *C. difficile* [233], proving that CD1597 has a very low abundance. However, CD1597 has been identified in larger quantities in the spore coat/exosporium layer of *C. difficile* spores [91]. In the context of *C. difficile* spores, two other pseudoproteases, namely CspA and CspB, have been identified, and they play crucial roles in spore germination [227]. The seemingly exclusive presence of CD1597 in spores could indicate a potential involvement in spore germination, either as a pseudoprotease or as a zymogen that requires specific stimuli for activation. Further exploration is needed to unravel the precise role of CD1597 in the context of *C. difficile* spores and to elucidate the factors influencing its apparent inactivity in our library assays.

Specificity profiling of a novel PPEP from *Anoxybacillus tepidamans*

Bioinformatic analysis predicted the presence of a PPEP homolog in the thermophilic bacterium *Anoxybacillus tepidamans* (gene: HNQ34_002771, UniProt ID: A0A7W8IRZ3) [148]. This protein, hereafter designated as PPEP-4, showed a close phylogenetic relationship to the previously described PPEP-3 from *Geobacillus thermodenitrificans* [148]. In fact, *A. tepidamans* was initially proposed to be a member of the genus *Geobacillus*, further demonstrating the close relationship of these organisms [287].

Prior investigations into the prime-side specificity of PPEP-3 indicated a strong preference for prolines at the P2' and P3' positions, in contrast to the more versatile P2' specificity of PPEP-1 and -2 (**Figure 4** and [230]). The primary structure of PPEP-4 very closely resembles that of PPEP-3, although some differences are observed (**Figure 8A**). Notably, the substitution of Phe190 in PPEP-3 with Leu in PPEP-4, similar to PPEP-1, stood out. In PPEP-1, this residue is close to the P2' Val in the substrate VNP↓PVP [162]. Likely, the Phe190 in PPEP-3 causes steric hindrance for a Val and most other residues at the P2' position (**Figure 8B**) and causes the need for a Pro to redirect the substrate elsewhere.

We hypothesized that the substitution of Phe190 in PPEP-3 to Leu in PPEP-4 could render the P2' specificity of PPEP-4 more akin to that of PPEP-1. To test this, we profiled PPEP-4 specificity using the combinatorial peptide libraries. Following LC-MS/MS analysis, a database search was performed using the database containing all possible 9-mer product peptides (PTEDAVXXP and PXXGGLEEF). In the experiments described above, we only included the most abundant product peptides that together accounted for >90% of the total intensity. However, in the case of PPEP-4, it became apparent that the non-prime-side specificity was highly diverse, and that the >90% cut-off value included too many substrates for meaningful visualization while many of them were very minor substrates. For clarity, we therefore decided to include only product peptides that represented at least 1% of the total intensity for both the non-prime- and prime-side product peptides. Using these product peptides, an EIC and logo were created to depict the P3-P3' specificity of PPEP-4 (**Figure 9B**).

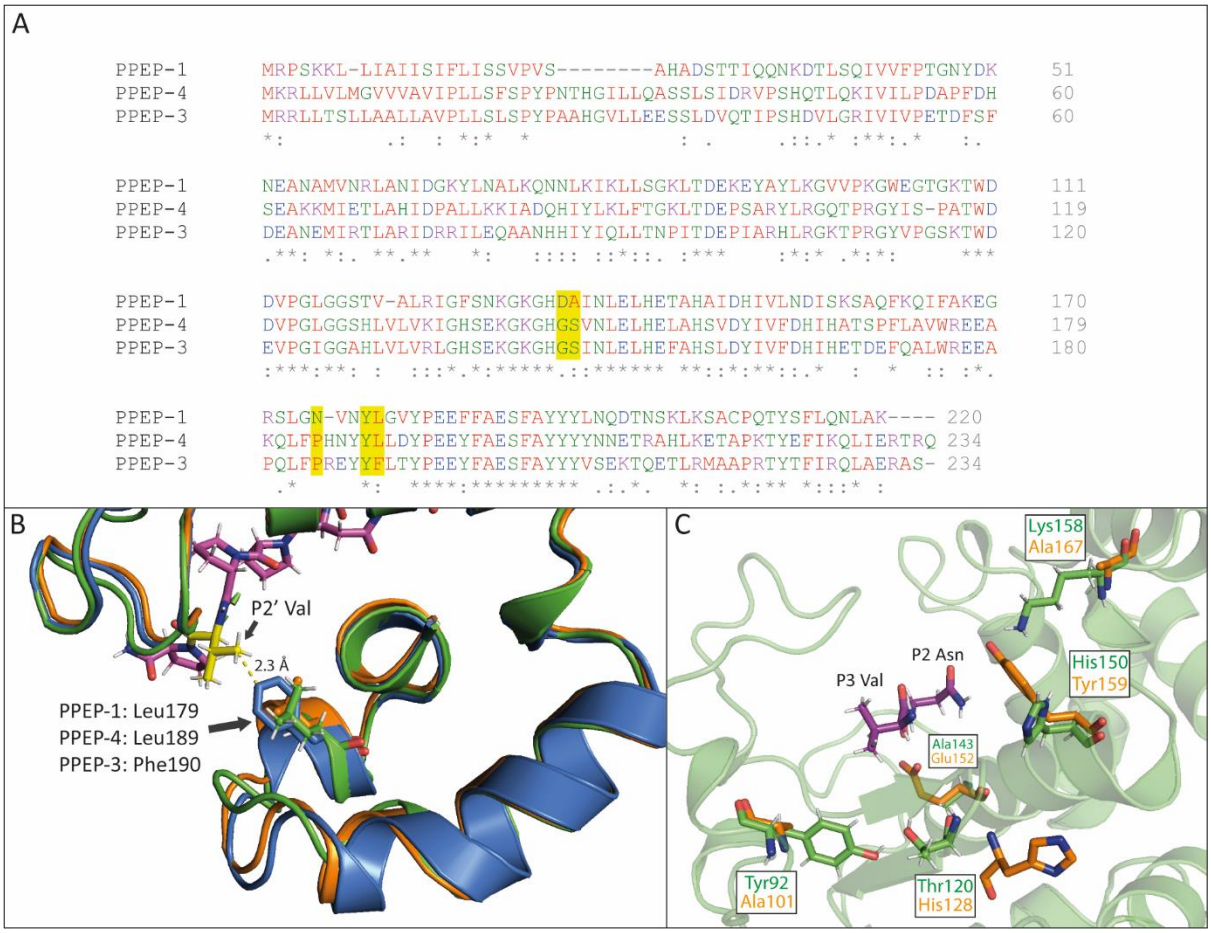


Figure 8. Comparison of PPEP-4 from *Anoxybacillus tepidamans* with PPEP-1 and PPEP-3. A) Sequence alignment of PPEP-1, PPEP-3, and PPEP-4 created using the Clustal Omega multiple sequence alignment tool. The residues in PPEP-1, and their corresponding residues in PPEP-3 and -4, that contact the Val at the P2' in the substrate peptide VNPPVP are highlighted in yellow. **B)** Structural comparison of co-crystal of PPEP-1 (green, PDB: 6R5C) with its substrate VNPPVP (magenta, Val=yellow) and the predicted structures of PPEP-3 (blue) and PPEP-4 (orange). **C)** Comparison of the residues within 5 Å of the P3 Val and P2 Asn that differ biochemically between PPEP-1 (green) and the predicted structure of PPEP-4 (orange).

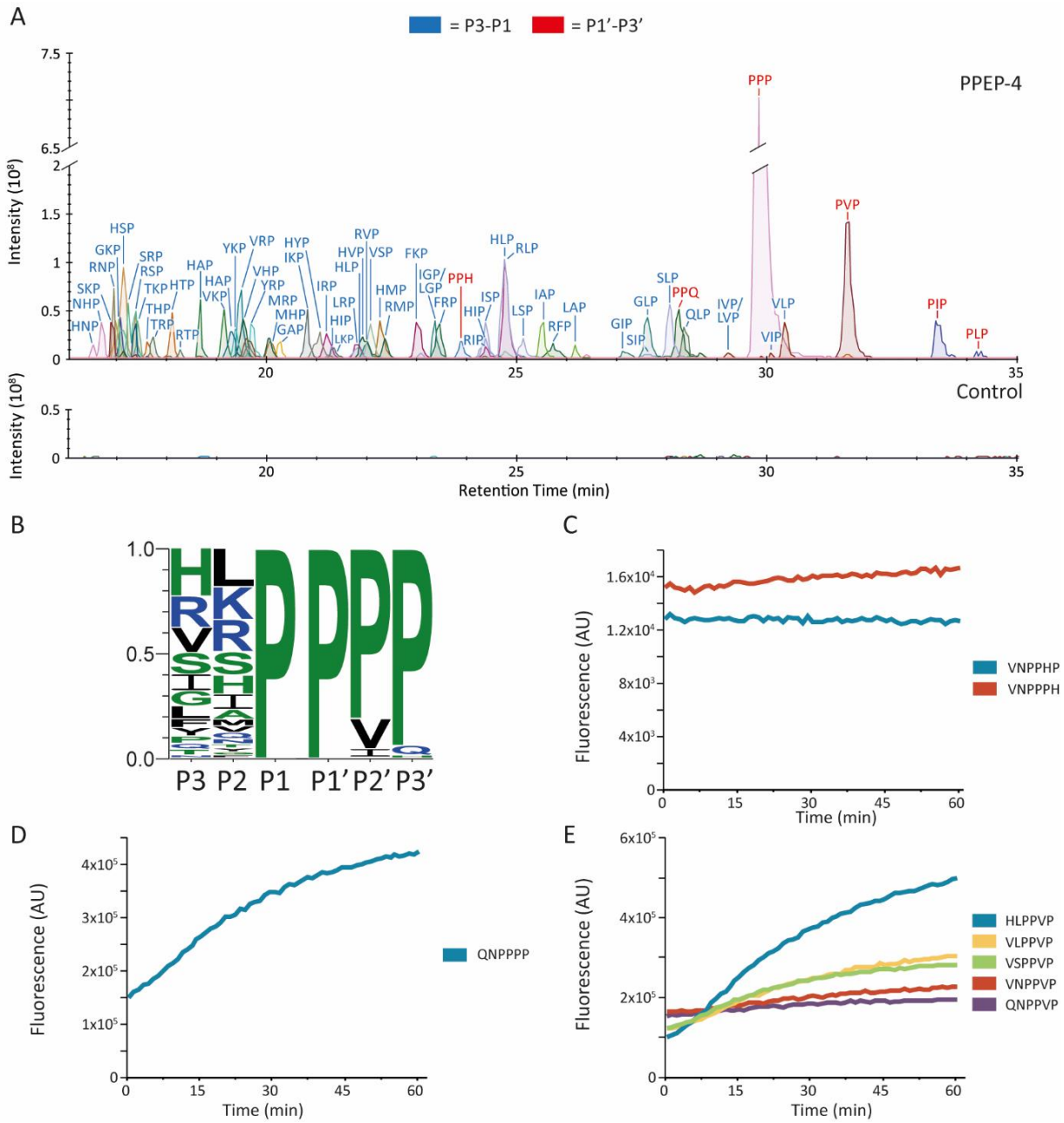


Figure 9. P3-P3' specificity of PPEP-4 from *Anoxybacillus tepidamans*. A) The non-prime- and prime-side libraries were mixed and the peptides were incubated with PPEP-4. The product peptides were analyzed using LC-MS/MS and a database search was performed to identify and quantify the products. Results were filtered for 9-mer product peptides and the products accounting for >1% of the total abundance were used to create the EICs. Mass tolerance was set to 5 ppm. An untreated control sample was included. **B)** The results from the EIC were used to create a logo that displays the observed frequency of a residue at positions P3-P3' for PPEP-4. **C)** Time course of PPEP-4 mediated cleavage of the synthetic FRET-quenched peptides Lys(Dabcyl)-EVNPPHPD-Glu(EDANS) and Lys(Dabcyl)-EVNPPPHD-Glu(EDANS). **D)** Time course of PPEP-4 mediated cleavage of the synthetic FRET-quenched peptide Lys(Dabcyl)-EQNPPPP-Glu(EDANS). **E)** Time course of PPEP-4 mediated cleavage of the synthetic FRET-quenched peptides Lys(Dabcyl)-E(QN/VN/VL/HL/VS)PPVP-Glu(EDANS).

For the prime-side specificity of PPEP-4, PPP (P1'-P3') is by far the most abundant product peptide, but also the PVP, PIP, and PLP product peptides are formed (**Figure 9A**), resembling a PPEP-1-like specificity profile [230]. Since the substitution of the Phe190 in PPEP-3 to the Leu in PPEP-4 is the only difference between the P2' contacting residues, we find it likely that the residue at this position is a large determinant for P2' specificity. Furthermore, a product peptide containing the PPQ sequence (P1'-P3') was found for PPEP-4, characteristic of PPEP-3 but not PPEP-1 [230]. Lastly, PPH (P1'-P3') is tolerated by PPEP-4, similar to both PPEP-1 and PPEP-3. Initially, MS/MS spectra were inconclusive on the identity of the product peptide, i.e., it was either PPH or PHP. However, incubation of PPEP-4 with the FRET-quenched peptides containing VNPPPH and VNPPHP demonstrated that PPEP-4 exclusively cleaves VNPPPH, similar to PPEP-1 (**Figure 9C**) [230]. Collectively, although PPEP-4 is highly similar to PPEP-3, PPEP-4's prime-side specificity shows characteristics of both PPEP-1 and PPEP-3.

Although PPEP-1 tolerates many residues at the P3 and P2 positions (**Figures 4A,C**), PPEP-4 displays even greater flexibility (**Figures 9A,B**). A comparison of the residues that are proximal to the P3 and P2 position in the PPEP-1 co-crystal with substrate VNPPVP and that differ biochemically between PPEP-1 and PPEP-4 (Alphafold model) revealed several residues that can influence P3-P2 specificity (**Figure 8C**). First, Tyr92 in PPEP-1 is substituted by an Ala in PPEP-4. In addition, Thr120 is substituted by a His that is predicted to be further removed from the active site. Together, these differences might reduce the steric hindrances at the P3 (and possibly P4) positions. Furthermore, several other residues in PPEP-1 that are close to the P3-P2 residues are substituted in PPEP-4. Although these substitutions do not seem to reduce steric hindrances, their different biochemical properties might influence the intra- and intermolecular interactions that make the non-prime-side specificity of PPEP-4 more permissible. Further investigations of the structure-function relationships in PPEPs necessitate additional co-crystal structures.

A search for candidate substrates in the proteome of *A. tepidamans* focused on the motifs P↓PPP, P↓PVP, P↓PIP, P↓PPH, and P↓PPQ (↓=supposed cleavage site), resulting in the identification of 14 candidate substrates. Given the predicted secretion of PPEP-4, an analysis of these 14 candidates using SignalP 6.0 identified a single secreted protein, a predicted lipoprotein (gene: HNQ34_001056, UniProt ID: A0A7W8IQP4). This lipoprotein is predicted with high confidence to possess a Sec/SPII signal sequence for integration in the lipid membrane and contains its putative cleavage site (QNP↓PPP) close to the location of lipid insertion. Although a FRET-quenched peptide with the core sequence QNPPPP is cleaved, proteolysis was incomplete after 1 h of incubation (**Figure 9D**). A comparison with other peptides in our collection that were selected based on their high abundance at the non-prime-sides in the logo in **Figure 9B** shows that QNP

(P3-P1) is indeed poorly tolerated (**Figure 9E**). However, the endogenous substrate of PPEP-4 does not necessarily need to possess the most optimal cleavage sequence, as is the case for PPEP-1 [146].

PPEP-1 and PPEP-2 are involved in adhesion/motility by cleaving large adhesive surface proteins, thereby releasing the cells. The lipoprotein in *A. tepidamans* differs from these substrates due to its small size and the lack of any predicted domains (or any other structural elements aside from the signal peptide). Another surface protein that does possess adhesion domains in the same organism and that contain PPEP-like cleavage motifs is the Penicillin-binding protein 1A (HNQ34_000435, UniProt ID: A0A7W8IMR6). The Penicillin-binding protein 1A is predicted to possess a transpeptidase domain, but also a Fibronectin type III (FN3) domain, which is known to be capable of binding components of the extracellular matrix, integrins, and possibly carbohydrates [288,289]. Interestingly, this protein contains a putative PPEP cleavage site directly upstream of the FN3 domain (EQPPAP) and two putative cleavage sites downstream of the FN3 domain (PTPPAP and TNPPAP), although these sites seem to represent poor substrates under our experimental conditions. Additional experiments such as bacterial surface-shaving [290] of *A. tepidamans* with PPEP-4 could provide more insight into the endogenous substrates of this member of the PPEP family.

Conclusion

We developed an approach to characterize both the non-prime- and prime-side-specificity of PPEPs by combining the use of synthetic combinatorial peptide libraries with LC-MS/MS. Using this method, we deepened our understanding of the specificity of the previously characterized PPEP-1 and PPEP-2. Importantly, we were able to identify PPEP-2's endogenous substrate sequence PLPPVP as the optimal substrate using our library method. In addition, we profiled the specificity of a novel PPEP from *A. tepidamans*, which we termed PPEP-4. Based on structural comparisons of PPEP-4 with other PPEPs, we predicted a P2' specificity that resembles that of PPEP-1, which was confirmed by our data. Moreover, investigation of mutants of PPEP-1 and PPEP-2 that had their β3/β4 loop swapped showed that the non-prime-side specificity shifted towards each other, demonstrating the involvement of this loop in determining substrate specificity. For a second putative PPEP from *C. difficile*, however, no Pro-Pro endoproteolytic activity was observed. Finally, after including the non-prime-side peptide library, the prime-side profiles of PPEP-1 and -2 were in line with previously reported data, thereby demonstrating that the synthetic combinatorial library approach is a robust method with excellent reproducibility.

Experimental procedures

Expression and purification of recombinant PPEPs

PPEP-1, PPEP-2, PPEP-1_{SERV}, and PPEP-2_{GGST} were expressed and purified as previously described [146,157]. For the expression of PPEP-4, a pET-16b vector containing an *E. coli* codon-optimized 10xHis-PPEP-4 (lacking the signal peptide) construct was ordered from GenScript (Rijswijk, The Netherlands). The pET-16b 10xHis-PPEP-4 plasmid was transformed to *E. coli* strain C43 and PPEP-4 expression was induced using 1 mM IPTG for 4 h. Bacterial pellets were resuspended in lysis buffer (20 mM NaH_2PO_4, 500 mM NaCl, 40 mM imidazole, pH 7.4) and lysozyme was added to the suspension to 1 mg/mL and incubated for 30 min on ice before disruption by 5 30 s rounds of sonication. The lysates were loaded onto a 1 mL HisTrap HP column (GE Healthcare, Chicago, Illinois, United States) coupled to an ÄKTA Pure FPLC system (GE Healthcare). The column was washed using wash buffer (20 mM NaH_2PO_4, 500 mM NaCl, 40 mM imidazole, pH 7.4) and 10xHis-PPEP-4 was eluted using a linear gradient with elution buffer (20 mM NaH_2PO_4, 500 mM NaCl, 500 mM imidazole, pH 7.4). Buffer exchange to PBS was performed using an Amicon® Ultra-15 Centrifugal Filter Unit (Merck Life Science NV, Amsterdam, The Netherlands) with a 10 kDa cut-off membrane.

For CD1597 (and the predicted catalytic domain, AA 211-416), a pET-28a vector containing an *E. coli* codon-optimized 6xHis-CD1597 (lacking the signal peptide) construct was ordered from Twist Bioscience (San Francisco, California, United States). Expression and purification were performed as described above but with different lysis (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0), wash (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 8.0), and elution (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, pH 8.0) buffers.

Combinatorial peptide library assays

The combinatorial peptide libraries were synthesized and assays were performed as previously described [230]. In short, approximately 10 nmol of precleaned (on avidin column) peptide mixture was incubated with 200 ng PPEP for 3 h at 37 °C in PBS. For PPEP-1_{SERV} and PPEP-2_{GGST}, 500 ng was used in combination with an incubation time of 16 h. A non-treated control was included. After incubation, the samples were loaded onto an in-house constructed column consisting of a 200 µL pipet tip containing a filter and a packed column of 100 µL of Pierce High Capacity Streptavidin Agarose beads (Fisher Scientific, Landsmeer, The Netherlands), which was washed four times with 150 µL of PBS before use), to remove the biotinylated peptides. The flow-through and four additional washes with 125 µL PBS were collected. The product peptides were desalted

using reversed-phase solid-phase extraction cartridges (Oasis HLB 1 cm3 30 mg, Waters) and eluted with 400 µL of 30% acetonitrile (v/v) in 0.1% formic acid. Samples were dried by vacuum concentration and stored at −20 °C until further use. For the peptide library assays in which the non-prime- and prime-side libraries were combined, approximately 5 nmol of each library was used (10 nmol in total).

LC-MS/MS analyses

For the analyses of the product peptides of P3=Val non-prime-side sublibrary after incubation with PPEP-1 and those of the non-prime-side library after incubation with PPEP-1 and -2, product peptides were analyzed as previously described [230] by online C18 nanoHPLC MS/MS with an Ultimate3000nano gradient HPLC system (Thermo, Bremen, Germany), and an Exploris480 mass spectrometer (Thermo). Peptides were injected onto a precolumn (300 µm × 5 mm, C18 PepMap, 5 µm, 100 A), and eluted via a homemade analytical nano-HPLC column (30 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm, 120 A; Dr. Maisch, Ammerbuch, Germany). The gradient was run with a gradient of 2% to 36% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 52 min. The nano-HPLC column was drawn to a tip of ∼10 µm and acted as the electrospray needle of the MS source. The mass spectrometer was operated in data-dependent MS/MS mode for a cycle time of 3 s, with HCD collision energies at both 17V and 23V and recording of the MS2 spectrum in the Orbitrap, with a quadrupole isolation width of 1.2 m/z. In the master scan (MS1) the resolution was 120,000, the scan range 350–1600, at standard AGC target at a maximum fill time of 50 ms. A lock mass correction on the background ion m/z = 445.12003 was used. Precursors were dynamically excluded after n = 1 with an exclusion duration of 10 s and with a precursor range of 10 ppm. Charge states 1–5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 30,000 at an AGC target of 100% @maximum fill time of 60 ms.

For the analyses of the product peptides of the mixed non-prime- and prime-side libraries following incubation with PPEPs (and separate analyses for CD1597), product peptides were analyzed as described previously with minor adjustments [230,291] by online C18 nano-HPLC MS/MS with a system consisting of an Easy nLC 1200 gradient HPLC system (Thermo) and an Orbitrap Fusion LUMOS mass spectrometer (Thermo). Peptides were injected onto a homemade precolumn (100 µm × 15 mm; Reprosil-Pur C18-AQ 3 µm, Dr Maisch, Ammerbuch, Germany) and eluted via a homemade analytical nano-HPLC column (30 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm). The gradient was run from 2% to 40% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 52 min. The nano-HPLC column was drawn to a tip of ∼5 µm and acted as the electrospray needle of the MS source. The LUMOS mass spectrometer was operated in data-

dependent MS/MS mode for a cycle time of 3 s, with HCD collision energies at 20 V, 25V, and 30V and recording of the MS2 spectrum in the orbitrap, with a quadrupole isolation width of 1.2 m/z. In the master scan (MS1) the resolution was 120,000, the scan range 350–1600, at an AGC target of 400,000 at a maximum fill time of 50 ms. A lock mass correction on the background ion m/z = 445.12003 was used. Precursors were dynamically excluded after n = 1 with an exclusion duration of 10 s and with a precursor range of 10 ppm. Charge states 1–5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 30,000 at an AGC target of 100% @maximum fill time of 60 ms.

LC-MS/MS data analysis

To identify product peptides in a database search after analysis of the non-prime-side library with PPEP-1 and -2, we generated a database containing all 130,321 possible peptides in the library, i.e., PTEDAVXXPPXXE-Ahx-Ahx-K (biotin was included as a variable modification in the database searches). The Ahx in all peptide sequences was replaced by a Leu (they have an identical mass). For the identification of product peptides after analysis of the mixed non-prime- and prime-side libraries, a database was generated containing all 9-mer product peptides that are possible based on Pro-Pro cleavage (i.e., PTEDAVXXP and PXXGGLEEF).

The post-analysis process was performed as previously described [230]. Raw data were converted to peak lists using Proteome Discoverer version 2.5.0.400 (Thermo Electron, Waltham, Massachusetts, United States) and submitted to the in-house created databases using Mascot v. 2.2.7 (www.matrixscience.com) for peptide identification, using the Fixed Value PSM Validator. Mascot searches were with 5 ppm and 0.02 Da deviation for precursor and fragment mass, respectively, and no enzyme specificity was selected. Biotin on the protein N-terminus was set as a variable modification.

The database search results were filtered for product peptides that contained either PTEDAV or GGLEEF, were 9 residues in length, and contained no biotin. The resulting peptide lists were transported to Microsoft Excel, where duplicate masses and corresponding abundances were removed (e.g., the abundances of isomers PLPGGLEEF and PIPGGLEEF are listed twice, while this abundance is the total abundance of the two variants). The most abundant product peptides that together accounted for >90% of the total abundance were selected for further analysis (except for PPEP-4; see Results and Discussion). Further analysis was performed in Skyline 23.1.0.268 by importing the product peptides as FASTA along with the raw data files [292]. The Extracted Ion Chromatograms (EICs) displaying the product peptides were created by plotting the intensities of the signals corresponding to the monoisotopic m/z values of both 1+ and

2+ charged peptides with a mass tolerance of 5 ppm. Peptide annotation in Skyline was refined by manual inspection of MS/MS spectra and peak areas were exported from Skyline and used to create the sequence logos using WegLogo 3.7.12 [293].

FRET peptide cleavage assays

Time course kinetic experiments with PPEPs were performed as previously described [230] using fluorescent FRET-quenched peptides. FRET peptides consisted of Lys(Dabcyl)-EXXPPXXD-Glu(EDANS), in which each X varied between the different peptides tested. Proteolysis of FRET peptides by PPEPs was tested in 150 µL PBS containing 50 mM FRET peptide and 200 ng enzyme. Peptide cleavage was measured using the Envision 2105 Multimode Plate Reader. Fluorescence intensity was measured each minute for 1 h, with 10 flashes per measurement. The excitation and emission wavelengths were 350 nm and 510 nm, respectively. For the assay at different pH, buffers were prepared by mixing 0.2 M NaH_2PO_4 and 0.2 M Na_2HPO_4 and adding dH_2O to dilute the buffer 2x. The resulting buffers had a pH of 5.8, 6.4, 7.0, 7.5, and 8.0.

Bioinformatic analyses

Signal peptide predictions were performed using SignalP 6.0 [269]. Sequence alignments were performed using the Clustal Omega Multiple Sequence Alignment tool [272]. The predicted structures of PPEP-3 and PPEP-4 were retrieved from the Alphafold DB (https://alphafold.ebi.ac.uk/). Structures were analyzed using PyMOL (The PyMOL Molecular Graphics System, Version 2.5.5 Schrödinger, LLC).

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [200] via the PRIDE [201] partner repository with the dataset identifier **PXD050236**.

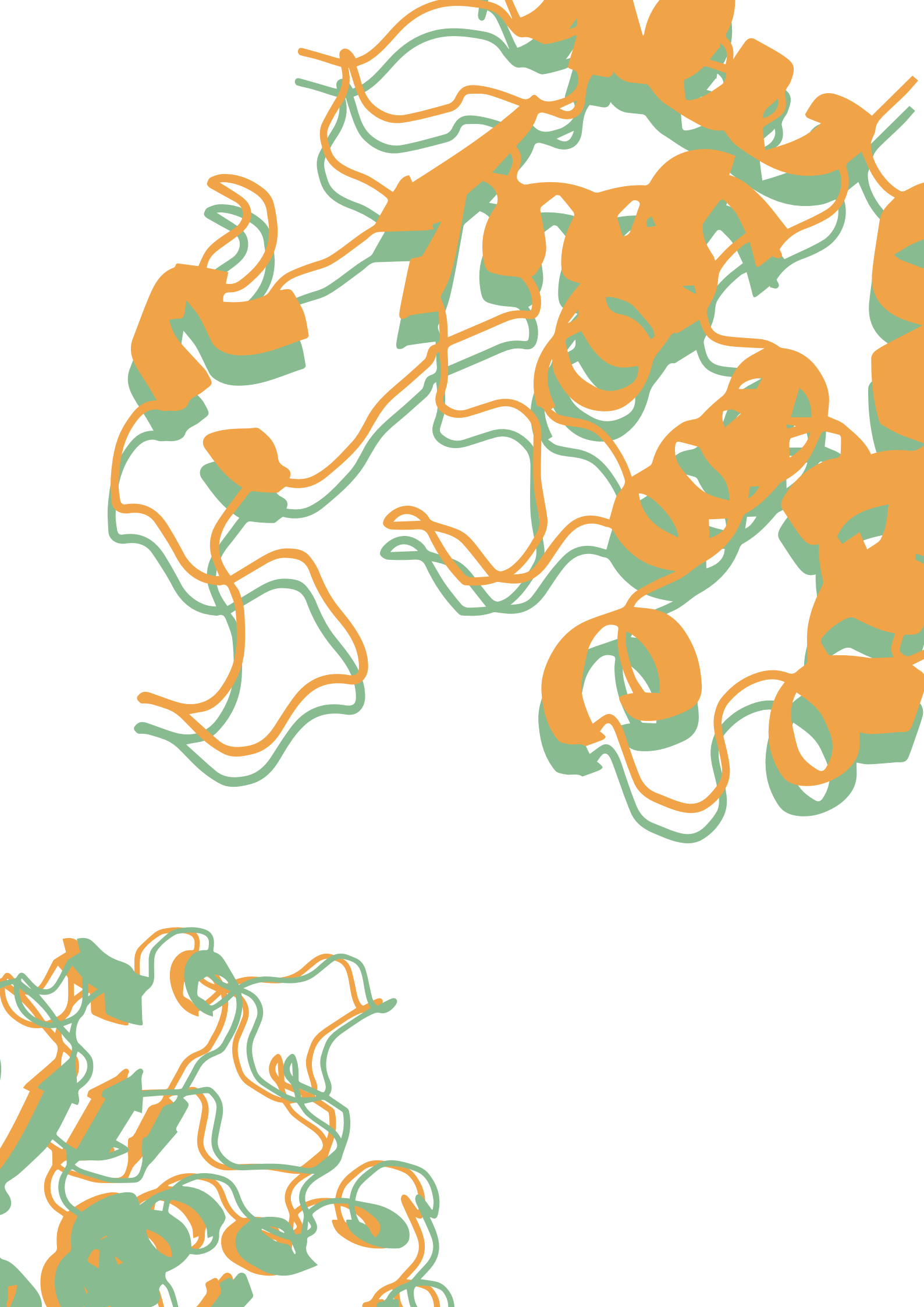
Acknowledgements

This research was supported by an ENW-M Grant (OCENW.KLEIN.103) from the Dutch Research Council (NWO) and by the research program Investment Grant NWO Medium with project number 91116004, which is (partially) financed by ZonMw. We thank Oleg I. Klychnikov and Stephen D. Weeks for the cloning and purification of PPEP-2, PPEP-1_{SERV}, and PPEP-2_{GGST}.

Supporting information

Table S1: The supplemental table can be found online

https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17160



6

Biochemical and structural characterization of PPEP-3 from *Geobacillus thermodenitrificans*

Bart Claushuis[1§], Fabian Wojtalla[2§], Hans C. van Leeuwen[3], Jeroen Corver[4], Ulrich Baumann[2¶], Paul J. Hensbergen[1¶]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] Department of Chemistry, Institute of Biochemistry, University of Cologne, Cologne, 50674, Germany
[3] Department of CBRN Protection, Netherlands Organization for Applied Scientific Research TNO, Rijswijk, 2280 AA, The Netherlands
[4] Leiden University Center of Infectious Diseases, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[§] and [¶]: Authors contributed equally

Manuscript in preparation

Abstract

The members of the group of Pro-Pro endopeptidases (PPEPs) are secreted bacterial endoproteases that display a unique preference to hydrolyze their substrates between two proline residues. The active site cleft of PPEPs accommodates six substrate residues, the P3-P3' residues, and the interactions between the protease and these substrate residues determine PPEP-3 specificity. In this study, we present the unbound and substrate-bound structures of PPEP-3 from the thermophilic bacterium *Geobacillus thermodenitrificans*. We describe the interactions in the protease-substrate complex on an atomic level. Most notably Tyr-112 and Phe-190, which differ from the corresponding residues in PPEP-1, greatly influence the P2 and P2' specificity, respectively. In addition, we characterized the substrate specificity in detail using synthetic combinatorial peptide libraries in combination with LC-MS/MS analyses. By correlating the substrate specificity profile to the structure, we explore the various mechanisms that determine PPEP-3 specificity and highlight differences with other PPEPs.

Introduction

The Pro-Pro endopeptidases (PPEPs) form a group of bacterial zinc metalloproteases characterized by the unique specificity to hydrolyze the peptide bond between two proline residues. In addition, PPEPs are extracellular proteases, either secreted in the environment or attached to the cell wall through additional domains [148]. The first identified PPEP, PPEP-1 from the human pathogen *Clostridioides difficile,* acts as a switch between adhesion and motility by cleaving two adhesion proteins [146,147]. The second characterized PPEP, PPEP-2, is believed to play a similar role in *Paenibacillus alvei* [157]. For both these PPEPs, the endogenous substrates are encoded by genes adjacent to the PPEP gene. In the case of two other PPEPs, PPEP-3 from *Geobacillus thermodenitrificans* and PPEP-4 from the closely related organism *Anoxybacillus tepidamans*, no endogenous substrates and function have been identified so far [206,230]. Interestingly, a PPEP homolog from *C. difficile*, CD1597, possesses a PPEP-like domain but exhibits no (Pro-Pro) proteolytic activity [206,294].

Previously, atomic structures have been experimentally determined for PPEP-1 and PPEP-2 [157,160–162]. Overall, these structures display highly similar structural elements. The proteases consist of an N-terminal (NTD) and C-terminal domain (CTD), which are divided by an active site helix containing the HExxH motif of metalloproteases [157,160]. For PPEP-1, cocrystal structures in complex with substrate peptides have been resolved [160,162]. In these cocrystals, the substrate binds in a double-kinked conformation that is produced by X-Pro bonds in the peptide [160]. This conformation is necessitated by a structural element called the diverting loop that otherwise restricts the substrate from exiting the active site cleft and therefore greatly impacts PPEP specificity [160]. Another important structural feature is the flexible S-loop which closes upon substrate binding and thereby covers a part of the active site cleft [160,162].

The active site cleft of PPEPs accommodates six substrate residues. Due to this, the substrate specificity depends on the six residues surrounding the cleavage: P3, P2, P1, P1’, P2’, and P3’ according to the nomenclature developed by Schechter and Berger [17]. Previously, we developed a method to characterize PPEP specificity in detail using synthetic combinatorial peptide libraries and liquid chromatography combined with tandem mass spectrometry (LC-MS/MS) [206,230]. Using this method, the complete specificity has been profiled for PPEP-1, PPEP-2, and PPEP-4 [206], while for PPEP-3 only the prime-side specificity has been profiled [230]. A remarkable feature of the prime-side specificity of PPEP-3 is the preference for all prolines at the P1’-P3’ positions, whereas other PPEPs display more variability [206,230]. A detailed understanding of the substrate specificity of PPEPs in combination with substrate-bound protease structures allows us to describe the structure-function relationship at an atomic level. However, to

identify both the general mechanisms and the unique determinants of PPEP specificity, additional cocrystal structures are needed.

In this study, we resolved the atomic structure of PPEP-3 from *G. thermodenitrificans*. In addition, we performed cocrystallization experiments with a known substrate peptide. By employing synthetic combinatorial peptide libraries combined with LC-MS/MS analyses, we were able to characterize both the non-prime- and prime-side specificity of PPEP-3. By combining the structural and specificity data, we shed light on the structure-function relationship of PPEP-3.

Results

Atomic structure of PPEP-3

The wild-type PPEP-3 structure was determined at 1.7 Å resolution in the tetragonal space group $P4_12_12$ with two monomers in the asymmetric unit (**Supplemental Table S1** and **Supplemental Figure S1**). The two crystallographically independent copies are virtually identical (RMSD of about 0.4 Å) except the termini: in chain A there are eight more residues resolved at the N-terminus (thus starting at residue Pro 20) and at the C-terminus there are two additional residues visible (up to Ser234) in the electron density map. Like PPEP-1 and PPEP-2, the overall PPEP-3 structure consists of an N-terminal (NTD) and a C-terminal domain (CTD) divided by the central active-site α-helix α4 (**Figure 1A**). The α/β NTD consists of eight helices $3_{10}1$ – $3_{10}5$ and α1-α3, respectively. A flexible substrate-binding loop (termed S-loop) interconnecting helices α3 and $3_{10}5$ covers the active site cleft. Both crystallographically independent molecules exhibit well-resolved S-loop segments in an identical open conformation, similar to the conformation of one molecule (chain A) of PPEP-1 in the substrate-unbound crystal form (PDB: 5A0P) and in the substrate-free structure of PPEP-2 (PDB: 6FPC). PPEP-3 differs from PPEP-1 and PPEP-2 by the elongated N-terminus, where helix α1 is extended by a loop and two 3_{10} helical segments (**Figure 1B**). Additional differences of significance are observed in the active site of PPEP-3. In PPEP-1, the side chain of His-150 is positioned away from the active site cleft, while Tyr-160 in PPEP-1 is rotated inwards and therefore likely affects the substrate specificity (**Figure 1C**). Furthermore, the S2' pocket residue Leu-179 in PPEP-1 is substituted by a more sterically demanding Phe in PPEP-3 (Phe-190) (**Figure 1C**).

The active site helix α4 in PPEP-3 harbors the two histidine residues (His-152, His-156) coordinating the catalytic zinc ion and the catalytic base Glu-153, which collectively form the conserved characteristic HExxH motif of the zincin family [13]. The C-terminal domain is formed by the six helices $3_{10}6$ and α5 – α9 which carry Glu-196 (α7) and Tyr-189 (α6). Glu-196 coordinates the zinc ion, whereas the oxy-anion hole providing Tyr-189 in the PPEP-3 crystal structure rotated away from the active site. This discrepancy might be explained by the observation that there is an unidentified ligand, presumably originating from the crystallization buffer, coordinating to the catalytic zinc ion. Possible candidates are the ethylene glycols being present in the crystallization buffer, or a Bicine buffer molecule. However, the electron density does not allow a definite identification, which is why it was interpreted as water molecules.

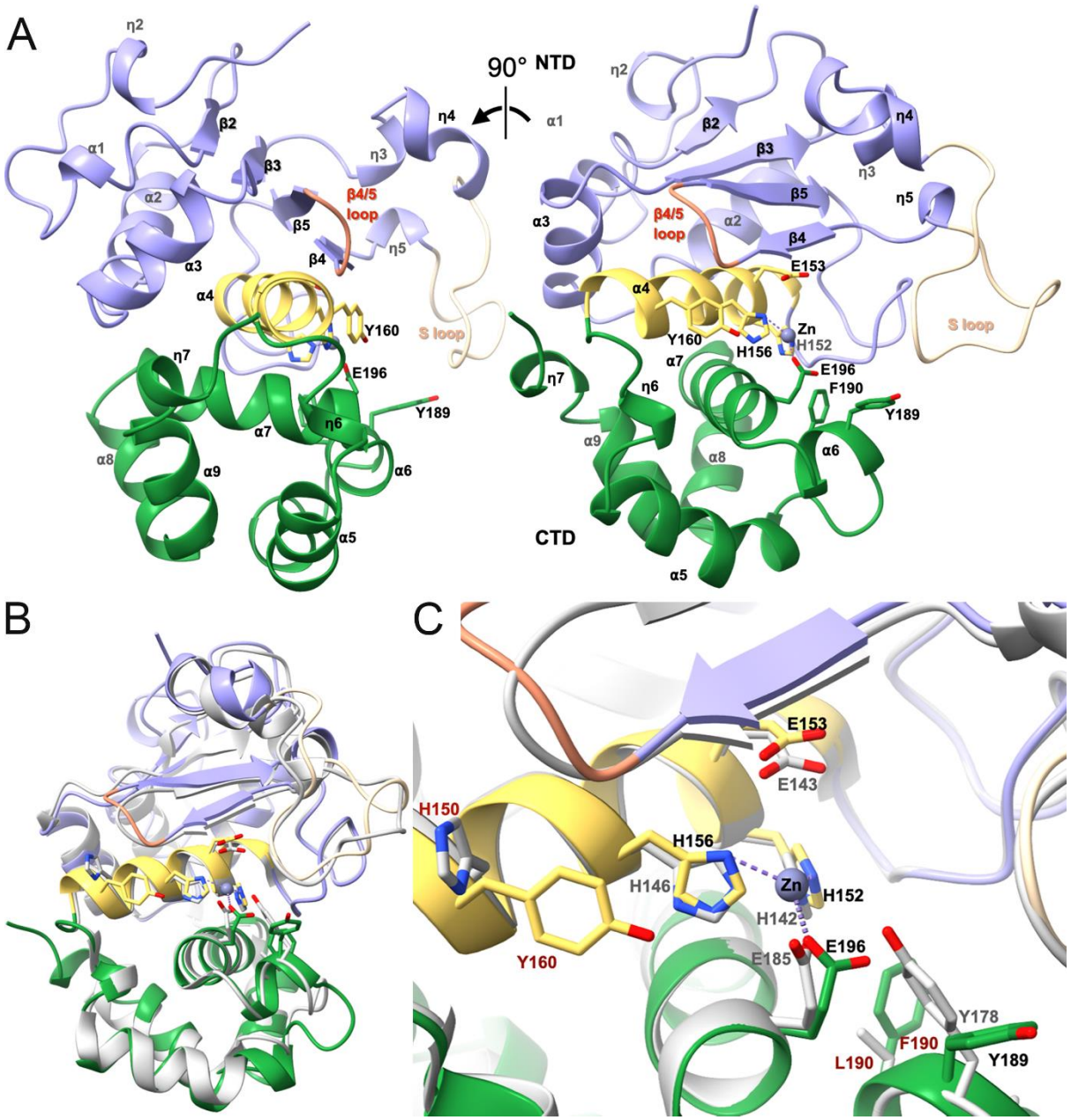


Figure 1. The overall structure of PPEP-3. (A) X-ray crystallographic structure of PPEP-3 (residues 27-234) in cartoon representation in two orthogonal views. Shown are the N-terminal domain (NTD in blue, the active site helix (yellow), the C-terminal domain (CTD) in green, the S-loop in amber, the β2 loop in red, and the zinc ion in grey. Zinc-coordinating and catalytically involved residue side chains as well as the two residues that alter specificity compared to PPEP-1 (Tyr-160, Phe-190) are depicted as sticks. **(B)** Superposition of PPEP-3 (colors as in A) with PPEP-1 (grey). **(C)** Detailed view of the catalytic site of PPEP-3 (colors as in A) and PPEP-1 (grey).

Substrate recognition by PPEP-3

We chose the peptide Ac-EPLPPPP-NH_2 (with PLPPPP acting as the P3-P3' residues) for our cocrystallization experiments, since this peptide was a known substrate of PPEP-3 [230]. In addition, it possessed all prolines at the prime-side, which is the optimal prime-side sequence for PPEP-3 [230]. Furthermore, a Leu at the P2 position is preferred by PPEP-2 and the closely related PPEP-4 [206].

To create an inactive mutant of PPEP-3, the catalytic base Glu-153 and the tetrahedral state stabilizing residue Tyr-189 were mutated to generate an E153A/Y189F double mutant. The structure of PPEP-3 complexed with Ac-EPLPPPP-NH_2 was determined at a resolution of 2.2 Å. The protein crystallized in space group $P2_12_12_1$ with four copies in the asymmetric unit. In all four copies, the bound substrate peptide is well-defined in the electron density map, and the catalytic zinc ion is still present with about 50 % occupancy.

Similar to PPEP-1, PPEP-3 has a beads-on-a-string like network of aromatic and aliphatic residues on the S-loop and bulge edge segment (**Figure 2**). These residues are His-103, Leu-104, Trp-119, Pro-109, and Tyr-112. His-103 and Tyr-112 are the Tyr-94 and Trp-103 residues in PPEP-1, respectively (**Supplemental Figure S2**).

The specificity of PPEPs to hydrolyze Pro-Pro originates from the interactions between the P1-P1' prolines and the S1 and S1' pockets (**Figure 2**). These interactions in PPEP-3 are highly similar to those in PPEP-1, and most of these residues are conserved in the other PPEPs (**Supplemental Figure S2**). The P1 Pro residue (Pro4*) is enclosed in the hydrophobic S1 pocket formed by Pro-109, Trp-119, Tyr-112, and Val-122. In addition, the main chain carbonyl oxygen interacts through hydrogen bonding with the His-152 and His-156 residues that coordinate the catalytic zinc ion. The P1' Pro (Pro5*) side chain interacts with Leu-149, Tyr-112, His-144, Ser-146 and His-152. Of all the residues involved in these protease-substrate interactions, only the Tyr-112 and Ser-146 residues differ from those in PPEP-1. In PPEP-1, the Tyr-112 and Ser-146 of PPEP-3 are Trp-103 and Ala-136, respectively (**Supplemental Figure S2**).

The five-membered ring of the P3 Pro residue (Pro2*) interacts through van der Waals interactions in the hydrophobic S3 pocket with Leu-104, Trp-119, and Ile-125 (**Figure 2**). The relatively small size and the restricted movement of the proline ring likely decrease the amount and strength of the van der Waals interactions compared to larger aliphatic residues.

The P2 Leu residue (Leu3*) interacts with PPEP-3 through both hydrogen bonds and van der Waals interactions (**Figure 2**). As in the PPEP-1 cocrystal, the main chain carbonyl oxygen and amide nitrogen form hydrogen bonds with the amide nitrogen and carbonyl oxygen, respectively, of Gly-126. The leucine side chain is snugly embedded in the S2 pocket, which is formed by the residues Arg-110, Gly-127, His-156, Tyr-160, Glu-195 and Glu-196.

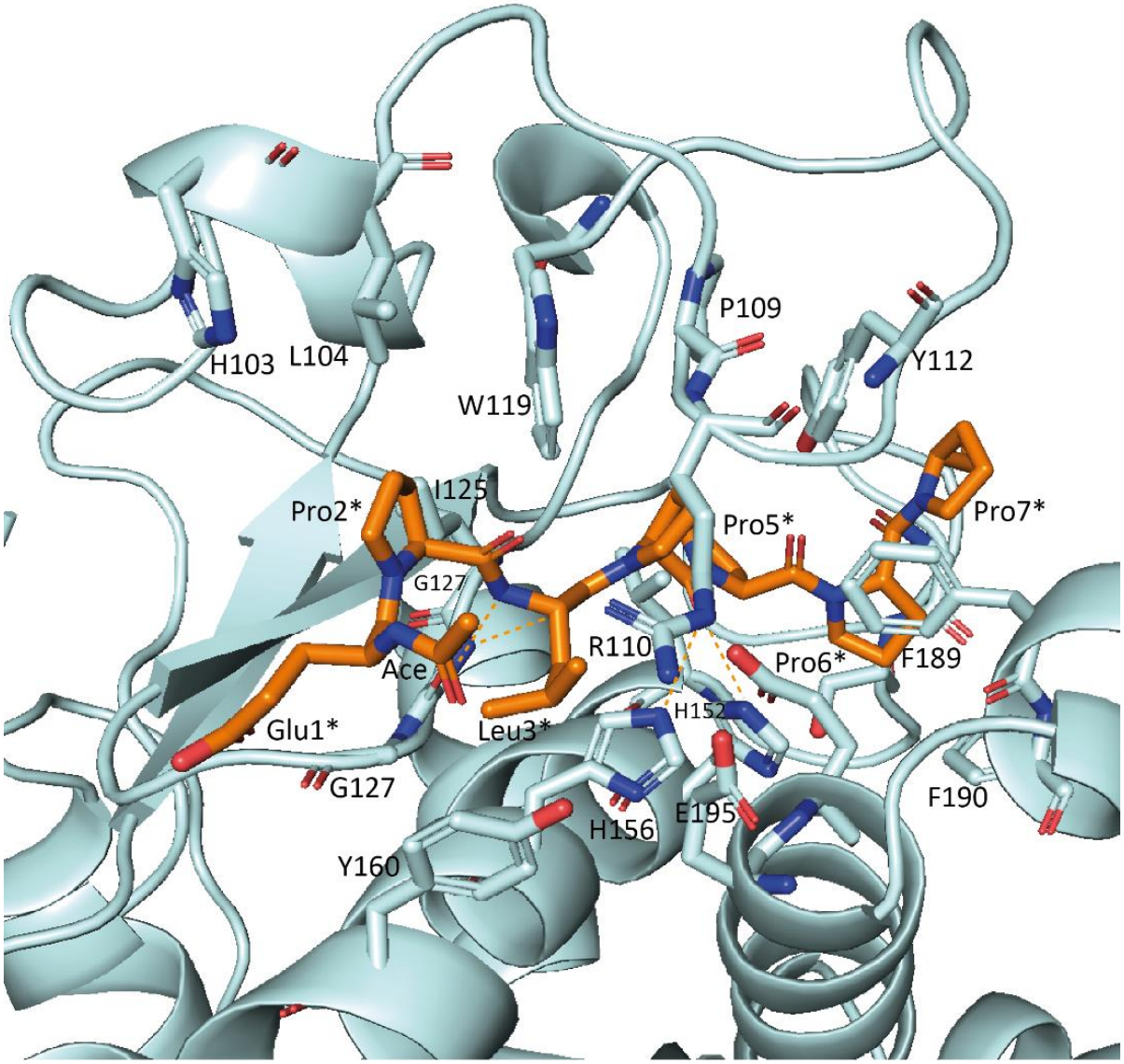


Figure 2. PPEP-3 in complex with the peptide Ac-EPLPPPP-NH_2. The cartoon representation of PPEP-3 is shown in cyan and the interacting residues are shown as sticks. The substrate peptide is shown as orange sticks. Hydrogen bonds are shown as orange dashed lines. Note: His-152 and His-156 form hydrogen bonds with the carbonyl group of Pro5*.

Substrate binding-induced conformational changes

In the PPEP-1 cocrystal with VNPPVP (P3-P3'), substrate binding causes the closure of the S-loop (**Figure 3A**). S-loop closure proceeds similarly in PPEP-3, however, additional conformational changes are observed for this PPEP (**Figure 3B**). Compared to PPEP-1, more movement is observed at the η3/η4 loop. This movement brings the Ile-100 of the η3/η4 loop in closer proximity to the aliphatic residues Leu-130 and Leu-132 located on the β5 strand located directly beneath, thereby increasing the van der Waals interactions between these elements and possibly aiding in the closure of the neighboring S-loop. Another substrate binding-induced change is observed at the α6 helix, which moves away from the active site. The α6 helix contains Phe-190, which is part of the S2' pocket and restricts the size of P2' residues in the substrate (**Figure 3C**). The movement of Phe-190 increases the S2' pocket's size and thereby accommodates the presence of a Pro residue at the P2' position (**Figure 3C**). However, its large side chain may still interfere with sterically more demanding residues at the P2' position, e.g. the valine of the PPEP-1 substrate, as is further discussed below.

In PPEP-1, the Lys-101 located in the S-loop forms hydrogen bonds with Glu-184 and Glu-185 found on the lower rim of the substrate-binding cleft [162]. In addition, Lys-101 hydrogen bonds with the Asn at the P2 position in the substrate. In PPEP-3, the interactions resulting from the S-loop closure differ from those in PPEP-1. First, the Lys-101 in PPEP-1 is substituted by Arg-110. In the substrate-bound conformation, this Arg-110 interacts with the γ-carboxyl group of Glu-195 in PPEP-3 through the Nε nitrogen and N$^{\eta 2}$ nitrogen atoms (**Figure 3D**). Unlike PPEP-1, no interactions are formed between the Arg and the neighboring Glu residue (Glu-196). In addition, no hydrogen bonds are formed between Arg-110 and the Leu at the P2 due to their physiochemical characteristics.

The backbone of the substrate peptide Ac-EPLPPPP-NH_2 adopts a conformation highly similar to the PPEP-1 substrate Ac-EVNPPVP-NH_2 (PDB: 6R5C), i.e., it adopts a double-kinked conformation (**Figure 3E**). In both substrates, the X-Pro bonds are in the *trans* conformation. The P1 and P1' Pro residues overlap perfectly, and also the Leu and Asn at the P2 are comparable in size and orientation. Furthermore, the P2' Val from the PPEP-1 substrate is sterically more demanding than the Pro in the PPEP-3 substrate.

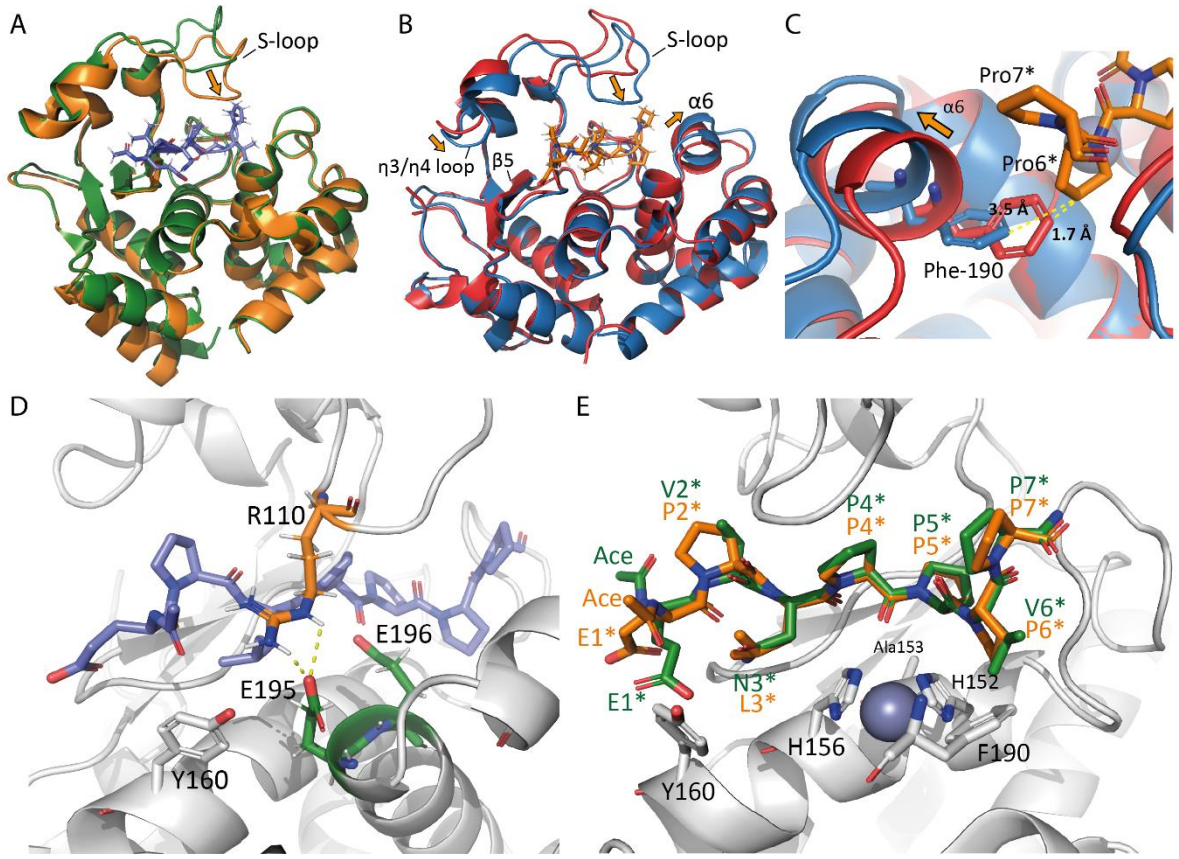


Figure 3. Substrate binding of PPEP-3. (A) Overlay of unbound (green, PDB: 5A0P) and substrate-bound (orange, PDB: 6R5C) PPEP-1. The substrate from the bound form is represented as purple sticks. **(B)** Overlay of unbound (red) and substrate-bound (blue) PPEP-3. The substrate from the bound form is represented as orange sticks. **(C)** Movement of the α6 helix increases the S2' pocket. The Phe-190 in the substrate-bound PPEP-3 (blue) moves away from the P2' Pro (Pro6*) compared to the unbound form (red). The closest distances between the Phe-190 and Pro6* are shown as dashed yellow lines. **(D)** The hydrogen bonds (yellow dashed line) formed between Arg-110 and Glu-195 following the closure of the S-loop. Arg-110, Tyr-160, Glu-195, and Glu-196 are shown as sticks. PPEP-3 is shown in grey and the substrate as purple sticks. **(E)** Overlay of the PPEP-1 substrate Ac-EVNPPVP-NH_2 (green) and the PPEP-3 substrate Ac-EPLPPPP-NH_2 (orange). PPEP-3 is shown in grey.

Profiling the substrate specificity of PPEP-3 using synthetic combinatorial peptide libraries

We determined the non-prime- and prime-side specificity of PPEP-3 using synthetic combinatorial peptide libraries specifically designed for PPEPs [206]. These libraries have two consecutive prolines in their core, while the surrounding positions are varied. We used two peptide libraries: one for determining the non-prime-side specificity and the other for the prime-side specificity. The non-prime-side library contains sequences with a PTEDAVXXPPXXEZZO motif (X = any residue except Cys, Z = 6-aminohexanoic acid, O = Lys(biotin)-amide). The prime-side library contains sequences with a JZEXXPPXXGGLEEF motif (X = any residue except Cys, Z = 6-aminohexanoic acid, J = biotin). The approach to profile the P3-P3' specificity has been previously described [206]. In short, the libraries were mixed and incubated with a PPEP. Non-biotinylated product peptides originating from Pro-Pro cleavage are PTEDAVXXP (non-prime-side) or PXXGGLEEF (prime-side). Non-biotinylated product peptides were enriched by negative selection on a streptavidin column and analyzed by LC-MS/MS. Extracted ion chromatograms (EICs) were produced, showing the intensities of the product peptides (**Figure 4**). Based on these intensities, a logo was constructed that shows the relative occurrence of a residue at a position surrounding the cleavage site (**Figure 4**).

We inspected the MS2 spectra to correctly annotate any ambiguous signals in the EIC. However, based on the MS2 spectra, we could not discriminate between the isomeric residues Leu and Ile. However, peptides with Ile residues elute before peptides with Leu [230,262], enabling peptide assignment based on the retention time. However, four signals were observed in the EIC that could originate from the PTEDAVIIP, PTEDAVLLP, PTEDAVLIP, or PTEDAVILP product peptides, with one signal larger than the others. We synthesized the four peptides and analyzed their retention on a C18 column using LC-MS/MS to annotate this signal. The isomeric peptides were completely resolved in time (**Supplemental Figure S3**), allowing us to annotate PTEDAVLIP in **Figure 4**.

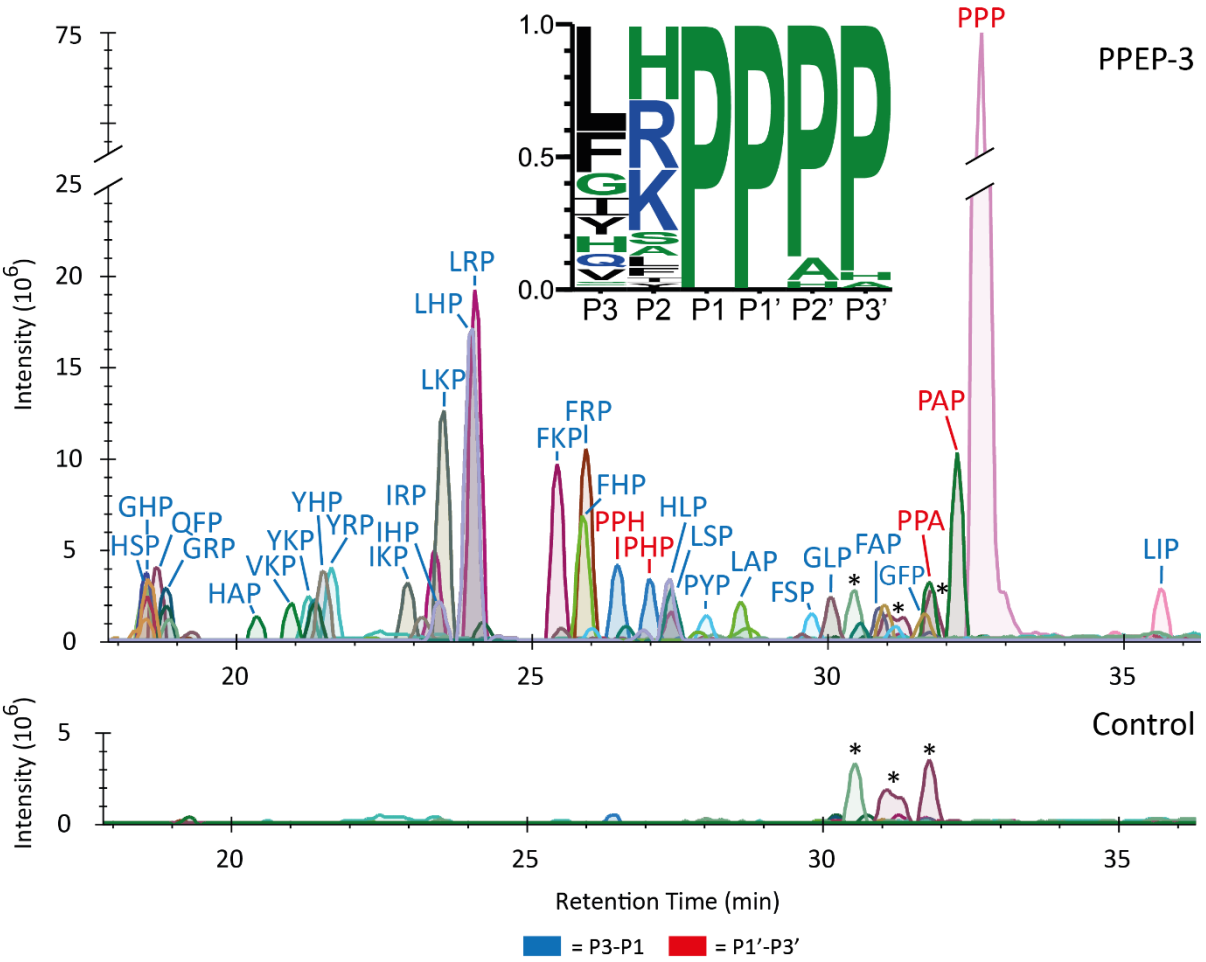


Figure 4. PPEP-3 specificity for amino acids surrounding the Pro-Pro cleavage site. A combinatorial peptide library was incubated with PPEP-3, product peptides were analyzed by LC-MS/MS, and a database search was performed to identify and quantify the products. Results were filtered for 9-mer product peptides and the most abundant products that collectively account for >90% of the total abundance per library were used to create the EICs. The PTEDAVXXP (non-prime-side) and PXXGGLEEF (prime-side) product peptides are shown in blue and red respectively. Mass tolerance was set to 5 ppm. An untreated control sample was included. A logo was created based on the product peptides to show the relative occurrence of the residue at a position surrounding the cleavage site. *MS/MS spectra did not indicate the presence of PTEDAVXXP or PXXGGLEEF product peptides.

PPEP-3 displays a strong preference for Pro at the P2’ and P3’ positions

The results from the combinatorial peptide library are in good agreement with previous data on the prime-side specificity, which showed a high preference for Pro residues at the P2’ and P3’ positions [230]. The small differences observed between the logo in **Figure 4** and the previously reported logo are mainly due to differences in the inclusion criteria of product peptides. The preference for a Pro at the P3’ position is a shared characteristic of PPEPs [206,230]. In PPEP-1, the Trp-103 interacts with the Pro at the P3’ in a parallel aliphatic-aromatic stacking interaction and forms a hydrogen bond to the carbonyl oxygen of the P1’ proline [162]. This Trp residue is crucial for activity and mutation to a Tyr diminishes PPEP-1 activity greatly [162]. In PPEP-3, the corresponding residue is Tyr-112, which does not prevent PPEP-3 activity as it does in PPEP-1. In fact, Tyr-112 interacts with the P3’ Pro similar to Trp-103 in PPEP-1, i.e., through an aliphatic-aromatic CH/π interaction (**Figure 5A**) [295]. In addition, the Pro7* residue is oriented at a 90° angle to Phe-189, which is introduced during mutagenesis to create the proteolytically inactive PPEP-3 (**Figure 5A**). The partially positive carbon of the pyrrolidine ring (Cδ) could interact with the negative electrostatic potential of the aromatic ring of Phe-189 (Tyr-189 in the wild-type) in a second CH/π interaction [296].

While all characterized PPEPs tolerate a Pro at the P2’ position, PPEP-3 displays the strongest preference for this residue (**Figure 4** and [206,230]). Notable differences are observed when comparing the S2’ pocket of PPEP-3 to that of PPEP-1 (**Figure 5B**). The most significant difference impacting P2’ specificity is the presence of a sterically demanding Phe residue (Phe190) in the S2’ pocket of PPEP-3. Substituting the Pro6* with a Cβ-branched Val residue, which is well tolerated by other PPEPs, results in a steric clash in PPEP-3 (**Figure 5C**).

In addition to the Pro, the logo in **Figure 4** shows the presence of Ala and His at the P2’ position. A modeled substitution of the Pro6* with Ala does not cause a steric clash but reduces the amount of van der Waals interactions between the substrate residue and Phe190 (**Figure 5C**). In addition, Pro residues increase backbone rigidness owing to the restricted φ angle compared to other residues, thereby reducing the entropy loss of the substrate upon binding. A substrate with an Ala at the P2’ loses more entropy to adopt the right conformation to fit the active site.

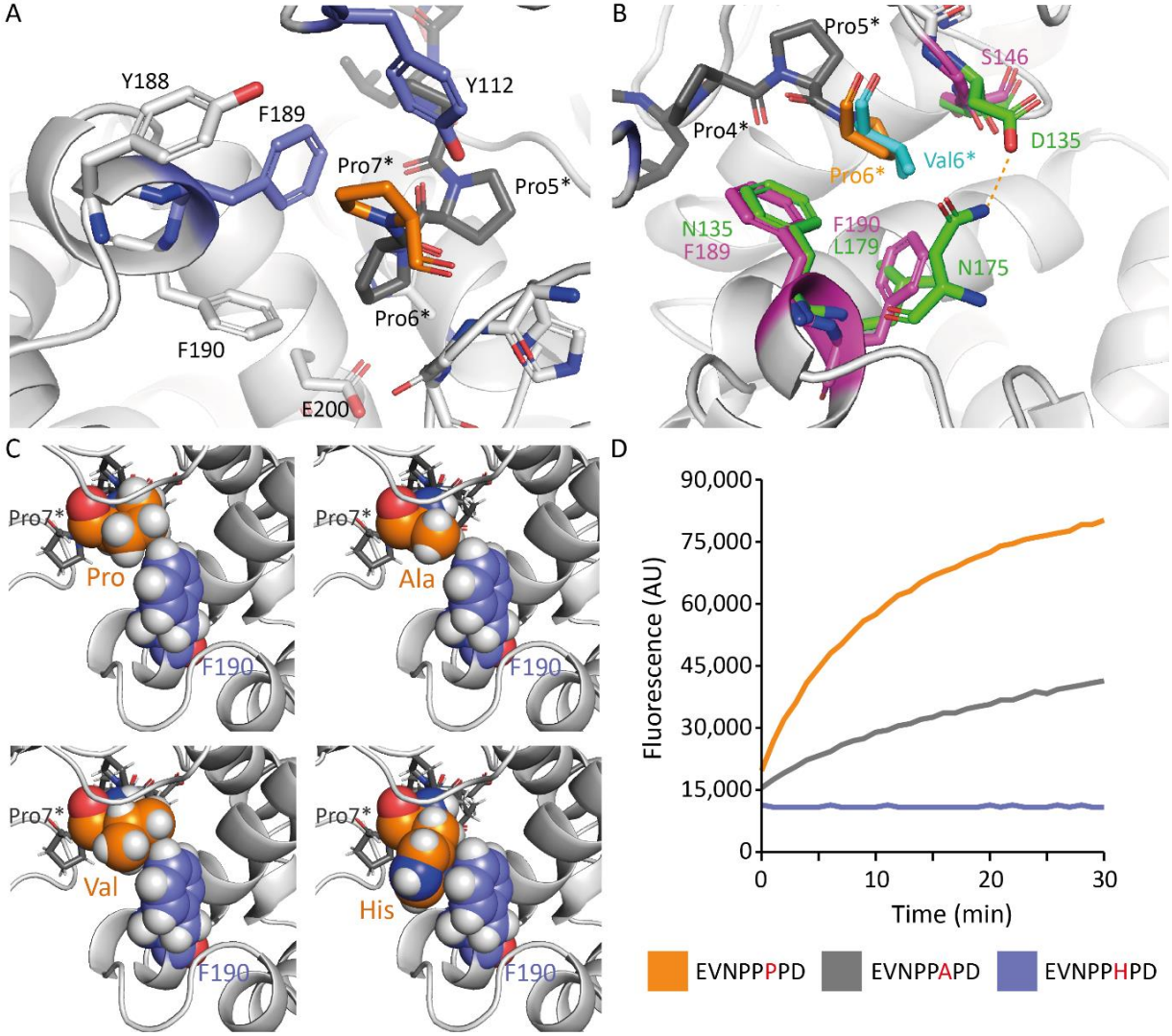


Figure 5. Structural analyses of the prime-side specificity of PPEP-3. (A) The Pro7* interacts with Tyr-112 and Phe-189 through aliphatic-aromatic interactions. Tyr-112 and Phe-189 are colored purple. The substrate peptide Ac-EPLPPPP-NH_2 is colored dark grey and the P3' Pro7* residue is colored orange. PPEP-3 is shown in grey. **(B)** A comparison of the S2' pocket residues of PPEP-3 (grey) and PPEP-1. For PPEP-1, only the S2' pocket residues are shown as green sticks. The PPEP-3 S2' residues are colored magenta. The substrate peptide Ac-EPLPPPP-NH_2 is colored dark grey with it P2' residue Pro7* in orange. The P2' Val6* residue from the PPEP-1 substrate Ac-EVNPPVP-NH_2 is colored cyan. The hydrogen bond between D135 and Asn175 from PPEP-1 is shown as a dashed line. **(C)** Modeled substitution of the P2' Pro (upper left) with Ala (upper right), Val (lower left), and His (lower right). The Phe-190 (purple) and the P2' residue (orange) are shown as spheres. **(D)** Time-course of PPEP-3 mediated cleavage of FRET-quenched peptides with the sequence Lys(Dabcyl)-EVNPP(P/A/H)PD-Glu(EDANS).

The presence of His at the P2' is surprising due to its size and this residue produces steric clashes when trying to model its side chain in between the Phe-190 and Pro7* (**Figure 5C**). However, the signal for the PHP (P1'-P3') product peptide is low compared to PPP and PAP (P1'-P3') peptides. Additionally, histidine residues are overrepresented due to the efficient ionization of His-containing peptides [230,266]. An assay using

FRET-quenched peptides showed a preference for a Pro at the P2', a lower activity for Ala at the P2', and no activity when a His occupied the P2' position (**Figure 5D**).

Additional differences in the S2' pocket between PPEP-1 and PPEP-3 involve Asp-135 and Asn-175 in PPEP-1. In PPEP-1, these hydrogen bonding residues are part of the S1'-wall-forming segment and the diverting loop (**Figure 5B**), necessitating the double-kinked conformation of the substrate peptide [160]. In PPEP-3, the residue corresponding to the Asn-175 of PPEP-1 is not part of the S2' pocket, while the Asp-135 in PPEP-1 is substituted by a Gly residue (Gly-145) in PPEP-3 (**Figure 5B**). While these structural differences might allow for more flexibility at the P2' position, Phe-190 in PPEP-3 restricts this flexibility and is, therefore, the primary determinant of P2' specificity.

The preference of PPEP-3 for hydrophobic residues at the P3 position

PPEP-3 mostly tolerates hydrophobic residues at the P3 position, although also His, Gly, and Gln are observed. The S3 pocket consists of His-103, Leu-104, and Trp-119 and is backed up by Ile-125 (**Figure 6**). The many hydrophobic residues in this pocket, together with the location at the surface of the protein, explain the preference for hydrophobic residues due to both hydrophobic and van der Waals interactions at the P3 position. For example, Leu is the most preferred residue at the P3 position, a preference that can be explained by its hydrophobic character and the many van der Waals interactions that are possible between the P3 Leu and His-103, Leu-104, and Trp-119 (**Figure 6A**).

In PPEP-1, -2, -3, and -4, residues Leu-104 and Trp-119 are well conserved. In addition, the Ile-125 that closes the S3 pocket is either a Leu or Val, residues with similar physicochemical properties. The main difference is the His-103 in PPEP-3, which is a Tyr residue in the other PPEPs (**Supplemental Figure S2**). When comparing the P3 specificity between PPEP-3 and other PPEPs, the high occurrence of the Phe residue in the logo stands out (**Figure 4**). In PPEP-3, the Phe at the P3 position can form extensive van der Waals interactions with Leu-104 (**Figure 6B**). In the PPEP-1 cocrystal, the hydroxyl group of the Tyr-94 residue in the PPEP-1 cocrystal structure (PDB: 6R5C) restricts the size of the hydrophobic pocket compared to the His-103 in PPEP-3. Modeling of a Phe residue in the PPEP-1 cocrystal indicates that the Tyr-94 and Trp-110 clamp the Phe residue, thereby increasing the distance to Leu-95, which reduces and weakens the van der Waals interactions with the Leu-95 (**Figure 6C**).

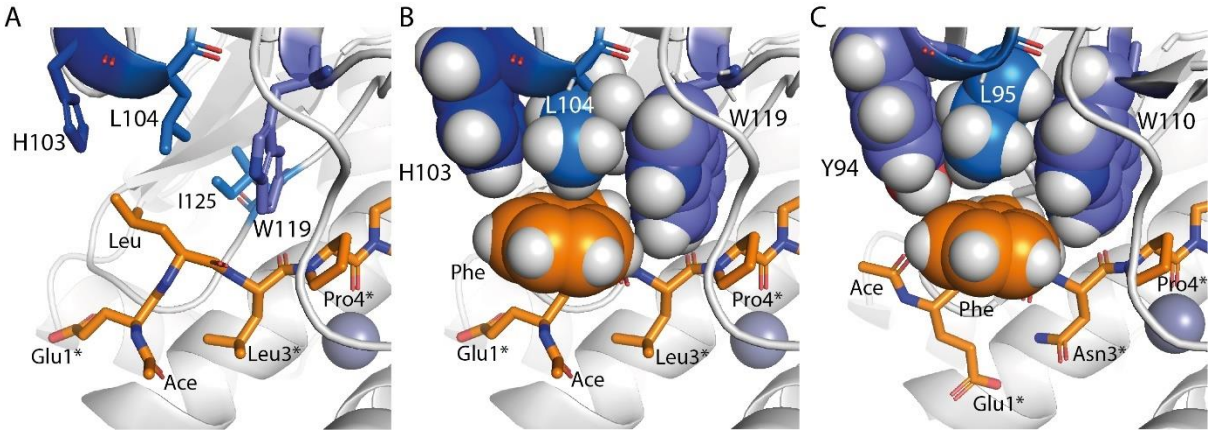


Figure 6. P3 specificity of PPEP-3. (A) The S3 pocket of PPEP-3 (grey). The P3 position is substituted for a Leu by modeling. The S3 residues His-103 (dark blue), Leu-104 (blue), Trp-119 (purple), Ile-125 (blue), and the substrate peptide (orange) are shown as sticks. **(B)** A Phe residue is modeled at the P3 position. The P3 Phe, His-103, Leu-104, and Trp-119 are shown as spheres. **(C)** A Phe residue is modeled at the P3 position of the PPEP-1 (grey) substrate peptide Ac-EFNPPVP-NH_2 (orange). Tyr-94 (purple), Leu-95 (blue), Trp-110, and the P3 Phe are shown as spheres).

The importance of Tyr-160 for the P2 specificity of PPEP-3

The P2 specificity of PPEP-3 is characterized by a preference for the basic residues His, Arg, and Lys (**Figure 4**). However, these residues are overrepresented in the EICs and logo due to their efficient ionization in LC-MS/MS [230,266]. Still, cleavage assays using FRET-quenched peptides showed a preference for the basic residues at the P2 over the Leu (**Figure 7A**) which was used to produce the cocrystal and is also observed in the logo in **Figure 4**.

Following the basic residues, Ser is the next most abundant residue observed at the P2 position (**Figure 4**). Comparison of FRET-quenched peptides with either Ser or His at the P2 position shows a preference for Ser (**Figure 7B**). The preference for Ser can be explained by the hydrogen bonding with Tyr-160 observed after modeling the Ser at the P2 (**Figure 7C**). However, this hydrogen bond is only present when the Tyr-160 adopts the rotamer observed in the apo structure. A Leu at the P2 position necessitates a larger S2 pocket which causes the Tyr-160 side chain to move away from the active site in the PPEP-3 cocrystal with Ac-EPLPPPP-NH_2 by mainly adopting a different the χ_2 dihedral angle (**Figure 7C**).

The increase in S2 pocket size by side chain conformations of Tyr-160 that differ from the apo crystal structure is also needed to explain the presence of the basic residues at the P2. Substitution of the P2 Leu for His, Arg, and Lys reveals steric clashes with Tyr-160. Possibly, Tyr-160 can adopt a similar conformation as His-150 in PPEP-1 (**Figure 1C**), thereby allowing the basic residues to fit the S2 pocket.

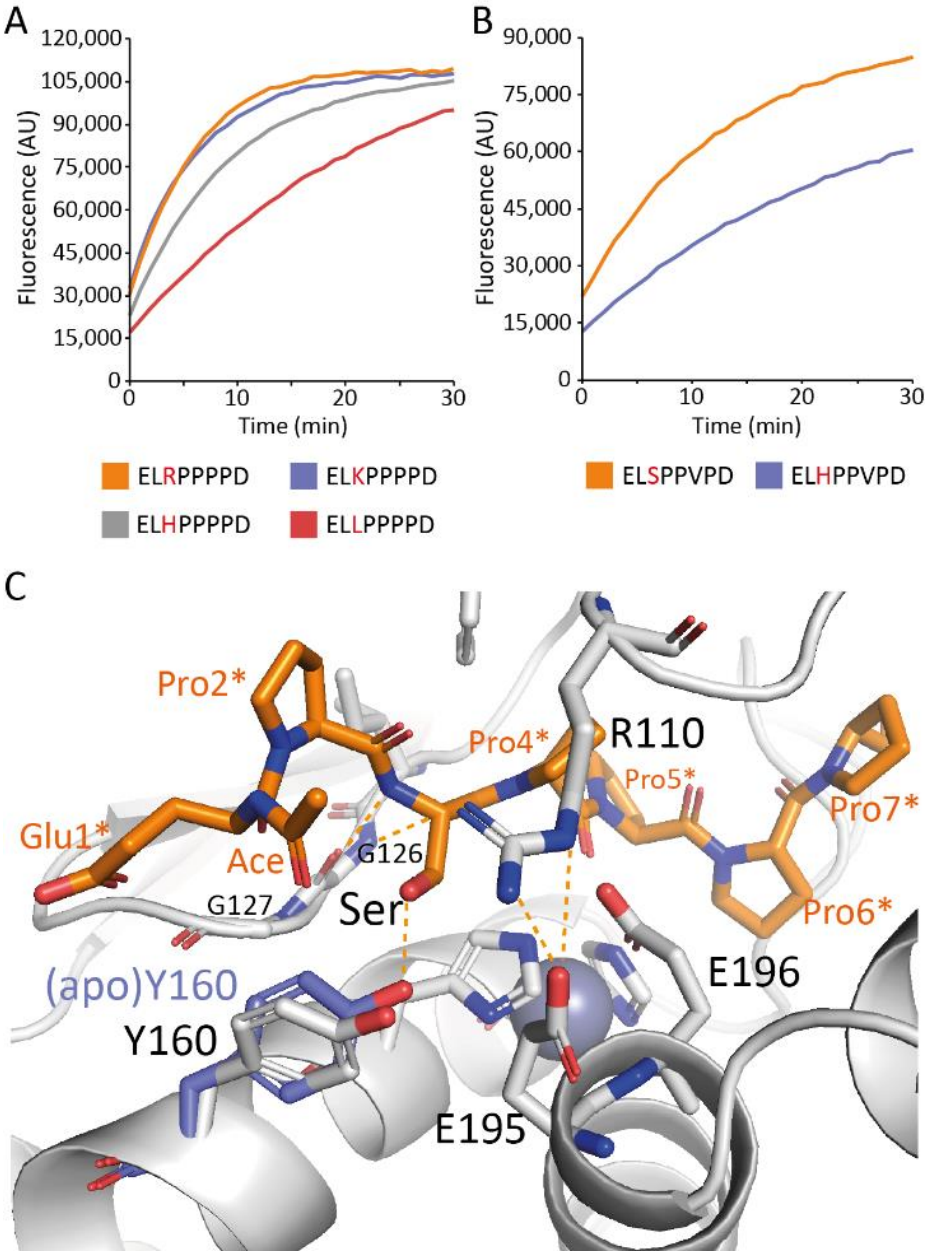


Figure 7. P2 specificity of PPEP-3. (**A**) Time-course of PPEP-3 mediated cleavage of FRET-quenched peptides with the sequence Lys(Dabcyl)-EL(R/H/K/L)PPPPD-Glu(EDANS). (**B**) Time-course of PPEP-3 mediated cleavage of FRET-quenched peptides with the sequence Lys(Dabcyl)-EL(S/H)PPVPD-Glu(EDANS). (**C**) The S2 pocket of PPEP-3. Modeling of Ser at the P2 produces a hydrogen bond with Tyr-160 in the conformation of the apo structure. Gly-126, Gly-127, Arg-110, Y-160, Glu-195, Glu-196, and the substrate peptide (orange) are shown as sticks. PPEP-3 is colored grey. Tyr-160 from the apo structure is colored purple. Hydrogen bonds in the S2 pocket are shown as orange dashed lines.

Discussion

Based on the structures and the cleavage motif presented here, we have provided mechanistic insights into the substrate specificity of PPEP-3. Our current and previous data showed a preference for all prolines at the prime-side positions in the substrates, which we could attribute to both size restrictions in the substrate pockets and extensive interactions between the Pro residues at the P2' and P3' positions. In addition, Pro residues are preferred at the P2' position due to the increased rigidity of the peptide backbone, which reduces the entropy loss of the substrate upon binding.

The non-prime-side specificity of PPEP-3 was less explored. Previously, FRET-quenched peptide cleavage assays using PPEP-3 showed that this protease tolerated VNP, PLP, PSP, and to a lesser extent DNP at the P3-P1 positions in the context of all prolines at the prime-side [230]. Since both PPEP-2 and PPEP-4 prefer a Leu residue at the P2 position, the peptide Ac-PLPPPP-NH_2 was a rational candidate for cocrystallization experiments. Although cocrystallization with the peptide Ac-EPLPPPP-NH_2 was successful, based on our current understanding of PPEP-3 substrate specificity, a different substrate peptide might bind more efficiently. However, by using Ac-EPLPPPP-NH_2, we observed movement of the Tyr-160 side chain, which is likely a common mechanism to increase the size of the S2 pocket, thereby allowing larger residues such as His, Arg, and Lys at the P2 position.

Although PLPPPP (P3-P3') is cleaved in assays using FRET-quenched peptides [230], we did not identify the PTEDAVPLP product peptide in our combinatorial peptide library experiment. Previously, we showed that in the context of PLPPPP substituting the P2' Pro for a Val residue eliminates PPEP-3 activity [230]. However, the FRET-quenched peptides containing LSPPVP and LHPPVP (P3-P3') were cleaved by PPEP-3 (**Figure 7B**). This indicates that a Val is exclusively tolerated at the P2' when the non-prime-side residues are highly favored by PPEP-3. Since the less favored motif PLP (P3-P1) is most likely only tolerated in the context of PPP and possibly PAP (P1'-P3'), the resulting PTEDAVPLP product peptides in our combinatorial peptide library assay do not exceed the limit of detection. This phenomenon is especially observed for PPEP-3. For the other PPEPs that display a more variable prime-side specificity, more peptides displaying a specific non-prime-side sequence are cleaved, which increases the non-prime-side product peptide signals in our LC-MS/MS analyses.

In PPEP-1, Trp-103 is essential for activity due to its interactions with the P3' Pro and the hydrogen bonding of the side chain nitrogen with the carbonyl oxygen of the P1' residue in the protease-substrate complex [162]. Mutation of this residue to Ala, His, Phe, and Tyr greatly diminished PPEP-1 activity [162]. In PPEP-3, the corresponding residue Tyr-112 interacts similarly with the Pro at the P3' but might also produce a similar hydrogen

bond with the P1’ carbonyl oxygen. The distance between the two oxygen atoms involved measures 2.6 Å, which is just outside the range of about 2.7 to 3.0 Å for asymmetric hydrogen bonds [297]. However, although we cannot determine the exact position of the hydrogen atom of the hydroxyl group and with the current resolution of 2.2 Å, hydrogen bonding between the Tyr-112 and the P1’ Pro carbonyl, similar to the interaction in the PPEP-1 cocrystal, is likely to occur.

Based on the preference of PPEP-3 for all prolines at the prime-side, we searched for secreted proteins possessing four consecutive prolines (P↓PPP, P1-P3’) in the *G. thermodenitrificans* proteome [230]. This search identified two proteins with either PSP↓PPP or DNP↓PPP as the putative PPEP-3 cleavage site, with PSP↓PPP being the far better substrate [230]. However, strong binding between PPEP-3 and the non-prime-side residues allows for more flexibility at the P2’ position (**Figure 7B**). Based on our new combinatorial peptide library results, we performed a search for endogenous substrates that included proteins that contained the motif (L/F)(H/R/K/S)P↓P (P3-P1’), resulting in the identification of 50 proteins. Of these 50 proteins, only a single protein, GTNG_0399, was predicted to possess a signal peptide for secretion by SignalP 6.0 (https://services.healthtech.dtu.dk/services/SignalP-6.0/). GTNG_0399, a spore coat N-acetylmuramic acid deacetylase, contains an LRPPRG site. Given the peptide library results shown in **Figure 4**, the combination of an Arg at the P2’ and a Gly at the P3’ is most likely not tolerated by PPEP-3. In addition, our LC-MS/MS analysis does not indicate the presence of a PRGGGLEEF product peptide. Therefore, the protein GTNG_0956 containing the putative cleavage site PSP↓PPP (P3-P3’) [230] remains the most likely endogenous candidate, especially since we can explain the preference for a Ser residue at the P2 position due to the hydrogen bonding with Tyr-160 (**Figure 7B,C**). Alternatively, the biological PPEP-3 substrate could also originate from a different organism.

The unique ability to specifically hydrolyze Pro-Pro bonds could be advantageous in applications that necessitate precise proteolysis, such as the removal of affinity tags [298]. In addition, several industrial processes, e.g., the breakdown of collagen for meat tenderization, require proteolysis of proline-rich proteins [299–301]. Although PPEP specificity is too strict to degrade a variety of proteins, directed mutagenesis could render these proteases more promiscuous while retaining the Pro-Pro specificity. A detailed understanding of the factors that determine PPEP specificity can aid in the development of PPEPs suitable for industrial applications. In this study, we shed more light on the structure-function relationship of PPEPs by combining an experimentally determined protease-substrate complex with an in-depth substrate specificity profile. This combination of techniques can be a valuable tool to study the mechanisms governing substrate specificity in other PPEPs or, with some adaptations to the peptide libraries, other proteases.

Experimental procedures

Expression and purification of recombinant PPEP-3

The truncated version (amino acids 26-234, lacking the N-terminal predicted signal peptide) of the PPEP-3 gene (GTNG_1672) from *Geobacillus thermodenitrificans* strain NG80-2, codon optimized for Escherichia coli, was obtained in a pET28a vector using the restriction sites NdeI / XhoI. An active site double mutant was generated via the one-step site-directed mutagenesis protocol [302]. For the E153A mutant the PCR was performed using pET28a-PPEP3 as template and oligonucleotides JGP614-GeoPPEP_E153A_f: 5'-CTGCACGCATTCGCGCACTCTCTGG-3' as well as JGP613-GeoPPEP_E153A_r: 5'-CGAATGCGTGCAGTTCCAGGTTG-3'. For the construct pET28a-PPEP3(E153A/Y189F) the construct pET28a-PPEP3(E153A) was used as template and the oligonucleotides JGP615-GeoPPEP_Y189F_f: 5'-GAATACTTCTTCCTGACCTACCCGG-3' and JGP616-GeoPPEP_Y189F_r: 5'-CAGGAAGAAGTATTCACGCGGGAAC-3' were used to introduce the second mutation. A reaction was performed using 16 cycles with 98 °C denaturation for 30 sec, 65 °C annealing for 30 sec and 72 °C elongation for 6 min followed by a 2 min final elongation step. Subsequently a DpnI digest was conducted using 1 U DpnI (NEB) at 37 °C for 1 h. 2 µl of the reaction were transformed into chemically competent *E. Coli* DH5a cells (Thermo Fisher Scientific), plated on LB-agar selection plates supplemented with 50 µg/ml kanamycin and incubated overnight at 37 °C. Isolated vectors were sequenced to identify positive clones.

The vectors pET28a-PPEP3 wild-type and pET28a-PPEP3(E153A/Y189F) were transformed into chemically competent *E. coli* BL21 (DE3) cells (Invitrogen) via heat shock protocol, plated on LB-agar selection plates supplemented with 50 µg/ml kanamycin and incubated overnight at 37 °C. A preculture grown overnight at 37 °C from a single colony was used to inoculate 6 x 1 L expression cultures (LB supplemented with 50 µg/ml kanamycin) to an optical density (OD_{600}) of 0.1. After incubation at 37 °C and reaching an OD_{600} of 0.7 expression was induced with 0.5 mM Isopropyl 1-thio-Beta-D-galactopyranoside (IPTG, BIOTREND). Protein expression was performed at 20 °C overnight. Cells were harvested by centrifugation at 4,000 x g, 4 °C for 20 min. Cell pellets were washed with Tris-buffered saline (TBS) (20 mM Tris [pH 7.5], 200 mM NaCl). Cells were pelleted again and stored at −80°C until further use.

The proteins were purified as previously described with minor adjustments [160]. The cell pellet from 2 L of culture was resuspended in TBS buffer (20 mM Tris [pH 7.5], 300 mM NaCl) supplemented with 10 µg/ml DNaseI (AppliChem). Cells were lysed by running the suspension two times through a Celldisruptor (I&L Biosystems) at 2.5 kbar. Cellular debris was pelleted by centrifugation at 10,000 x g, 4 °C for 10 min. The supernatant was cleared by ultracentrifugation at 165,000 x g, 4 °C for 30 min. The supernatant was

adjusted with 1 M imidazole (pH 7.5) to a final concentration of 10 mM and loaded onto 2 ml NiNTA superflow resin (Qiagen). After two wash steps with TBS supplemented with first 10 mM and then 30 mM imidazole, the protein was eluted with TBS containing 250 mM imidazole. Protein concentration was determined at 280 nm using the molar extinction coefficient of 27,390 M^{-1} cm^{-1} (wild-type) and 25,900 M^{-1} cm^{-1} (double mutant), respectively. In addition to dialysis against 50x the elution volume in TBS, 2 U of thrombin (Sigma Aldrich) was added to the protein solution to cleave the His_6-tag during dialysis at 4 °C overnight. The protein solution was passed through the same NiNTA superflow column (equilibrated to TBS with 10 mM imidazole), collecting the cleaved protein in the flow-through. The protein was concentrated and applied on a HiLoad Superdex 200 16/600 column (Cytiva) equilibrated with TBS. Protein fractions were collected, concentrated, and stored at -80 °C until further use.

Crystallization of PPEP-3

Single crystals of substrate-unbound wild-type, double mutant E153A/Y189F in unbound and Ac-PLPPPP-NH_2 were obtained by broad screening using sitting drop vapor diffusion crystallization with drop sizes of 300 nl. Protein (381 µM, 10 mg/ml) was pipetted in ratios of 1:2, 1:1 and 2:1 (protein to precipitant) in commercially available crystallization screens (Hampton Research). For substrate complex formation, the catalytic Zn^{2+} ion was removed by dialyzing the protein solution against buffer containing about 6 mM EDTA and 6 mM ortho-phenanthroline in order to avoid proteolysis, which occurs even in the double mutant albeit slowly. Crystal formation was observed in conditions Morpheus C1, C5, C9, E9 and H9. Best diffracting crystals were obtained from Morpheus E9 containing 10% w/v PEG 20 000, 20% v/v PEG MME 550, 0.3 M diethyleneglycol, 0.3 M triethyleneglycol, 0.3 M tetraethyleneglycol, 0.3 M pentaethyleneglycol, 0.1 M bicine/Trizma base pH 8.5. Single crystals were cryoprotected in a mixture of a precipitant solution containing 50% sucrose and flash-frozen in liquid nitrogen.

Data collection and structure determination

High-resolution data for structure determination were collected at ESRF on the beamline ID30A-3 using an Eiger X 4M detector (Dectris) or at beam line ID30B. Datasets were processed with XDS (Kabsch, 2010). The structure was solved using molecular replacement employing the PPEP-1 coordinates (PDB: 5A0P) as a search model. Phasing and refinement were performed using the PHENIX package [303] and model building with Coot [304]. Data collection and refinement statistics are shown in **Supplemental Table S1**.

Combinatorial peptide library assays

The combinatorial peptide libraries were synthesized, and assays were performed as previously described [230]. In short, approximately 10 nmol of precleaned (on avidin column) peptides was incubated with 200 ng PPEP-3 for 3 h at 37 °C in PBS. A nontreated control was included. After incubation, the samples were loaded onto an in-house constructed column consisting of a 200 µL pipet tip containing a filter and a packed column of 100 µL of Pierce High Capacity Streptavidin Agarose beads (Thermo, the column was washed four times with 150 µL of PBS before use) to remove the biotinylated peptides. The flow-through and four additional washes with 125 µL H_2O were collected. The product peptides were desalted using reversed-phase solid-phase extraction cartridges (Oasis HLB 1 cm3 10 mg, Waters) and eluted with 200 µL of 30% acetonitrile (v/v) in 0.1% formic acid. Samples were dried by vacuum concentration and stored at −20 °C until further use. For the peptide library assays in which the non-prime- and prime-side libraries were combined, approximately 5 nmol of each library was used (10 nmol in total).

LC-MS/MS analyses

PPEP-3 product peptides analyzed as previously described [206] by online C18 nano-HPLC MS/MS with a system consisting of an Easy nLC 1200 gradient HPLC system (Thermo, Bremen, Germany) and an Orbitrap Fusion LUMOS mass spectrometer (Thermo). Peptides were injected onto a homemade precolumn (100 µm × 15 mm; Reprosil-Pur C18-AQ 3 µm, Dr Maisch, Ammerbuch, Germany) and eluted via a homemade analytical nano-HPLC column (30 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 µm). The gradient was run from 2% to 40% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 52 min. The nano-HPLC column was drawn to a tip of ∼5 µm and acted as the electrospray needle of the MS source. The LUMOS mass spectrometer was operated in data-dependent MS/MS mode for a cycle time of 3 s, with HCD collision energies at 20 V, 25V, and 30V and recording of the MS2 spectrum in the orbitrap, with a quadrupole isolation width of 1.2 m/z. In the master scan (MS1) the resolution was 120,000, the scan range 350–1600, at an AGC target of 400,000 at a maximum fill time of 50 ms. A lock mass correction on the background ion m/z = 445.12003 was used. Precursors were dynamically excluded after n = 1 with an exclusion duration of 10 s and with a precursor range of 10 ppm. Charge states 1–5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 30,000 at an AGC target of 100% @maximum fill time of 60 ms.

LC-MS/MS data analysis

The LC-MS/MS data were analyzed as previously described [206]. For the identification of product peptides after analysis of the mixed non-prime- and prime-side libraries, a database was generated containing all possible 9-mer product peptides that can be expected based on Pro-Pro cleavage (i.e., PTEDAVXXP and PXXGGLEEF).

Raw data were converted to peak lists using Proteome Discoverer version 2.5.0.400 (Thermo Electron) and submitted to the in-house created databases using Mascot v. 2.2.7 (www.matrixscience.com) for peptide identification, using the Fixed Value PSM Validator. Mascot searches were with 5 ppm and 0.02 Da deviation for precursor and fragment mass, respectively, and no enzyme specificity was selected. Biotin on the protein N-terminus was set as a variable modification.

The database search results were filtered for product peptides that contained either PTEDAV or GGLEEF, were 9 residues in length, and contained no biotin. The resulting peptide lists were transported to Microsoft Excel, where duplicate masses and corresponding abundances were removed (e.g., the abundances of isomers PLPGGLEEF and PIPGGLEEF are listed twice, while this abundance is the total abundance of the two). The most abundant product peptides that together accounted for >90% of the total abundance were selected for further. Further analysis was performed in Skyline 23.1.0.268 by importing the product peptides as FASTA along with the raw data files [292]. The Extracted Ion Chromatograms (EICs) displaying the product peptides were created by plotting the intensities of the signals corresponding to the monoisotopic m/z values of both 1+ and 2+ charged peptides with a mass tolerance of 5 ppm.

FRET peptide cleavage assays

FRET-quenched peptide cleavage assays with PPEP-3 were performed using peptides with a Lys_{Dabcyl}-EXXPPXXD-Glu_{Edans} (the X positions varied between peptides). Assays were performed in 150 µl PBS containing 200 ng enzyme and 50 mM FRET peptide. Peptide cleavage was analyzed using an Envision 2105 Multimode Plate Reader at 37 °C. Fluorescence intensity was measured every minute for 30 min, with 10 flashes per measurement. The excitation and emission wavelengths were 350 nm and 510 nm, respectively.

Bioinformatic analyses

Structures were analyzed using PyMOL (The PyMOL Molecular Graphics System, Version 2.5.5 Schrödinger, LLC) and USCF ChimeraX [305].

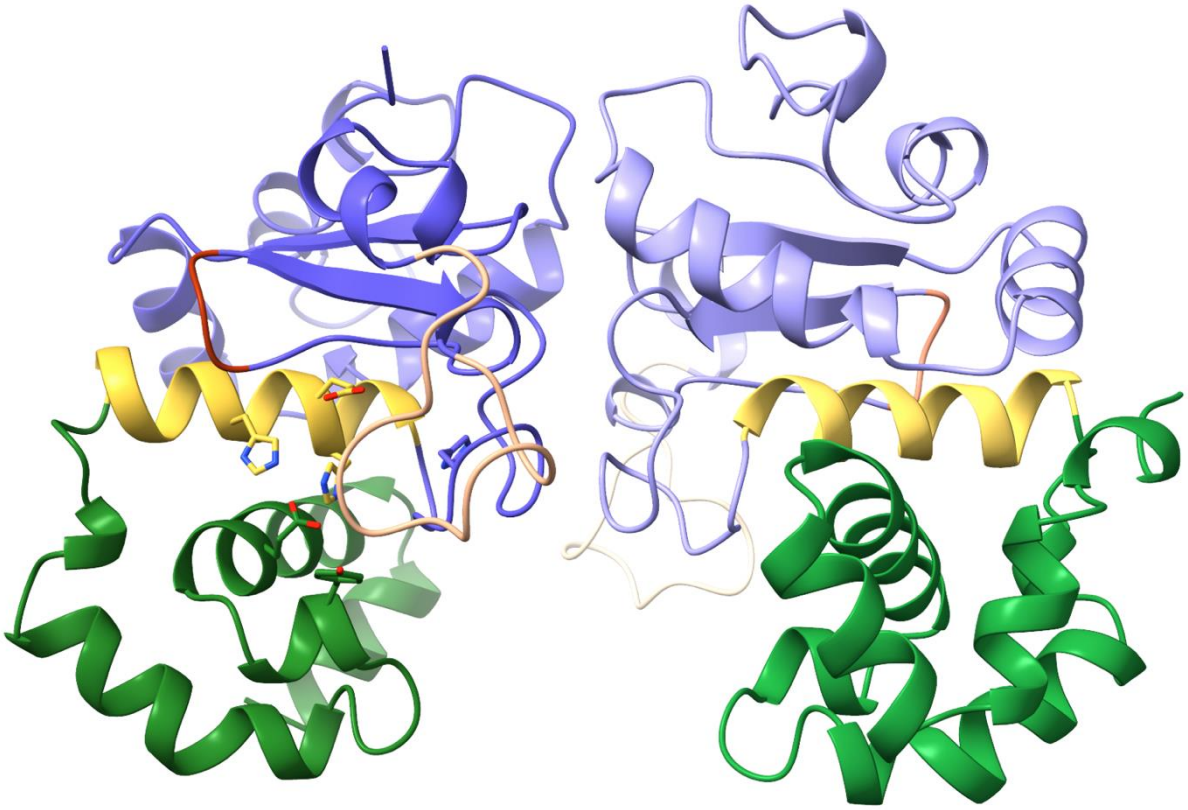
Supporting information

Table S1. Data collection and refinement statistics.

	PPEP-3 WT	PPEP-3 E153A Y189F	PPEP-3 E153A Y189F EPLPPPP
Wavelength			
Resolution range	45.03 - 1.696 (1.74 - 1.7)	44.96 - 2.094 (2.16 - 2.09)	47.91 - 2.201 (2.26 - 2.2)
Space group	P 41 21 2	P 41 21 2	P 21 21 21
Unit cell	126.5 126.5 64.12 90 90 90	126.26 126.26 64.06 90 90 90	73.625 95.822 127.432 90 90 90
Total reflections	1167460 (71030)	444888 (6900)	314143 (22548)
Unique reflections	109522 (7262)	59935 (4518)	88244 (6247)
Multiplicity	10.7 (9.8)	7.4 (1.5)	3.6 (3.6)
Completeness (%)	99.45 (92.29)	96.64 (74.05)	99.87 (99.02)
Mean I/sigma(I)	6.96 (1.08)	8.37 (1.86)	5.90 (1.03)
Wilson B-factor	27.50	24.30	35.67
R-merge	0.1541 (1.891)	0.2168 (1.026)	0.172 (1.272)
R-meas	0.1617 (1.993)	0.2273 (1.373)	0.203 (1.499)
R-pim	0.04849 (0.6143)	0.06609 (0.9058)	0.1066 (0.7849)
CC1/2	0.997 (0.437)	0.994 (0.262)	0.989 (0.431)
CC*	0.999 (0.78)	0.999 (0.644)	0.997 (0.776)
Reflections used in refinement	57697 (3769)	30006 (2040)	46404 (3230)
Reflections used for R-free	2000 (130)	1501 (102)	1999 (140)
R-work	0.1941 (0.2597)	0.1882 (0.2422)	0.1773 (0.2674)
R-free	0.2179 (0.2976)	0.2259 (0.2928)	0.2250 (0.3564)
Number of non-	3571	3471	7308

hydrogen atoms			
macromolecules	3451	3425	6990
ligands	2	2	7
solvent	118	44	311
Protein residues	422	421	860
RMS(bonds)	0.007	0.008	0.005
RMS(angles)	0.78	0.91	0.78
Ramachandran favored (%)	99.76	98.32	98.33
Ramachandran allowed (%)	0.24	1.68	1.67
Ramachandran outliers (%)	0.00	0.00	0.00
Rotamer outliers (%)	0.00	0.28	0.14
Clashscore	1.47	1.78	5.38
Average B-factor	31.12	26.93	41.68
macromolecules	31.00	26.95	41.52
ligands	27.38	26.45	49.04
solvent	34.84	25.28	45.19

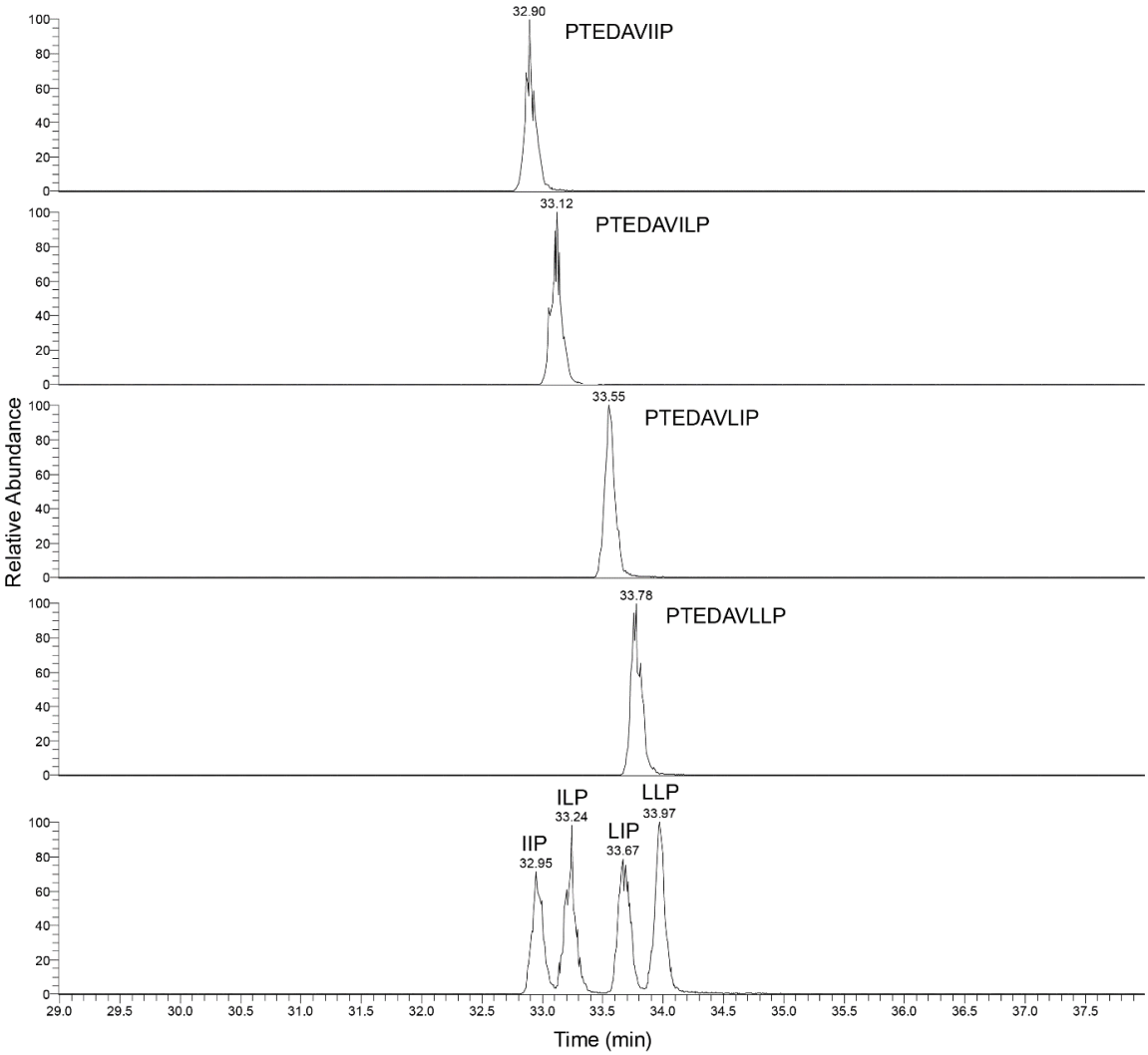
Statistics for the highest-resolution shell are shown in parentheses.



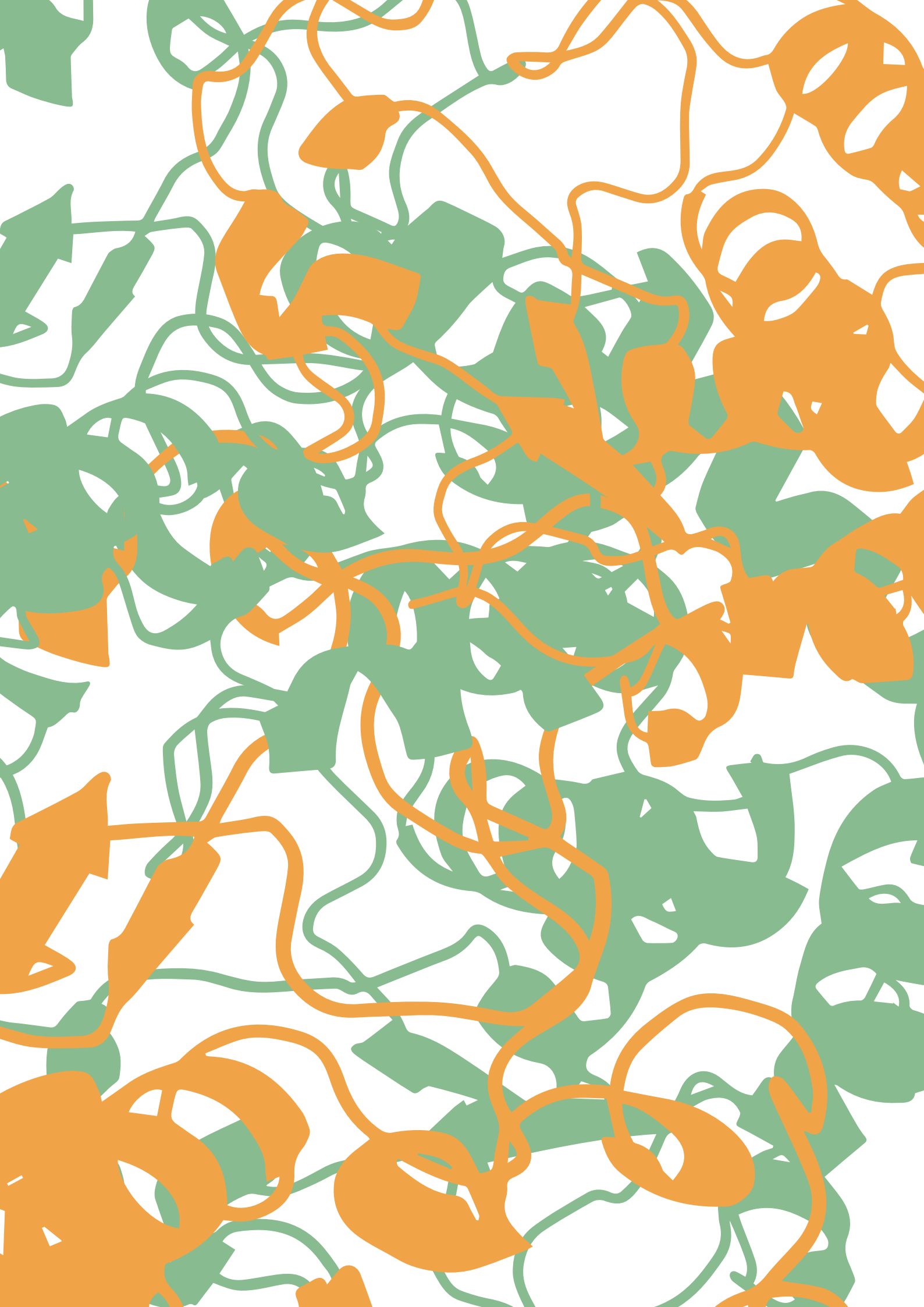
Supplemental Figure S1: PPEP-3 asymmetric unit. X-ray crystallographic structure of PPEP-3 in cartoon representation. Shown are the N-terminal domain (NTD in blue, the active site helix (yellow), the C-terminal domain (CTD) in green, the S-loop in amber, and the β2 loop in red. Zinc-coordinating and catalytically involved residue side chains as well as the two residues that alter specificity compared to PPEP-1 (Tyr-160, Phe-190) are depicted as sticks in the left monomer.

```
PPEP-1      MRPSKKLLIAIISIFLI------SSVPVS------AHADSTTIQQNKDTLSQIVVFPTGN       48
PPEP-2      MKWDKRVVALILAVMIV------CP--L---------FAAPAHAQEQSILDKLVVLPSGE       43
PPEP-3      ---MRRLLTSLLAALLAVPLLSLSPYPAAHGVLLEESSLDVQTIPSHDVLGRIVIVPETD       57
PPEP-4      ---MKRLLVLMGVVVAVIPLLSFSPYPNTHGILLQASSLSIDRVPSHQTLQKIVILPDAP       57
                ::::  :   .        .                    .:. * ::*:.*

PPEP-1      YDKNEANAMVNRLANIDGKYLNALKQNNLKIKLLSGKLTDEKEYAYLKGVVPKGWEGTGK       108
PPEP-2      YNHSEAAAMKQRLEKIPTSILDALYSKGVKIKLTQGAITNEPELAYLKGVVPRGWEGTGL       103
PPEP-3      FSFDEANEMIRTLARIDRRILEQAANHHIYIQLLTNPITDEPIARHLRGKTPRGYVPGSK       117
PPEP-4      FDHSEAKKMIETLAHIDPALLKKIADQHIYLKLFTGKLTDEPSARYLRGQTPRGYIS-PA       116
            :. .**  * . * .*    *.   .: : ::*  . :*:*    :*:* .*:*:

PPEP-1      TWDDVPGLGGST-VALRIGFSNKGKGHDAINLELHETAHAIDHIVLNDISKSAQFKQIFA       167
PPEP-2      TWDDVPGVSERV-VAVRIGYSEKGKGHNSLNLEIHETLHAVDRLVLNEVSGTDEFINIFN       162
PPEP-3      TWDEVPGIGGAHLVLVRLGHSEKGKGHGSINLELHEFAHSLDYIVFDHIHETDEFQALWR       177
PPEP-4      TWDDVPGLGGSHLVLVKIGHSEKGKGHGSVNLELHELAHSVDYIVFDHIHATSPFLAVWR       176
            ***:***:.    * :::*.*:*****.::***:**  *::* :*::.:  :  *  ::

PPEP-1      KEGR-SLGNVNYLGVYPEEFFAESFAYYYLNQDTNSKLKSACPQTYSFLQNLAK----  220
PPEP-2      KEASVKYKGDGYVSAYPTEYFAEAASLYLYSDATRSDLKDSMPLTYEFMAKLFAN---  217
PPEP-3      EEAPQLFPREYYFLTYPEEYFAESFAYYYVSEKTQETLRMAAPRTYTFIRQLAERAS-  234
PPEP-4      EEAKQLFPHNYYLLDYPEEYFAESFAYYYYNNETRAHLKETAPKTYEFIKQLIERTRQ  234
            :*.        *.  ** *:***: : *  .: *.  *: : * ** *: :*
```

Supplemental Figure S2. Sequence alignment of PPEP-3 and PPEP-1, -2, and -4. The multiple sequence alignment was produced using ClustalW.



Supplemental Figure S3. Separation of the product peptides PTEDAVIIP, PTEDAVILP, PTEDAVLIP, and PTEDAVLLP. The retention times of synthetic peptides were analyzed on a C18 column using LC-MS/MS. EICs were produced by including m/z = 954.5142 ($[M + H]^{+}$) and m/z 477.7608 ($[M + 2H]^{2+}$) with a mass tolerance of 10 ppm.



7

General Discussion

Post-translational modifications (PTMs) are crucial for the correct functioning of proteins and regulate numerous processes in bacteria. In the pathogenic bacterium *Clostridioides difficile*, the PTMs proteolysis and glycosylation significantly influence adhesion and motility. Proteolytic activity by PPEP-1 detaches the cells from the intestinal epithelium by cleaving adhesion proteins, while glycosylation of FliC with the Type A glycan is essential for motility.

This thesis focused on the enzymes involved in proteolysis and glycosylation, particularly examining their specificity, structure, and function. Special attention was given to the group of proteases known as Pro-Pro endoproteases, known for their unique substrate specificity. Our findings were obtained using a diverse array of methods, including various molecular biology techniques, extensive mass spectrometry-based methods, structural analyses, and an innovative synthetic combinatorial peptide library approach. Collectively, this allowed us to propose a revised model for the Type A glycan biosynthesis, study the enigmatic protein CD1597, and characterize PPEP specificity in great detail.

In this chapter, we will reflect on our findings, discuss additional insights, explore applications, and provide a framework for future research.

Glycosylation of FliC in *Clostridioides difficile*: A more thorough understanding of the biosynthetic pathway and the implications for further research

New insights into the Type A glycan biosynthesis

In **Chapter 2** we proposed a model for the biosynthesis of the Type A glycan that is present on FliC in *Clostridioides difficile* 630Δ*erm*. This model predicts the enzymatic activities of the proteins CD0241-CD0244 and the biosynthetic intermediates. Therefore, the model can be regarded as a collection of hypotheses that can be tested. The core intermediate of the predicted biosynthetic pathway is CDP-threonine, a molecule that has not been previously described. In the model, CDP-threonine is formed by CD0242, which is expected to transfer a phosphothreonine to CTP while releasing inorganic pyrophosphate. *In vitro* assays in which CD0242, phosphothreonine, and CTP are incubated together can be used to substantiate our model. Either the formation of CDP-threonine or inorganic pyrophosphate can be used as a readout.

A strong indication that CDP-threonine is indeed a biosynthetic intermediate in the Type A pathway comes from recent studies with CD0244. CD0244 is predicted to transfer the phosphothreonine moiety (with or without the methyl group) from CDP-threonine to the

GlcNAcs on FliC. Preliminary data from an *in vitro* assay indeed showed that recombinant CD0244 can transfer phosphothreonine to the GlcNAc on a synthetic peptide, using synthetic CDP-threonine as the donor substrate. These results strengthen the idea that CDP-threonine is a biosynthetic intermediate in the Type A pathway.

Interestingly, additional preliminary data from a similar *in vitro* assay using synthetic CDP-*N*-methylthreonine as the donor substrate showed that this reaction is not only possible, but also proceeded more efficiently than with CDP-threonine. In the context of our model, this suggests that CD0243 first methylates the threonine and that CD0244 subsequently attaches the *N*-methyl-phosphothreonine to the GlcNAc. In **Chapter 2**, we identified both peptides with a nonmethylated Type A and peptides with only the GlcNAc moiety in the *cd0243* mutant. We argued that the incomplete transfer of the phosphothreonine to the GlcNAc is due to the polar effects on *cd0244* resulting from the insertional mutagenesis of the *cd0243* gene, which still holds, but an additional reason could be the reduced efficiency of the reaction by CD0244 due to the absence of methylation by CD0243.

The structure of the Type B glycan in *Pseudomonas aeruginosa* strain PAO1 is similar to Type A from *C. difficile* [140]. In addition, the *P. aeruginosa* gene cluster encoding the biosynthetic proteins resembles the one in *C. difficile* 630Δ*erm* (**Chapter 2** and [140,142]). The Type B glycan in *P. aeruginosa* consists of a sugar (a deoxyhexose) that is linked through a phosphodiester to a hitherto unknown moiety [140,142]. Since the biosynthetic gene clusters in both organisms are very similar, one could expect this unknown moiety to be an *N*-methylthreonine. However, this unknown moiety has an additional mass of 14 Da compared to *N*-methylthreonine [140]. Recent MS analysis of FliC in *P. aeruginosa* suggests that the unknown moiety is an *N,N*-dimethylthreonine (unpublished data). A *P. aeruginosa* mutant strain that is unable to dimethylate the threonine, similar to the *cd2043* mutant of *C. difficile*, only carries the sugar moiety (unpublished data). Therefore, it appears that dimethylation of the threonine occurs before the transfer of the *N,N*-dimethyl-phosphothreonine to the sugar in *P. aeruginosa*, supporting the idea that in *C. difficile* the methylation by CD0243 occurs prior to the transferase activity of CD0244.

The role of glycosylation of flagellin in bacteria

Bacterial flagella are mostly recognized for their involvement in motility. However, flagella are also involved in adhesion [306–308], secretion of effector molecules and proteins [309,310], biofilm formation [311], and immunomodulation [310,312]. Glycosylation of the flagellin (FliC in *C. difficile*) is often involved in these processes. For example, in *Pseudomonas syringae* pv. *tabaci*, glycosylation of flagellin stabilizes the

flagella and is necessary for the swimming ability, adherence to polystyrene surfaces, and the ability to cause disease in tobacco plants [313–315]. In *C. difficile* strains that glycosylate FliC with the Type B glycan, such as the R20291 strain, glycosylation of FliC promotes motility and adherence but reduces biofilm formation [144]. Moreover, glycosylation of flagellin is also important for immune evasion. In *P. syringae*, glycosylation of the flagellin suppresses the immune response in tobacco plants by shielding the immunogenic flg22 region of flagellin [315]. In *Campylobacter jejuni*, immune evasion is achieved through mutations in the Toll-like receptor 5 (TLR5) epitope of flagellin, which is otherwise recognized by the host's TLR5 [316]. These mutations weaken the subunit-subunit interactions of flagellin and therefore necessitate interactions through other domains of flagellin, which are stabilized by the many glycosylation sites [316]. Thus, in the case of *C. jejuni*, glycosylation provides a solution for the destabilizing mutations that aid in immune evasion.

In *C. difficile* strain 630Δ*erm*, glycosylation of FliC is essential for motility, since mutation of the genes responsible for the synthesis of the Type A glycan renders the bacteria non-motile [142]. In addition, loss of the Type A glycan causes cell aggregation, increases binding to abiotic surfaces, and attenuates colonization and relapse in mice [142]. These findings demonstrate the importance of the glycosylation of FliC with the Type A glycan, but do not offer a reason why the glycosylation is needed for the correct functioning of the flagella. There are two general explanations. First, the Type A glycan might be beneficial for the bacteria, but glycosylation led to (or allowed) an aberrant FliC that renders the bacteria non-motile. Alternatively, the aberrant FliC benefits the bacteria, and glycosylation is needed for the correct functioning of the flagella, similar to *C. jejuni*.

C. difficile FliC has a similar TLR5 epitope as *Bacillus subtilis*, which is bound by TLR5 and causes activation of pro-inflammatory gene expression [317,318]. In *C. difficile* R20291, FliC glycosylation with the Type B glycan does not affect TLR5 activation [144], and since *C. difficile* 630Δ*erm* FliC is nearly identical, this might also be true for the Type A glycan. However, the glycosylation of FliC may suppress immune reactions through other mechanisms. Structural studies of glycosylated FliC might provide insights into the effects of glycosylation on flagellar structure, stability, and function. In order to perform such studies, a thorough understanding of the Type A biosynthesis is crucial. When one wants to produce the glycosylated FliC *in vitro*, FliC needs to be glycosylated by CD0240, after which synthetically produced phospho-*N*-methylthreonine can be attached by CD0244.

In vitro glycosylation of FliC with the Type A glycan might also benefit immunization studies. The role of flagellin in the innate and adaptive immune system is well-documented [319–321]. However, immunization studies using FliC from *C. difficile* employ recombinantly produced FliC, which is not glycosylated [322–325]. FliC consists

of two domains; one that forms the core of the filament and another that constitutes the outermost layer of the filament [326,327]. Importantly, the Type A glycan is found on the outermost layer of the flagellar filament and we therefore expect it to play a significant role in antigenicity. The use of glycosylated FliC in future immunization studies might prove effective in developing a vaccine against *C. difficile* infection. To develop vaccines against (hypervirulent) strains that carry the Type B glycan, studies that elucidate the biosynthetic pathway for this glycan are essential.

To cleave or not to cleave: the enigmatic PPEP-homolog CD1597

In **Chapters 3 and 5** we investigated the activity and function of CD1597 from *C. difficile*. Information on the specificity of proteases can aid in identifying their endogenous substrates and therefore their biological role. However, since no proteolytic activity was observed for CD1597, we could not predict a substrate and function for this protein using these approaches. In addition to the assays using FRET-quenched peptides and the combinatorial peptide library, we employed Terminal Amine Isotopic Labeling of Substrates (TAILS) to test the proteolytic activity of CD1597. The TAILS method aims to identify proteolytically generated neo-N-termini in a WT (or protease-treated) proteome that are absent in a protease knockout (or untreated) proteome [23]. In short, this is achieved by labeling the (neo-)N-termini of the two proteomes using isotopically different labels, e.g., light and heavy dimethyl labels. The differently labeled samples are mixed, digested, enriched for (neo-)N-termini by negative selection, and analyzed by LC-MS/MS. The aim is to identify peptides that are only present in the WT (or protease treated) sample, since these should result from cleavage by the protease of interest.

We used a Δ*cwp84* strain as a positive control in our TAILS experiments. Cwp84 cleaves the highly abundant SlpA precursor protein and the products of this reaction form the surface layer of *C. difficile* [328]. After LC-MS/MS analysis, we were able to readily identify the peptide that originated from cleavage by Cwp84. In our experiments with the *cd1597* mutant strain, however, we were unable to identify neo-N-terminal peptides that were generated by CD1597. The reason for this could be that CD1597 is not expressed or expressed in very low quantities during our experimental conditions. To overcome this issue, we also performed an *in vitro* TAILS experiment in which we compared CD1597 treated and untreated *C. difficile* proteomes. Although many neo-N-terminal peptides were identified as only present in the protease-treated sample by our analysis software, manual inspection of the raw data revealed that these were false positives. In addition, we identified several Pro-Pro cleavages, but these did not result from CD1597 activity since the peptides were present in both the WT (or protease-treated) and knockout (or

untreated) samples. However, this indicates that there are other proteases present in *C. difficile* that are able to hydrolyze Pro-Pro bonds.

The hypothesis that CD1597 is a PPEP-like protease was based on the similarity of the C-terminal domain of CD1597 and other PPEPs. However, CD1597 possesses a large N-terminal domain that could play an inhibitory function. Therefore, we recombinantly produced the C-terminal PPEP-like domain separately, which was often used alongside the full-length protein in our assays. However, no proteolytic activity was observed for the PPEP-like domain, thus it remains unclear whether the N-terminal domain could play an inhibiting function. However, the N-terminal domain is predicted to be attached to the PPEP-like domain through a long and flexible stretch of residues, which does not suggest an inhibitory function, since the PPEP-like domain is likely accessible to a potential substrate. However, the positioning of the N-terminal domain, and therefore the inhibitory function, could depend on a cofactor or stimulus.

Compared to other PPEPs, CD1597 displays several mutations and insertions that might influence the activity (**Chapter 3**). In **Chapter 5** we have shown the influence of the β3/β4 loop in PPEP-1 and PPEP-2 on the non-prime-side specificity. In CD1597, this loop is ^{303}YRNN^{306} and differs largely from ^{117}GGST^{120} in PPEP-1 regarding the size and biochemical properties of the side chains (**Figure 1**). Based on the alignment of the primary structure and the predicted structure, we constructed a ^{303}YRN^{305} → GG mutant of the PPEP-like domain, but this mutant displayed no proteolytic activity in assays using Pro-Pro-containing FRET-quenched peptides (data not shown). Furthermore, when comparing the apo structure of PPEP-1 with the PPEP-like domain of CD1597, we observed differences due to insertions in CD1597. Insertions of two residues in the diverting loop and four residues in a nearby part called the S1'-wall-forming segment [160] are predicted to enlarge the outer edges of CD1597 (**Figure 1**). In PPEP-1, these loops are involved in the redirection and exiting of the substrate [160]. The insertions in CD1597 might influence the prime-side specificity or prohibit the substrate from exiting the active site. To test whether the differences in the diverting loop and S1'-wall-forming segment render CD1597 inactive, a mutant of the PPEP-like domain of CD1597 was produced containing the PPEP-1-like β3/β4 loop, diverting loop, and S1'-wall-forming segment. However, this more PPEP-1-like mutant was also proteolytically inactive towards Pro-Pro-containing FRET-quenched peptides.

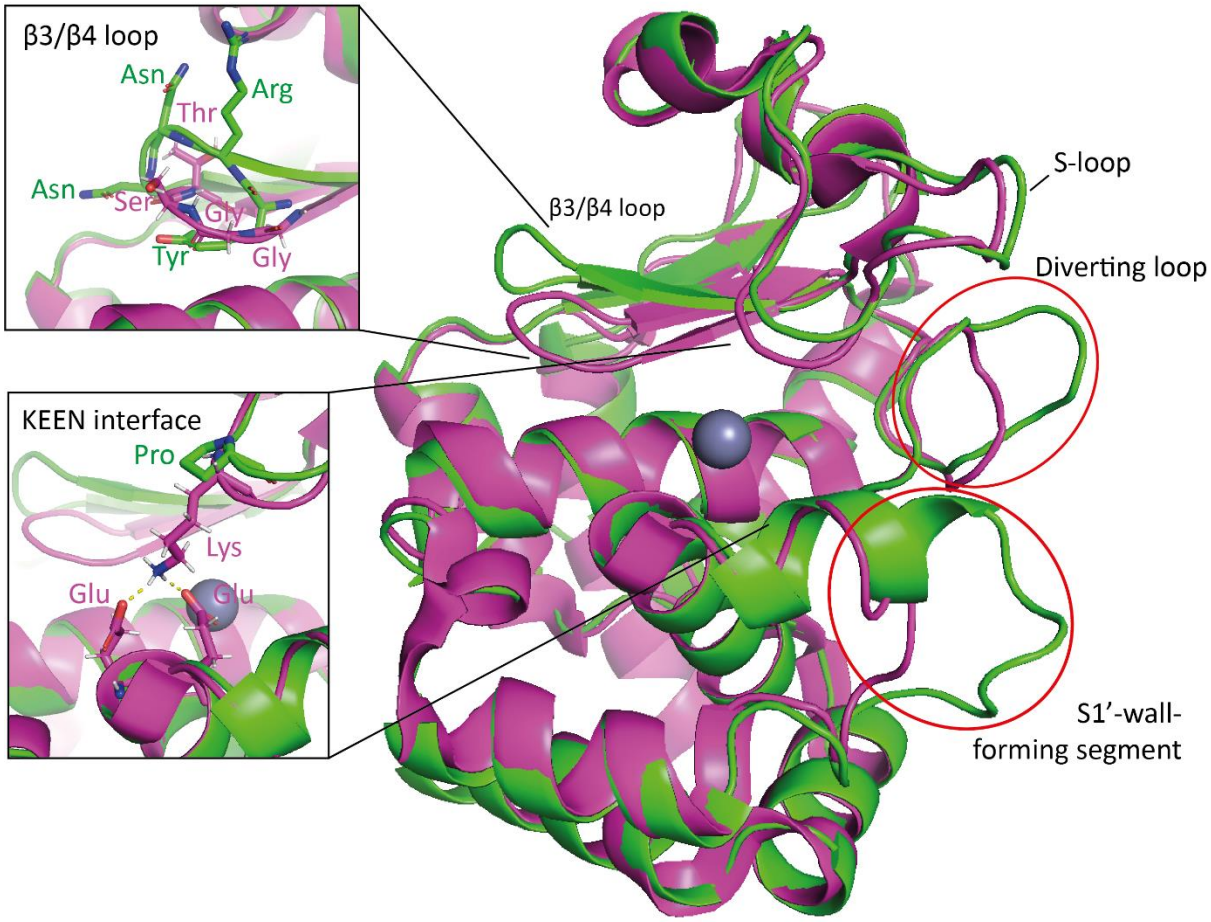


Figure 1. Comparison of the PPEP-like domain of CD1597 and the apo structure of PPEP-1. Insertions in the diverting loop and S1'-wall-forming segment are highlighted. Upper left panel: Close-up of the β3/β4 loop highlighting structural differences between CD1597 (green, AlphaFold, UniProt ID: Q186F3) and PPEP-1 (magenta, PDB: 5A0P). ^{303}YRNN^{306} from CD1597 and ^{117}GGST^{120} from PPEP-1 are shown as sticks. Lower left panel: Close-up of the KEEN interaction interface of PPEP-1. In CD1597, Lys101 in PPEP-1 is substituted for a Pro residue. Glu184 and Glu185 are conserved in CD1597. Of note: the Asn at the P2 position of the PPEP-1 substrate involved in the KEEN interface is absent in the apo structure.

Another interesting difference between CD1597 and other PPEPs is a mutation of the residue Lys101 in PPEP-1 and Arg96 in PPEP-2. This residue is located in the S-loop that closes upon binding of a substrate [160]. This conformational change allows the Lys101 in PPEP-1 to hydrogen bond with the Glu184, Glu185, and the Asn at the P2 in the PPEP-1 substrate [160,162]. These interactions, collectively called the KEEN interface, are essential for proteolytic activity of PPEP-1 [162]. Moreover, substitution of the Lys101 with an Arg residue did not alter the activity due to the similar physicochemical properties of this residue [162]. In CD1597, this residue is substituted for a Pro, which is unable to produce similar interactions (**Figure 1**). Nevertheless, the two Glu residues

involved in the KEEN interface are conserved in CD1597. Substitution of the Pro with a Lys did not produce a proteolytically active enzyme in the WT PPEP-like domain and the other mutants described above (data not shown).

In the end, we tested CD1597 for activity using the full-length protein, the C-terminal PPEP-like domain, and mutants of this domain that render CD1597 more PPEP-1-like. However, we observed no activity in the assays with Pro-Pro-containing FRET-quenched peptides, casein, XXPPXX (X=any residue except Cys) combinatorial peptide libraries, and XXXXXX (X=any residue except Cys) combinatorial peptide libraries (**Chapters 3 and 5** and unpublished results). Also, the experiments using the TAILS method did not reveal any activity.

Based on our assays, we conclude that CD1597 does not exhibit (Pro-Pro) endoproteolytic activity. However, we did not rule out the possibility that CD1597 might possess exoproteolytic activity. To investigate this, future studies could, for example, incubate CD1597 with peptides of six or fewer residues in length, which should fit in the active site. Additionally, an experimentally determined crystal structure of CD1597 could help in predicting substrates, but efforts to obtain this structure have so far been unsuccessful. Alternatively, CD1597 could be a ligand-binding protein that lacks proteolytic activity altogether. To investigate this, pull-down experiments using CD1597 might identify a binding partner.

Surprisingly, a recent bottom-up MS analysis of *C. difficile* grown in *C. difficile* minimal medium (CDMM) identified 4 peptides of CD1597 (unpublished data). Before, we only identified CD1597 in analyses of spores. Possibly, CD1597 expression is higher in CDMM, but our results using a reporter assay for promoter activity did not indicate increased expression in **Chapter 3**. However, the identification of only 4 peptides indicates a low expression that might not be discernible from the reporter signals in the control strain. CDMM is a defined minimal medium that contains the minimum amount of nutrients needed for survival [329] and could lead to nutrient stress more quickly than rich media. Since sporulation is regulated by nutrient availability and CD1597 is identified in spores ([64,78] and **Chapter 3**), CDMM might promote sporulation and therefore the expression of CD1597.

Applications of Pro-Pro endopeptidases in research, clinical settings, and industry

Applications in research

In research, highly specific proteases are often used to remove affinity tags during protein purification. The proteases used for this application need to be highly specific in order to prevent degradation of the purified protein. In addition, protein stability, broad compatibility with reagents, and activity over a wide pH range are favorable characteristics. Factor Xa, thrombin, TEV protease, and enteropeptidase are among the most widely used proteases for certain applications. However, factor Xa and thrombin, both serine proteases involved in blood coagulation, are promiscuous and can therefore degrade the protein of interest in purification routines [330]. In addition, factor Xa and thrombin need to be activated post-purification to produce an active enzyme [331,332], complicating the production of these enzymes. Enteropeptidase (also known as enterokinase) has the advantage that the P1' position tolerates any residue except proline and tryptophan, allowing complete removal of an affinity tag [333]. However, promiscuity is also a problem for enteropeptidase since it shows proteolytic activity towards unexpected sequences [334,335]. In addition, the protein contains four disulfide bonds that are essential for proteolytic activity, thus the use of enteropeptidase is incompatible with DTT [336]. The viral TEV protease cleaves the consensus sequence Glu-Xaa-Xaa-Tyr-Xaa-Gln↓Ser/Gly (P6-P1'), although several studies indicate more stringency for the P4 and P2 positions while less stringency has been observed for the P1' position [337–339]. Nevertheless, the many positions surrounding the cleavage site make the TEV protease highly specific.

The high specificity of PPEPs makes them suitable candidates for the removal of affinity tags. For example, the prime-side specificity of PPEP-3 almost exclusively tolerates prolines (**Chapter 4**). In addition, these proteases are easy to produce recombinantly, do not necessitate activation, are regarded as thermostable, and do not contain disulfide bonds. However, since specificity depends on the P3-P3' positions, PPEP-1 cleavage would leave a remnant of three residues attached to the purified protein. Additional investigations regarding the active pH range, optimal temperature, suitable buffers, compatibility with other reagents, and enzyme kinetics are needed to determine if PPEPs are more suitable for the removal of affinity tags than the current proteases used for this purpose.

Applications in a clinical setting

PPEP-1 is the most interesting PPEP for use in a clinical setting since it originates from the clinically relevant gut pathogen *C. difficile*. *C. difficile* attaches itself to the intestinal epithelium through CD2831 and CD3246, although the binding ligand for CD3246 remains unknown [147]. A PPEP-1 deficient strain demonstrated attenuated virulence in a hamster infection model [147]. This observation was likely an underestimation due to the growth advantage of ClosTron-generated mutants compared to wild-type cells in animals pre-treated with clindamycin [147,224]. ClosTron mutagenesis introduces an erythromycin resistance gene that confers cross-resistance to clindamycin, which likely increases the survival rate *in vivo* in clindamycin pretreated animals. The attenuated virulence is thought to result from decreased colonization efficiency, as the strain cannot detach from the gut wall. Conversely, constitutive expression of PPEP-1 might also result in reduced virulence by preventing adherence through CD2831 and CD3246. Future research using infection models with a *C. difficile* strain that constitutively expresses PPEP-1 might provide evidence supporting this hypothesis. If high levels of PPEP-1 lead to reduced virulence, therapy involving the administration of recombinantly produced PPEP-1 might alleviate the symptoms of *C. difficile* infection (CDI).

Since PPEP-1 is secreted in the colon, stool samples of CDI patients might contain the active enzyme, which would allow for the use of PPEP-1 as a biomarker for CDI. For example, the soluble fraction of stool samples of CDI patients could be added to a peptide labeled with chromogenic or fluorogenic groups. Cleavage of such peptides should indicate the presence of PPEP-1 by a change in color or fluorescence/luminescence. One study that investigated the use of PPEP-1 as a biomarker using a quenched NanoLuciferase showed the detection of luminescence at PPEP-1 concentrations as low as 10 nM [340]. In this study, NP↓PVPP (P2-P4') was used as a linker between the luciferase molecule and the quencher. However, our current understanding of PPEP-1 specificity would suggest a linker consisting of VNP↓PPP (P3-P3') for optimal sensitivity. Thus, to design such diagnostic assays, a deep understanding of the substrate specificity is needed. On the other hand, the reaction must be very selective for PPEP-1. Nevertheless, when using a FRET-quenched substrate containing the optimal PPEP-1 cleavage site, i.e., VNPPPP (P3-P3'), it is unlikely that other proteases will show activity towards this peptide. In order to develop such a diagnostic tool for the detection of CDI, it is important to identify the presence of active PPEP-1 in stool samples of patients with CDI. In addition, stool samples of healthy people should not be active toward the peptide used for the detection of PPEP-1.

Applications in industry

In industry, proteases are used for many processes. In the food industry, proteases are used to tenderize meat, clarify beer, degrade gluten, and produce dairy products [299–301,341]. In addition, proteases are used in the leather and textile industry to improve materials [342–344]. In detergents, proteases are used to degrade protein-based stains, enhancing the efficiency of laundry and dishwashing products [345]. Their application extends to waste management as well, where proteases aid in the decomposition of organic waste, contributing to more efficient and environmentally friendly waste processing methods [346]. For many of these processes, the proteases must degrade multiple substrates as much as possible. For example, the widely used papain, a promiscuous protease cleaving after Lys and Arg residues, is used for many applications in the food, pharmaceutical, cosmetics, leather, and textile industry due to its low substrate specificity [347].

The high specificity of PPEPs makes these proteases unlikely candidates to be used in industry for general applications that require broad substrate degradation. PPEPs cleave at specific proline-rich sites, limiting their use in processes where the breakdown of a wide range of proteins is necessary. There are, however, several groups of industrially relevant proline-rich proteins. For example, the gliadins and glutenins that form the group of gluten, the extracellular matrix proteins elastin and collagen, and the hordeins in beer contain many prolines in their sequences. The use of PPEPs might be useful for the industrial degradation of some of these proteins. Glutenin in wheat (*Triticum aestivum*) consists of high- and low-molecular-weight subunits. Inspection of these subunits reveals many Pro-Pro sequences, although no obvious PPEP cleavage sites are present (i.e., PPXP). In addition, bovine and porcine collagen contains many Pro-Pro sites, and several PPXP sites. However, the use of PPEPs in industrial applications such as gluten degradation and meat tenderization most likely necessitate mutations that increase the flexibility at the P3, P2, P2', and P3' positions. A thorough understanding of PPEP specificity could aid in predicting the mutations that render the PPEP specificity more permissible, e.g., by reducing steric hindrances. Also, directed evolution of proteases could aid in obtaining a protease that displays flexibility at the P3, P2, P2', and P3' positions while remaining very specific for Pro-Pro cleavage [348]. Alternatively, the specificity of PPEPs can be advantageous in niche applications where precise proteolysis is required.

Profiling protease specificity using synthetic combinatorial peptide libraries and mass spectrometry

A comparison of synthetic combinatorial peptide libraries with alternative libraries

In **Chapters 4, 5, and 6** we characterized PPEP specificity in detail using a novel method that combines the use of synthetic combinatorial peptide libraries with mass spectrometry. However, other types of peptide libraries exist that can be employed for the purpose of protease specificity profiling, each with its strengths and weaknesses.

The PPEP-1 substrates were discovered using a collection of synthetic peptides [146]. These synthetic peptides can contain a FRET pair, allowing both detection and quantification of proteolytic activity by a protease of interest. However, the main drawbacks are the difficulty in creating a large and diverse library and the laborious process required to test each peptide individually.

Proteome-derived peptide libraries are produced by enzymatic digestion of a proteome and therefore the production is straightforward and inexpensive. Treatment of these libraries with a protease of interest yields product peptides which are analyzed by MS [24,252,253]. The product peptides can be enriched by chemical modification of the neo-N-termini [24] or separated based on hydrophobicity [252] or charge state [253]. An advantage of proteome-derived libraries is that they allow for the identification of non-prime-side sequences after the identification of prime-side sequences, which are automatically detected through database searches [24]. In addition, the identified substrates might represent biologically relevant substrates. However, proteome-derived libraries cannot be tailored to specific needs, e.g., to focus on proline-rich peptides. In addition, proteome-derived libraries do not offer a quantitative approach since the amount of a substrate peptide is determined by the amount of the protein in the organism.

Substrate phage display (SPD) peptide libraries have also been used to characterize protease specificity [169,170,349,350]. A typical approach to profile protease specificity using SPD libraries starts with the production of a library that contains phages displaying a large diversity of peptides on their surface. These peptides contain a C-terminal biotin tag, allowing the phages to be immobilized on streptavidin beads. Cleavage of the peptide releases the phages (the protease-sensitive pool), which can be amplified and used in repetitive rounds of selection for phages displaying the protease substrates. After multiple rounds of selection, PCR is used to amplify the substrate encoding DNA which is then analyzed by sequencing. The DNA is then translated to identify the cleavage sites.

The main advantage of SPD libraries is their high diversity, which can encompass billions of unique peptides. In addition, the libraries can be tailored to contain fully randomized (i.e., combinatorial) peptides or sequences originating from a proteome [349]. Furthermore, since the libraries are genetically encoded, propagation of the library is inexpensive and can be done using simple techniques. However, from designing the library to obtaining cleavage motifs is an arduous process with many steps. In addition, phage display libraries can exhibit biases in the displayed sequences due to the limitations of the host bacteria's processing machinery or phage characteristics [351,352], affecting the ability of SPD approaches to be truly quantitative. Furthermore, SPD approaches only identify the sequence that was cleaved but not the cleavage site, since DNA is used as a readout. Therefore, an additional challenge lies in the requirement for downstream validation, as identified sequences must be synthesized and tested in solution to confirm their (site of) proteolysis.

Synthetic combinatorial peptide libraries offer a diversity similar to that of SPD libraries, depending on the number of varied positions in the library. A major advantage of synthetic libraries is that they can be precisely designed for a protease of interest or a research question. Not only can certain positions be fixed residues, also different tags, glycosylated amino acids, and non-proteinogenic amino acids (e.g., hydroxyproline) can be incorporated.

Another benefit of synthesizing a library is that the peptides are produced in an equimolar manner, allowing for a quantitative approach. However, since the product peptides in our assays in **Chapters 4, 5, and 6** were analyzed using LC-MS/MS, quantification of the peptides is dependent on their ionization efficiency. For example, the logo in **Chapter 5** for PPEP-1 show the high abundance of the His residue (His is protonated in our experimental setup) at the P2 position, while assays using FRET-quenched peptides showed that an Asn residue is preferred at this site. The increased electrospray ionization (ESI) response factor of the His-, Arg-, and Lys-containing peptides poses a challenge for quantification in MS, especially for non-tryptic peptides (tryptic peptides possess a single Lys or Arg residue at the C-terminus by definition). Proteins containing many of these basic residues can be overrepresented in MS analyses due to the high ESI response factor of their peptides. The difference in ionization efficiency of peptides is a general limitation of quantitative MS-based approaches. Currently, models are being trained to predict the ESI response factor for peptides [353]. In the future, such models might be applied in post-analysis processes to normalize the signals based on their predicted ESI response factor.

A slight drawback of synthetic combinatorial peptide libraries is that, unlike proteome-derived and SPD libraries, they do not contain substrates derived from a biological source. However, detailed information on the substrate specificity could aid in

identifying endogenous substrates. This was true for PPEP-2, for which the logo in **Chapter 5** clearly showed a preference for the endogenous cleavage site PLPPVP (P3-P3'). However, for the newly characterized PPEPs, PPEP-3 and PPEP-4, substrate identification was not straightforward. Analysis of the *G. thermodenitrificans* and *A. tepidamans* proteome revealed two potential substrates for PPEP-3 and a single potential substrate for PPEP-4. These potential substrates do not reflect the PPEP-1 and PPEP-2 substrates, i.e., they are not (predicted) adhesins. Additional experiments such as proteolytic surface-shaving [290] of *G. thermodenitrificans* and *A. tepidamans* with PPEP-3 and PPEP-4, respectively, could confirm cleavage of these proteins or other proteins from these organisms. Alternatively, the true substrates of PPEP-3 and PPEP-4 might originate from other species, which will greatly complicate their identification.

Alternative application of synthetic combinatorial peptide libraries

Correct annotation of the product peptide signals in the EICs (**Chapters 4, 5 and 6**) was complicated by the presence of isomeric peptides. These peptides included peptides that had two residues switched (e.g., PHPGGLEEF and PPHGGLEEF) and peptides with different lengths (e.g., PPPGGLEEF and YKGGLEEF). To discriminate between these isomeric peptides, good quality MS2 spectra containing the discriminatory fragments were essential. However, to discriminate between the isomeric residues Leu and Ile, additional assays investigating the cleavage of FRET-quenched peptides and/or retention times of synthetic peptides were necessary.

In **Chapter 6**, we investigated the separation of the peptides PTEDAVLLP, PTEDAVILP, PTEDAVLIP, and PTEDAVIIP. To our surprise, we were able to fully separate these peptides on a C18 column in our LC-MS/MS analysis. However, the post-analysis process is currently not able to discriminate between the four peptides. Possibly, improved algorithms that predict the retention time of a peptide can allow for automatic annotation of these peptides. For this, synthetic combinatorial peptide libraries might be useful for the improvement of retention time prediction models.

The retention time of a (unmodified) peptide within a given experimental setup is determined by the sequence, which can differ in amino acid composition, modifications, length, and three-dimensional conformation [354–356]. The retention time can be used in post-analysis algorithms to increase peptide identification and to discriminate between correct and incorrect peptide-spectrum matches (PSMs) [357–359]. In these algorithms, the retention time is predicted, often by machine-learning methods which are trained on a set of training peptides [360]. The effectivity of a retention time prediction algorithm is dependent on the set of training peptides. For example, training an algorithm using solely tryptic peptides will optimize it for tryptic peptides, but can fall

short when predicting retention times for non-tryptic peptides. Combinatorial peptide libraries offer a very high degree of variability and might be a valuable source for peptides to train machine-learning-based prediction algorithms. An important feature of a suitable combinatorial peptide library should be the variation in the length of the peptide since this is a key determinant of the retention time.

The use of artificial intelligence in research

In **Chapters 2, 3, 4, and 5** we made use of protein structures that were predicted by AlphaFold, an artificial intelligence (AI) system developed by Google Deepmind that uses the primary structure of proteins to predict the three-dimensional conformation [361]. Although these structures are predictions, AlphaFold's accuracy is comparable to experimentally determined structures [361]. Currently, there are over 200 million predicted structures available in the AlphaFold database, a number that is impossible to obtain using experimental techniques. **Figure 2** shows a comparison between the cocrystal structure of PPEP-3 which was determined by X-ray crystallography (**Chapter 6**) and the predicted structure. Slight differences are observed between the two structures, for example in the position of the S-loop that closes upon substrate binding, yet the highly structured α-helices and β-sheets align perfectly. Advances in predictive protein folding methods could increase the accuracy of these systems even further. However, these advances might also prove valuable for predicting multimers [362], substrate docking [363], drug design [364], and *de novo* protein design [365].

The work presented in this thesis was performed during a bridging period between a world where the use of AI tools was only limited to experienced computer scientists and a world where AI applications are ubiquitous and used for everyday purposes. If implemented correctly, the scientific community can benefit immensely from the use of AI in several ways.

Research is getting more and more expensive. Materials have increased in price, cutting-edge instruments cost more than their predecessors, and wages tend to increase yearly. However, in the Netherlands, the coalition formed after the 2023 parliamentary elections is planning to reduce the budget for scientific research by 150 million euros (~10% of the total budget) per year until the year 2031. In public research organizations, the largest part of the expenses (62%) goes to personnel costs according to a 2011 study commissioned by the European Commission [366]. Therefore, to produce a similar scientific output, an increase in efficiency is necessary. The use of AI, especially in the form of large language models (LLMs) such as OpenAI's ChatGPT, Microsoft's Copilot, and Meta's LLaMa, could increase the efficiency of individual researchers by saving time on time-consuming tasks.

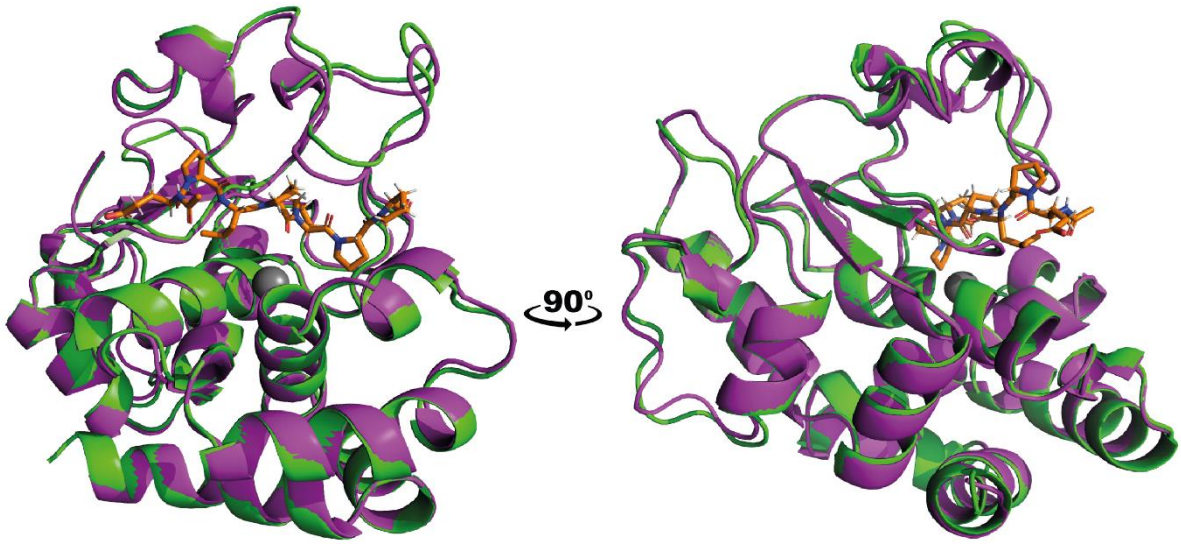


Figure 2. Comparison of the cocrystal structure and the AlphaFold prediction of PPEP-3. The experimentally determined cocrystal structure of PPEP-3 (**Chapter 6**) is shown in purple and the substrate peptide Ac-EPLPPPP as sticks in orange. The predicted structure from AlphaFold (UniProt ID: A4INY2) is shown in green.

Documentation of research in publicly available scientific articles, reviews, and theses is a fundamental part of science. However, writing can be a time-consuming activity, and texts need to be screened for correct grammar and spelling errors. First, we had Microsoft Word's autocorrect to help us with these tasks. Then, more sophisticated programs such as Grammarly and ProWritingAid could aid in improving clarity, readability, and engagement. Now, LLMs can take over these tasks by not only improving written text but can generate text themselves. Moreover, LLMs excel in scanning information and bringing this together, possibly making them ideal tools for writing systematic reviews. AI could not only save time for the writer but also the reader. A well-constructed sentence/paragraph is easier to read than a complicated one. In addition, AI-powered academic search engines such as Consensus scan millions of scientific articles and are useful for answering questions based on scientific literature, finding the right article to read, and providing references for a statement. Furthermore, AI can suggest lines of research, assist bioinformatics in writing and checking code, and can perform statistics based on a description of a study or experiment.

However, we must be wary of the downsides of this new technology. Possibly, AI can be used to generate fake data or articles easily. Also, delegating tasks to AI could result in a loss of skill when people no longer perform these tasks themselves. In addition, uncareful use of AI such as adding suggested references without validation can lead to sloppy work. Furthermore, AI algorithms could amplify biases present in data, the tasks

performed by AI can lack transparency, and there are unanswered questions regarding intellectual property and ownership.

Concluding remarks

The work presented in this thesis aimed to elucidate the function, structure, and specificity of both previously characterized and novel PPEPs. Our findings suggest that the roles of PPEPs extend beyond those of PPEP-1 and PPEP-2, which are known for cleaving adhesion proteins. Specifically, the candidate substrates for PPEP-3 and PPEP-4 do not include endogenous adhesion proteins or might originate from other organisms. Additionally, our investigations of the PPEP-homolog from *C. difficile*, CD1597, suggest a previously unknown function due to its presence in spores and the lack of (endo)proteolytic activity. Further research into PPEP-3, PPEP-4, CD1597, and currently uncharacterized PPEPs could shed more light on the diversity of roles played by these unique bacterial proteases.

The unique specificity of PPEPs prompted us to develop a novel PPEP-specific method that combined the use of synthetic combinatorial peptide libraries with MS. This approach offers several advantages over alternative methods, including the high diversity, customized design, and sensitive detection of product peptides. Moreover, experiments using our library method can be completed within two days and deliver excellent reproducibility. We believe that similar approaches can be readily adapted to study other groups of proteases and we look forward to seeing this happen in the future.

By relating the specificity of a protease to the atomic structure, either experimentally determined or algorithmically predicted, we explained the observed preferences for specific residues surrounding the cleavage site. Even minor structural differences, such as a single amino acid substitution, can significantly impact protease specificity. A thorough understanding of the structure-function relationship of proteases can aid in the design of enzymes with tailored specificities, which can be beneficial for applications in research, industry, and healthcare.

References

1 Nothaft H & Szymanski CM (2010) Protein glycosylation in bacteria: sweeter than ever. *Nature Reviews Microbiology 2010 8:11* **8**, 765–778.

2 Valguarnera E, Kinsella RL & Feldman MF (2016) Sugar and Spice Make Bacteria Not Nice: Protein Glycosylation and Its Influence in Pathogenesis. *J Mol Biol* **428**, 3206–3220.

3 Baker TA & Sauer RT (2006) ATP-dependent proteases of bacteria: recognition logic and operating principles. *Trends Biochem Sci* **31**, 647–653.

4 Queraltó C, Álvarez R, Ortega C, Díaz-Yáñez F, Paredes-Sabja D & Gil F (2023) Role and Regulation of Clp Proteases: A Target against Gram-Positive Bacteria. *Bacteria 2023, Vol 2, Pages 21-36* **2**, 21–36.

5 Caminero A, Guzman M, Libertucci J & Lomax AE (2023) The emerging roles of bacterial proteases in intestinal diseases. *Gut Microbes* **15**.

6 Hartley BS (1960) Proteolytic Enzymes. *Annu Rev Biochem* **29**, 45–72.

7 Seemüller E, Lupas A, Stock D, Löwe J, Huber R & Baumeister W (1995) Proteasome from Thermoplasma acidophilum: A threonine protease. *Science (1979)* **268**, 579–582.

8 Blow DM, Birktoft JJ & Hartley BS (1969) Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature* **221**, 337–340.

9 Vernet T, Tessier DC, Chatellier J, Plouffe C, Tak Sing Lee, Thomas DY, Storer AC & Menard R (1995) Structural and functional roles of asparagine 175 in the cysteine protease papain. *Journal of Biological Chemistry*.

10 Huber EM, Heinemeyer W, Li X, Arendt CS, Hochstrasser M & Groll M (2016) A unified mechanism for proteolysis and autocatalytic activation in the 20S proteasome. *Nat Commun* **7**, 1–10.

11 Hedstrom L (2002) Serine protease mechanism and specificity. *Chem Rev* **102**, 4501–4523.

12 Polgár L (1987) The mechanism of action of aspartic proteases involves 'push-pull' catalysis. *FEBS Lett* **219**, 1–4.

13 Hooper NM (1994) Families of zinc metalloproteases. *FEBS Lett* **354**, 1–6.

14 Menach E, Hashida Y, Yasukawa K & Inouye K (2013) Effects of Conversion of the Zinc-Binding Motif Sequence of Thermolysin, HEXXH, to That of Dipeptidyl Peptidase III, HEXXXH, on the Activity and Stability of Thermolysin. *Biosci Biotechnol Biochem* **77**, 1901–1906.

15 Cerdà-Costa N & Gomis-Rüth FX (2014) Architecture and function of metallopeptidase catalytic domains. *Protein Science* **23**, 123–144.

16 Carrington JC & Dougherty WG (1988) A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. *Proceedings of the National Academy of Sciences* **85**, 3391–3395.

17 Schechter I & Berger A (1967) On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* **27**, 157–162.

18 Carmona AK, Juliano MA & Juliano L (2009) The use of Fluorescence Resonance Energy Transfer (FRET) peptides for measurement of clinically important proteolytic enzymes. *An Acad Bras Cienc* **81**, 381–392.

19 Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL & Wells JA (2008) Global Sequencing of Proteolytic Cleavage Sites in Apoptosis by Specific Labeling of Protein N Termini. *Cell* **134**, 866–876.

20 Staes A, van Damme P, Timmerman E, Ruttens B, Stes E, Gevaert K & Impens F (2017) Protease Substrate Profiling by N-Terminal COFRADIC. *Methods Mol Biol* **1574**, 51–76.

21 Rano TA, Timkey T, Peterson EP, Rotonda J, Nicholson DW, Becker JW, Chapman KT & Thornberry NA (1997) A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). *Chem Biol* **4**, 149–155.

22 O'Donoghue AJ, Alegra Eroy-Reveles AA, Knudsen GM, Ingram J, Zhou M, Statnekov JB, Greninger AL, Hostetter DR, Qu G, Maltby DA, Anderson MO, Derisi JL, McKerrow JH, Burlingame AL & Craik CS (2012) Global identification of peptidase specificity by multiplex substrate profiling. *Nature Methods 2012 9:11* **9**, 1095–1100.

23 Kleifeld O, Doucet A, Auf Dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN & Overall CM (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature Biotechnology 2010 28:3* **28**, 281–288.

24 Schilling O & Overall CM (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature Biotechnology 2008 26:6* **26**, 685–694.

25 Bekker-Jensen DB, Martínez-Val A, Steigerwald S, Rüther P, Fort KL, Arrey TN, Harder A, Makarov A & Olsen J V. (2020) A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Molecular & Cellular Proteomics* **19**, 716–729.

26 Lawson PA, Citron DM, Tyrrell KL & Finegold SM (2016) Reclassification of Clostridium difficile as Clostridioides difficile (Hall and O'Toole 1935) Prévot 1938. *Anaerobe* **40**, 95–99.

27 Smits WK, Lyras D, Lacy DB, Wilcox MH & Kuijper EJ (2016) Clostridium difficile infection. *Nature Reviews Disease Primers 2016 2:1* **2**, 1–20.

28 Czepiel J, Dróżdż M, Pituch H, Kuijper EJ, Perucki W, Mielimonka A, Goldman S, Wultańska D, Garlicki A & Biesiada G (2019) Clostridium difficile infection: review. *European Journal of Clinical Microbiology and Infectious Diseases* **38**, 1211–1221.

29 Leffler DA & Lamont JT (2015) Clostridium difficile Infection. *New England Journal of Medicine* **372**, 1539–1548.

30 Pépin J, Saheb N, Coulombe MA, Alary ME, Conriveau MP, Authier S, Leblanc M, Rivard G, Bettez M, Primeau V, Nguyen M, Jacob CÉ & Lanthier L (2005) Emergence of fluoroquinolones as the predominant risk factor for Clostridium difficile-associated diarrhea: A cohort study during an epidemic in Quebec. *Clinical Infectious Diseases* **41**, 1254–1260.

31 Álvarez-Pérez S, Blanco JL, Martínez-Nevado E, Peláez T, Harmanus C, Kuijper E & García ME (2014) Shedding of Clostridium difficile PCR ribotype 078 by zoo animals, and report of an unstable metronidazole-resistant isolate from a zebra foal (Equus quagga burchellii). *Vet Microbiol* **169**, 218–222.

32 Knetsch CW, Connor TR, Mutreja A, van Dorp SM, Sanders IM, Browne HP, Harris D, Lipman L, Keessen EC, Corver J, Kuijper EJ & Lawley TD (2014) Whole genome sequencing reveals potential spread of Clostridium difficile between humans and farm animals in the Netherlands, 2002 to 2011. *Euro Surveill* **19**, 1–12.

33 Schneeberg A, Rupnik M, Neubauer H & Seyboldt C (2012) Prevalence and distribution of Clostridium difficile PCR ribotypes in cats and dogs from animal shelters in Thuringia, Germany. *Anaerobe* **18**, 484–488.

34 Simango C & Mwakurudza S (2008) Clostridium difficile in broiler chickens sold at market places in Zimbabwe and their antimicrobial susceptibility. *Int J Food Microbiol* **124**, 268–270.

35 Shivaprasad HL (2003) Hepatitis associated with Clostridium difficile in an ostrich chick. *Avian Pathology* **32**, 57–62.

36 Kochan TJ, Shoshiev MS, Hastie JL, Somers MJ, Plotnick YM, Gutierrez-Munoz DF, Foss ED, Schubert AM, Smith AD, Zimmerman SK, Carlson PE & Hanna PC (2018) Germinant Synergy Facilitates Clostridium difficile Spore Germination under Physiological Conditions. *mSphere* **3**.
37 Borriello SP & Barclay FE (1986) An in-vitro model of colonisation resistance to Clostridium difficile infection. *J Med Microbiol* **21**, 299–309.
38 Vollaard EJ & Clasener HAL (1994) Colonization resistance. *Antimicrob Agents Chemother* **38**, 409.
39 Lyras D, O'Connor JR, Howarth PM, Sambol SP, Carter GP, Phumoonna T, Poon R, Adams V, Vedantam G, Johnson S, Gerding DN & Rood JI (2009) Toxin B is essential for virulence of Clostridium difficile. *Nature* **458**, 1176.
40 Kuehne SA, Cartman ST, Heap JT, Kelly ML, Cockayne A & Minton NP (2010) The role of toxin A and toxin B in Clostridium difficile infection. *Nature 2010 467:7316* **467**, 711–713.
41 Kuehne SA, Cartman ST & Minton NP (2011) Both, toxin A and toxin B, are important in Clostridium difficile infection. *Gut Microbes* **2**, 252–255.
42 Torres JF (1991) Purification and characterisation of toxin B from a strain of Clostridium difficile that does not produce toxin A. *J Med Microbiol* **35**, 40–44.
43 Alfa MJ, Kabani A, Lyerly D, Moncrief S, Neville LM, Al-Barrak A, Harding GKH, Dyck B, Olekson K & Embil JM (2000) Characterization of a toxin A-negative, toxin B-positive strain of Clostridium difficile responsible for a nosocomial outbreak of Clostridium difficile-associated diarrhea. *J Clin Microbiol* **38**, 2706–2714.
44 Chaves-Olarte E, Freer E, Parra A, Guzmán-Verri C, Moreno E & Thelestam M (2003) R-Ras glucosylation and transient RhoA activation determine the cytopathic effect produced by toxin B variants from toxin A-negative strains of Clostridium difficile. *J Biol Chem* **278**, 7956–7963.
45 Monot M, Eckert C, Lemire A, Hamiot A, Dubois T, Tessier C, Dumoulard B, Hamel B, Petit A, Lalande V, Ma L, Bouchier C, Barbut F & Dupuy B (2015) Clostridium difficile: New Insights into the Evolution of the Pathogenicity Locus. *Sci Rep* **5**.
46 Gerding DN, Johnson S, Rupnik M & Aktories K (2014) Clostridium difficile binary toxin CDT. *Gut Microbes* **5**, 15–27.
47 Aktories K, Papatheodorou P & Schwan C (2018) Binary Clostridium difficile toxin (CDT) - A virulence factor disturbing the cytoskeleton. *Anaerobe* **53**, 21–29.
48 Fluit AC, Wolfhagen MJHM, Verdonk GPHT, Jansze M, Torensma R & Verhoef J (1991) Nontoxigenic strains of Clostridium difficile lack the genes for both toxin A and toxin B. *J Clin Microbiol* **29**, 2666.
49 Natarajan M, Walk ST, Young VB & Aronoff DM (2013) A Clinical and Epidemiological Review of Non-toxigenic Clostridium difficile. *Anaerobe* **22**, 1.
50 Voth DE & Ballard JD (2005) Clostridium difficile Toxins: Mechanism of Action and Role in Disease. *Clin Microbiol Rev* **18**, 247.
51 Shen A (2012) Clostridium difficile toxins: mediators of inflammation. *J Innate Immun* **4**, 149–158.
52 Yu H, Chen K, Sun Y, Carter M, Garey KW, Savidge TC, Devaraj S, Tessier ME, Von Rosenvinge EC, Kelly CP, Pasetti MF & Feng H (2017) Cytokines Are Markers of the Clostridium difficile-Induced Inflammatory Response and Predict Disease Severity. *Clin Vaccine Immunol* **24**.
53 Kelly CP & Kyne L (2011) The host immune response to Clostridium difficile. *J Med Microbiol* **60**, 1070–1079.

54 Braun V, Hundsberger T, Leukel P, Sauerborn M & Von Eichel-Streiber C (1996) Definition of the single integration site of the pathogenicity locus in Clostridium difficile. *Gene* **181**, 29–38.
55 Mani N, Lyras D, Barroso L, Howarth P, Wilkins T, Rood JI, Sonenshein AL & Dupuy B (2002) Environmental response and autoregulation of Clostridium difficile TxeR, a sigma factor for toxin gene expression. *J Bacteriol* **184**, 5971–5978.
56 Govind R & Dupuy B (2012) Secretion of Clostridium difficile toxins A and B requires the holin-like protein TcdE. *PLoS Pathog* **8**.
57 Dupuy B, Govind R, Antunes A & Matamouros S (2008) Clostridium difficile toxin synthesis is negatively regulated by TcdC. *J Med Microbiol* **57**, 685–689.
58 Matamouros S, England P & Dupuy B (2007) Clostridium difficile toxin expression is inhibited by the novel regulator TcdC. *Mol Microbiol* **64**, 1274–1288.
59 Bakker D, Smits WK, Kuijper EJ & Corver J (2012) TcdC Does Not Significantly Repress Toxin Expression in Clostridium difficile 630ΔErm. *PLoS One* **7**, e43247.
60 Cartman ST, Kelly ML, Heeg D, Heap JT & Minton NP (2012) Precise manipulation of the Clostridium difficile chromosome reveals a lack of association between the tcdC genotype and toxin production. *Appl Environ Microbiol* **78**, 4683–4690.
61 Majumdar A & Govind R (2022) Regulation of Clostridioides difficile toxin production. *Curr Opin Microbiol* **65**, 95–100.
62 Underwood S, Guan S, Vijayasubhash V, Baines SD, Graham L, Lewis RJ, Wilcox MH & Stephenson K (2009) Characterization of the Sporulation Initiation Pathway of Clostridium difficile and Its Role in Toxin Production. *J Bacteriol* **191**, 7296.
63 El Meouche I, Peltier J, Monot M, Soutourina O, Pestel-Caron M, Dupuy B & Pons JL (2013) Characterization of the SigD Regulon of C. difficile and Its Positive Control of Toxin Production through the Regulation of tcdR. *PLoS One* **8**, e83748.
64 Martin-Verstraete I, Peltier J & Dupuy B (2016) The Regulatory Networks That Control Clostridium difficile Toxin Synthesis. *Toxins (Basel)* **8**.
65 Karlsson S, Dupuy B, Mukherjee K, Norin E, Burman LG & Åkerlund T (2003) Expression of Clostridium difficile toxins A and B and their sigma factor TcdD is controlled by temperature. *Infect Immun* **71**, 1784–1793.
66 Dineen SS, Villapakkam AC, Nordman JT & Sonenshein AL (2007) Repression of Clostridium difficile toxin gene expression by CodY. *Mol Microbiol* **66**, 206–219.
67 Dineen SS, McBride SM & Sonenshein AL (2010) Integration of metabolism and virulence by Clostridium difficile CodY. *J Bacteriol* **192**, 5350–5362.
68 Antunes A, Camiade E, Monot M, Courtois E, Barbut F, Sernova N V., Rodionov DA, Martin-Verstraete I & Dupuy B (2012) Global transcriptional control by glucose and carbon regulator CcpA in Clostridium difficile. *Nucleic Acids Res* **40**, 10701.
69 Antunes A, Martin-Verstraete I & Dupuy B (2011) CcpA-mediated repression of Clostridium difficile toxin gene expression. *Mol Microbiol* **79**, 882–899.
70 Deakin LJ, Clare S, Fagan RP, Dawson LF, Pickard DJ, West MR, Wren BW, Fairweather NF, Dougan G & Lawley TD (2012) The Clostridium difficile spo0A Gene Is a Persistence and Transmission Factor. *Infect Immun* **80**, 2704.
71 Castro-Córdova P, Mora-Uribe P, Reyes-Ramírez R, Cofré-Araneda G, Orozco-Aguilar J, Brito-Silva C, Mendoza-León MJ, Kuehne SA, Minton NP, Pizarro-Guajardo M & Paredes-Sabja D (2021) Entry of spores into intestinal epithelial cells contributes to recurrence of Clostridioides difficile infection. *Nature Communications 2021 12:1* **12**, 1–18.

72 Chiu CW, Tsai PJ, Lee CC, Ko WC & Hung YP (2021) Inhibition of spores to prevent the recurrence of Clostridioides difficile infection - A possibility or an improbability? *Journal of Microbiology, Immunology and Infection* **54**, 1011–1017.
73 Fimlaid KA, Bond JP, Schutz KC, Putnam EE, Leung JM, Lawley TD & Shen A (2013) Global Analysis of the Sporulation Pathway of Clostridium difficile. *PLoS Genet* **9**, e1003660.
74 Pettit LJ, Browne HP, Yu L, Smits WK, Fagan RP, Barquist L, Martin MJ, Goulding D, Duncan SH, Flint HJ, Dougan G, Choudhary JS & Lawley TD (2014) Functional genomics reveals that Clostridium difficile Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**, 1–15.
75 DiCandia MA, Edwards AN, Jones JB, Swaim GL, Mills BD & McBride SM (2022) Identification of functional Spo0A residues critical for sporulation in Clostridioides difficile. *J Mol Biol* **434**, 167641.
76 Edwards AN, Wetzel D, DiCandia MA & McBride SM (2022) Three Orphan Histidine Kinases Inhibit Clostridioides difficile Sporulation. *J Bacteriol* **204**.
77 Edwards AN & Mcbride SM (2014) Initiation of Sporulation in Clostridium difficile: a Twist on the Classic Model. *FEMS Microbiol Lett* **358**, 110.
78 Nawrocki KL, Edwards AN, Daou N, Bouillaut L & McBride SM (2016) CodY-Dependent Regulation of Sporulation in Clostridium difficile. *J Bacteriol* **198**, 2113.
79 Edwards AN, Tamayo R & McBride SM (2016) A novel regulator controls Clostridium difficile sporulation, motility and toxin production. *Mol Microbiol* **100**, 954–971.
80 Paredes-Sabja D, Shen A & Sorg JA (2014) Clostridium difficile spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol* **22**, 406.
81 Saujet L, Pereira FC, Henriques AO & Martin-Verstraete I (2014) The regulatory network controlling spore formation in Clostridium difficile. *FEMS Microbiol Lett* **358**, 1–10.
82 Setlow P (2006) Spores of Bacillus subtilis: their resistance to and killing by radiation, heat and chemicals. *J Appl Microbiol* **101**, 514–525.
83 Nerber HN & Sorg JA (2024) The small acid-soluble proteins of spore-forming organisms: Similarities and differences in function. *Anaerobe* **87**, 102844.
84 Paidhungat M, Setlow B, Driks A & Setlow P (2000) Characterization of spores of Bacillus subtilis which lack dipicolinic acid. *J Bacteriol* **182**, 5505–5512.
85 Lauren Donnelly M, Fimlaid KA & Shen A (2016) Characterization of Clostridium difficile spores lacking either SpoVAC or dipicolinic acid synthetase. *J Bacteriol* **198**, 1694–1707.
86 Cowan AE, Olivastro EM, Koppel DE, Loshon CA, Setlow B & Setlow P (2004) Lipids in the inner membrane of dormant spores of Bacillus species are largely immobile. *Proc Natl Acad Sci U S A* **101**, 7733.
87 Warth AD & Strominger JL (1972) Structure of the Peptidoglycan from Spores of Bacillus subtilis. *Biochemistry* **11**, 1389–1396.
88 Popham DL, Gilmore ME & Setlow P (1999) Roles of low-molecular-weight penicillin-binding proteins in Bacillus subtilis spore peptidoglycan synthesis and spore properties. *J Bacteriol* **181**, 126–132.
89 Permpoonpattana P, Phetcharaburanin J, Mikelsone A, Dembek M, Tan S, Brisson MC, La Ragione R, Brisson AR, Fairweather N, Hong HA & Cutting SM (2013) Functional characterization of Clostridium difficile spore coat proteins. *J Bacteriol* **195**, 1492–1503.
90 Zeng J, Wang H, Dong M & Tian GB (2022) Clostridioides difficile spore: coat assembly and formation. *Emerg Microbes Infect* **11**, 2340.

91 Alves Feliciano C, Douché T, Giai Gianetto Q, Matondo M, Martin-Verstraete I & Dupuy B (2019) CotL, a new morphogenetic spore coat protein of Clostridium difficile. *Environ Microbiol* **21**, 984–1003.
92 Phetcharaburanin J, Hong HA, Colenutt C, Bianconi I, Sempere L, Permpoonpattana P, Smith K, Dembek M, Tan S, Brisson MC, Brisson AR, Fairweather NF & Cutting SM (2014) The spore-associated protein BclA1 affects the susceptibility of animals to colonization and infection by Clostridium difficile. *Mol Microbiol* **92**, 1025–1038.
93 Calderón-Romero P, Castro-Córdova P, Reyes-Ramírez R, Milano-Céspedes M, Guerrero-Araya E, Pizarro-Guajardo M, Olguín-Araneda V, Gil F & Paredes-Sabja D (2018) Clostridium difficile exosporium cysteine-rich proteins are essential for the morphogenesis of the exosporium layer, spore resistance, and affect C. difficile pathogenesis. *PLoS Pathog* **14**.
94 Shrestha R & Sorg JA (2018) Hierarchical recognition of amino acid co-germinants during Clostridioides difficile spore germination. *Anaerobe* **49**, 41–47.
95 Kochan TJ, Somers MJ, Kaiser AM, Shoshiev MS, Hagan AK, Hastie JL, Giordano NP, Smith AD, Schubert AM, Carlson PE & Hanna PC (2017) Intestinal calcium and bile salts facilitate germination of Clostridium difficile spores. *PLoS Pathog* **13**, e1006443.
96 Francis MB, Allen CA, Shrestha R & Sorg JA (2013) Bile Acid Recognition by the Clostridium difficile Germinant Receptor, CspC, Is Important for Establishing Infection. *PLoS Pathog* **9**, e1003356.
97 Lawler AJ, Lambert PA & Worthington T (2020) A Revised Understanding of Clostridioides difficile Spore Germination. *Trends Microbiol* **28**, 744–752.
98 Kochan TJ, Somers MJ, Kaiser AM, Shoshiev MS, Hagan AK, Hastie JL, Giordano NP, Smith AD, Schubert AM, Carlson PE & Hanna PC (2017) Intestinal calcium and bile salts facilitate germination of Clostridium difficile spores. *PLoS Pathog* **13**, e1006443.
99 Shrestha R, Cochran AM & Sorg JA (2019) The requirement for co-germinants during Clostridium difficile spore germination is influenced by mutations in yabG and cspA. *PLoS Pathog* **15**, e1007681.
100 Diaz OR, Sayer C V., Popham DL & Shen A (2018) Clostridium difficile Lipoprotein GerS Is Required for Cortex Modification and Thus Spore Germination . *mSphere* **3**.
101 Francis MB & Sorg JA (2016) Dipicolinic Acid Release by Germinating Clostridium difficile Spores Occurs through a Mechanosensing Mechanism. *mSphere* **1**.
102 Darkoh C, Dupont HL, Norris SJ & Kaplan HB (2015) Toxin synthesis by Clostridium difficile is regulated through quorum signaling. *mBio* **6**.
103 Ahmed UKB & Ballard JD (2022) Autoinducing peptide-based quorum signaling systems in Clostridioides difficile. *Curr Opin Microbiol* **65**, 81–86.
104 Slater RT, Frost LR, Jossi SE, Millard AD & Unnikrishnan M (2019) Clostridioides difficile LuxS mediates inter-bacterial interactions within biofilms. *Scientific Reports 2019 9:1* **9**, 1–15.
105 Paredes-Sabja D & Sarker MR (2012) Adherence of Clostridium difficile spores to Caco-2 cells in culture. *J Med Microbiol* **61**, 1208–1218.
106 Calderón-Romero P, Castro-Córdova P, Reyes-Ramírez R, Milano-Céspedes M, Guerrero-Araya E, Pizarro-Guajardo M, Olguín-Araneda V, Gil F & Paredes-Sabja D (2018) Clostridium difficile exosporium cysteine-rich proteins are essential for the morphogenesis of the exosporium layer, spore resistance, and affect C. difficile pathogenesis. *PLoS Pathog* **14**, e1007199.
107 Mora-Uribe P, Miranda-Cárdenas C, Castro-Córdova P, Gil F, Calderón I, Fuentes JA, Rodas PI, Banawas S, Sarker MR & Paredes-Sabja D (2016) Characterization of the Adherence of

Clostridium difficile Spores: The Integrity of the Outermost Layer Affects Adherence Properties of Spores of the Epidemic Strain R20291 to Components of the Intestinal Mucosa. *Front Cell Infect Microbiol* **6**, 99.

108 Castro-Córdova P, Mora-Uribe P, Reyes-Ramírez R, Cofré-Araneda G, Orozco-Aguilar J, Brito-Silva C, Mendoza-León MJ, Kuehne SA, Minton NP, Pizarro-Guajardo M & Paredes-Sabja D (2021) Entry of spores into intestinal epithelial cells contributes to recurrence of Clostridioides difficile infection. *Nat Commun* **12**.

109 Castro-Córdova P, Otto-Medina M, Montes-Bravo N, Brito-Silva C, Lacy DB & Paredes-Sabja D (2023) Redistribution of the Novel Clostridioides difficile Spore Adherence Receptor E-Cadherin by TcdA and TcdB Increases Spore Binding to Adherens Junctions. *Infect Immun* **91**.

110 Purcell EB, McKee RW, McBride SM, Waters CM & Tamayo R (2012) Cyclic diguanylate inversely regulates motility and aggregation in Clostridium difficile. *J Bacteriol* **194**, 3307–3316.

111 Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH & Breaker RR (2008) Riboswitches in Eubacteria Sense the Second Messenger Cyclic Di-GMP. *Science* **321**, 411.

112 Breaker RR (2012) Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol* **4**, 3566–3567.

113 Montange RK & Batey RT (2008) Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys* **37**, 117–133.

114 Bordeleau E, Fortier LC, Malouin F & Burrus V (2011) c-di-GMP Turn-Over in Clostridium difficile Is Controlled by a Plethora of Diguanylate Cyclases and Phosphodiesterases. *PLoS Genet* **7**.

115 Corver J, Cordo' V, van Leeuwen HC, Klychnikov OI & Hensbergen PJ (2017) Covalent attachment and Pro-Pro endopeptidase (PPEP-1)-mediated release of Clostridium difficile cell surface proteins involved in adhesion. *Mol Microbiol* **105**, 663–673.

116 Purcell EB, McKee RW, Courson DS, Garrett EM, McBride SM, Cheney RE & Tamayo R (2017) A Nutrient-Regulated Cyclic Diguanylate Phosphodiesterase Controls Clostridium difficile Biofilm and Toxin Production during Stationary Phase. *Infect Immun* **85**.

117 Bordeleau E, Purcell EB, Lafontaine DA, Fortier LC, Tamayo R & Burrus V (2015) Cyclic Di-GMP riboswitch-regulated type IV pili contribute to aggregation of Clostridium difficile. *J Bacteriol* **197**, 819–832.

118 Hengge R (2009) Principles of c-di-GMP signalling in bacteria. *Nature Reviews Microbiology 2009 7:4* **7**, 263–273.

119 Soutourina OA, Monot M, Boudry P, Saujet L, Pichon C, Sismeiro O, Semenova E, Severinov K, Le Bouguenec C, Coppée JY, Dupuy B & Martin-Verstraete I (2013) Genome-Wide Identification of Regulatory RNAs in the Human Pathogen Clostridium difficile. *PLoS Genet* **9**, 1003493.

120 Calabi E, Calabi F, Phillips AD & Fairweather NF (2002) Binding of Clostridium difficile Surface Layer Proteins to Gastrointestinal Tissues. *Infect Immun* **70**, 5770.

121 Merrigan MM, Venugopal A, Roxas JL, Anwar F, Mallozzi MJ, Roxas BAP, Gerding DN, Viswanathan VK & Vedantam G (2013) Surface-Layer Protein A (SlpA) is a major contributor to host-cell adherence of clostridium difficile. *PLoS One*.

122 Spigaglia P, Barketi-Klai A, Collignon A, Mastrantonio P, Barbanti F, Rupnik M, Janezic S & Kansau I (2013) Surface-layer (S-layer) of human and animal Clostridium difficile strains and their behaviour in adherence to epithelial cells and intestinal colonization. *J Med Microbiol* **62**, 1386–1393.

123 Fagan RP, Janoir C, Collignon A, Mastrantonio P, Poxton IR & Fairweather NF (2011) A proposed nomenclature for cell wall proteins of Clostridium difficile. *J Med Microbiol* **60**, 1225–1228.
124 Fagan RP & Fairweather NF (2014) Biogenesis and functions of bacterial S-layers. *Nature Reviews Microbiology 2014 12:3* **12**, 211–222.
125 Bradshaw WJ, Kirby JM, Roberts AK, Shone CC & Acharya KR (2017) Cwp2 from Clostridium difficile exhibits an extended three domain fold and cell adhesion in vitro. *FEBS J* **284**, 2886–2898.
126 Zhou Q, Rao F, Chen Z, Cheng Y, Zhang Q, Zhang J, Guan Z, He Y, Yu W, Cui G, Qi X & Hong W (2022) The cwp66 Gene Affects Cell Adhesion, Stress Tolerance, and Antibiotic Resistance in Clostridioides difficile . *Microbiol Spectr* **10**.
127 McKee RW, Aleksanyan N, Garrett EM & Tamayo R (2018) Type IV pili promote Clostridium difficile adherence and persistence in a mouse model of infection. *Infect Immun* **86**.
128 Waligora AJ, Hennequin C, Mullany P, Bourlioux P, Collignon A & Karjalainen T (2001) Characterization of a cell surface protein of Clostridium difficile with adhesive properties. *Infect Immun* **69**, 2144–2153.
129 Tulli L, Marchi S, Petracca R, Shaw HA, Fairweather NF, Scarselli M, Soriani M & Leuzzi R (2013) CbpA: a novel surface exposed adhesin of Clostridium difficile targeting human collagen. *Cell Microbiol* **15**, 1674–1687.
130 Kovacs-Simon A, Leuzzi R, Kasendra M, Minton N, Titball RW & Michell SL (2014) Lipoprotein CD0873 Is a Novel Adhesin of Clostridium difficile. *J Infect Dis* **210**, 274–284.
131 Bradshaw WJ, Bruxelle JF, Kovacs-Simon A, Harmer NJ, Janoir C, Pechine S, Acharya KR & Michell SL (2019) Molecular features of lipoprotein CD0873: A potential vaccine against the human pathogen Clostridioides difficile. *J Biol Chem* **294**, 15850.
132 Haiko J & Westerlund-Wikström B (2013) The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology 2013, Vol 2, Pages 1242-1267* **2**, 1242–1267.
133 Chaban B, Hughes HV & Beeby M (2015) The flagellum in bacterial pathogens: For motility and a whole lot more. *Semin Cell Dev Biol* **46**, 91–103.
134 Song WS, Cho SY, Hong HJ, Park SC & Yoon S il (2017) Self-Oligomerizing Structure of the Flagellar Cap Protein FliD and Its Implication in Filament Assembly. *J Mol Biol* **429**, 847–857.
135 Goon S, Kelly JF, Logan SM, Ewing CP & Guerry P (2003) Pseudaminic acid, the major modification on Campylobacter flagellin, is synthesized via the Cj1293 gene. *Mol Microbiol* **50**, 659–671.
136 Schirm M, Soo EC, Aubry AJ, Austin J, Thibault P & Logan SM (2003) Structural, genetic and functional characterization of the flagellin glycosylation process in Helicobacter pylori. *Mol Microbiol* **48**, 1579–1592.
137 Salah Ud-Din AIM & Roujeinikova A (2017) Flagellin glycosylation with pseudaminic acid in Campylobacter and Helicobacter: prospects for development of novel therapeutics. *Cellular and Molecular Life Sciences 2017 75:7* **75**, 1163–1178.
138 Ardissone S, Kint N & Viollier PH (2020) Specificity in glycosylation of multiple flagellins by the modular and cell cycle regulated glycosyltransferase flmg. *Elife* **9**, 1–28.
139 Schirm M, Arora SK, Verma A, Vinogradov E, Thibault P, Ramphal R & Logan SM (2004) Structural and Genetic Characterization of Glycosylation of Type a Flagellin in Pseudomonas aeruginosa. *J Bacteriol* **186**, 2523–2531.

140 Verma A, Schirm M, Arora SK, Thibault P, Logan SM & Ramphal R (2006) Glycosylation of b-type flagellin of Pseudomonas aeruginosa: Structural and genetic basis. *J Bacteriol* **188**, 4395–4403.
141 Twine SM, Reid CW, Aubry A, McMullin DR, Fulton KM, Austin J & Logan SM (2009) Motility and flagellar glycosylation in Clostridium difficile. *J Bacteriol* **191**, 7050–7062.
142 Faulds-Pain A, Twine SM, Vinogradov E, Strong PCR, Dell A, Buckley AM, Douce GR, Valiente E, Logan SM & Wren BW (2014) The post-translational modification of the Clostridium difficile flagellin affects motility, cell surface properties and virulence. *Mol Microbiol* **94**, 272–289.
143 Bouché L, Panico M, Hitchen P, Binet D, Sastre F, Faulds-Pain A, Valiente E, Vinogradov E, Aubry A, Fulton K, Twine S, Logan SM, Wren BW, Dell A & Morris HR (2016) The Type B Flagellin of Hypervirulent Clostridium difficile Is Modified with Novel Sulfonated Peptidylamido-glycans. *Journal of Biological Chemistry* **291**, 25439–25449.
144 Valiente E, Bouché L, Hitchen P, Faulds-Pain A, Songane M, Dawson LF, Donahue E, Stabler RA, Panico M, Morris HR, Bajaj-Elliott M, Logan SM, Dell A & Wren BW (2016) Role of Glycosyltransferases Modifying Type B Flagellin of Emerging Hypervirulent Clostridium difficile Lineages and Their Impact on Motility and Biofilm Formation. *Journal of Biological Chemistry* **291**, 25450–25461.
145 Zong Y, Xu Y, Liang X, Keene DR, Höök A, Gurusiddappa S, Höök M & Narayana SVL (2005) A 'Collagen Hug" Model for Staphylococcus aureus CNA binding to collagen.' *EMBO J* **24**, 4224.
146 Hensbergen PJ, Klychnikov OI, Bakker D, Van Winden VJC, Ras N, Kemp AC, Cordfunke RA, Dragan I, Deelder AM, Kuijper EJ, Corver J, Drijfhout JW & Van Leeuwen HC (2014) A novel secreted metalloprotease (CD2830) from clostridium difficile cleaves specific proline sequences in LPXTG cell Surface Proteins. *Molecular and Cellular Proteomics* **13**, 1231–1244.
147 Hensbergen PJ, Klychnikov OI, Bakker D, Dragan I, Kelly ML, Minton NP, Corver J, Kuijper EJ, Drijfhout JW & Van Leeuwen HC (2015) Clostridium difficile secreted Pro-Pro endopeptidase PPEP-1 (ZMP1/CD2830) modulates adhesion through cleavage of the collagen binding protein CD2831. *FEBS Lett* **589**, 3952–3958.
148 van Leeuwen HC, Roelofs D, Corver J & Hensbergen P (2021) Phylogenetic analysis of the bacterial Pro-Pro-endopeptidase domain reveals a diverse family including secreted and membrane anchored proteins. *Curr Res Microb Sci* **2**, 100024.
149 Walden M, Edwards JM, Dziewulska AM, Bergmann R, Saalbach G, Rohde M, Schwarz-Linek U & Banfield MJ (2015) Covalent host-targeting by thioester domains of Gram-positive pathogens. *Acta Crystallogr A Found Adv* **71**, s29–s29.
150 Cafardi V, Biagini M, Martinelli M, Leuzzi R, Rubino JT, Cantini F, Norais N, Scarselli M, Serruto D & Unnikrishnan M (2013) Identification of a Novel Zinc Metalloprotease through a Global Analysis of Clostridium difficile Extracellular Proteins. *PLoS One* **8**, 81306.
151 Klimpel KR, Arora N & Leppla SH (1994) Anthrax toxin lethal factor contains a zinc metalloprotease consensus sequence which is required for lethal toxin activity. *Mol Microbiol* **13**, 1093–1100.
152 Yaron A, Naider F & Scharpe S (1993) Proline-dependent structural and biological properties of peptides and proteins. *Crit Rev Biochem Mol Biol* **28**, 31–81.
153 Vanhoof G, Goossens F, Meester I De, Hendriks D & Scharpé S (1995) Proline motifs in peptides and their biological processing. *The FASEB Journal* **9**, 736–744.
154 Cunningham DF & O'Connor B (1997) Proline specific peptidases. *Biochim Biophys Acta* **1343**, 160–186.

155 Gass J & Khosla C (2007) Prolyl endopeptidases. *Cellular and Molecular Life Sciences* **64**, 345–355.
156 Rodriguez J, Gupta N, Smith RD & Pevzner PA (2008) Does trypsin cut before proline? *J Proteome Res* **7**, 300–305.
157 Klychnikov OI, Shamorkina TM, Weeks SD, van Leeuwen HC, Corver J, Drijfhout JW, van Veelen PA, Sluchanko NN, Strelkov S V. & Hensbergen PJ (2018) Discovery of a new Pro-Pro endopeptidase, PPEP-2, provides mechanistic insights into the differences in substrate specificity within the PPEP family. *Journal of Biological Chemistry* **293**, 11154–11165.
158 Forsgren E (2010) European foulbrood in honey bees. *J Invertebr Pathol* **103**, S5–S9.
159 Mesnage S, Fontaine T, Mignot T, Delepierre M, Mock M & Fouet A (2000) Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation. *EMBO J* **19**, 4473.
160 Schacherl M, Pichlo C, Neundorf I & Baumann U (2015) Structural Basis of Proline-Proline Peptide Bond Specificity of the Metalloprotease Zmp1 Implicated in Motility of Clostridium difficile. *Structure* **23**, 1632–1642.
161 Rubino JT, Martinelli M, Cantini F, Castagnetti A, Leuzzi R, Banci L & Scarselli M (2016) Structural characterization of zinc-bound Zmp1, a zinc-dependent metalloprotease secreted by Clostridium difficile. *Journal of Biological Inorganic Chemistry* **21**, 185–196.
162 Pichlo C, Juetten L, Wojtalla F, Schacherl M, Diaz D & Baumann U (2019) Molecular determinants of the mechanism and substrate specificity of Clostridium difficile proline-proline endopeptidase-1. *Journal of Biological Chemistry* **294**, 11525–11535.
163 Gonzalez-Delgado LS, Walters-Morgan H, Salamaga B, Robertson AJ, Hounslow AM, Jagielska E, Sabała I, Williamson MP, Lovering AL & Mesnage S (2019) Two-site recognition of Staphylococcus aureus peptidoglycan by lysostaphin SH3b. *Nature Chemical Biology 2019 16:1* **16**, 24–30.
164 Valk V, Kaaij RM van der & Dijkhuizen L (2017) The evolutionary origin and possible functional roles of FNIII domains in two Microbacterium aurum B8.A granular starch degrading enzymes, and in other carbohydrate acting enzymes. *Amylase* **1**, 1–11.
165 Watanabe T, Ito Y, Yamada T, Hashimoto M, Sekine S & Tanaka H (1994) The roles of the C-terminal domain and type III domains of chitinase A1 from Bacillus circulans WL-12 in chitin degradation. *J Bacteriol* **176**, 4465–4472.
166 Wiita AP, Seaman JE & Wells JA (2014) Global Analysis of Cellular Proteolysis by Selective Enzymatic Labeling of Protein N-Termini. *Methods Enzymol* **544**, 327–358.
167 Griswold AR, Cifani P, Rao SD, Axelrod AJ, Miele MM, Hendrickson RC, Kentsis A & Bachovchin DA (2019) A Chemical Strategy for Protease Substrate Profiling. *Cell Chem Biol* **26**, 901-907.e6.
168 Ju S, Kwon Y, Kim JM, Park D, Lee S, Lee JW, Hwang CS & Lee C (2020) iNrich, Rapid and Robust Method to Enrich N-Terminal Proteome in a Highly Multiplexed Platform. *Anal Chem* **92**, 6462–6469.
169 Matthews DJ & Wells JA (1993) Substrate Phage: Selection of Protease Substrates by Monovalent Phage Display. *Science (1979)* **260**, 1113–1117.
170 Román-Meléndez GD, Venkataraman T, Monaco DR & Larman HB (2020) Protease Activity Profiling via Programmable Phage Display of Comprehensive Proteome-Scale Peptide Libraries. *Cell Syst* **11**, 375-381.e4.
171 Choe Y, Leonetti F, Greenbaum DC, Lecaille F, Bogyo M, Brömme D, Ellman JA & Craik CS (2006) Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *Journal of Biological Chemistry* **281**, 12824–12832.

172 Harris JL, Backes BJ, Leonetti F, Mahrus S, Ellman JA & Craik CS (2000) Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proceedings of the National Academy of Sciences* **97**, 7754–7759.
173 Thormann KM & Paulick A (2010) Tuning the flagellar motor. *Microbiology (N Y)* **156**, 1275–1283.
174 Josenhans C & Suerbaum S (2002) The role of motility as a virulence factor in bacteria. *International Journal of Medical Microbiology* **291**, 605–614.
175 Nakamura S & Minamino T (2019) Flagella-Driven Motility of Bacteria. *Biomolecules 2019, Vol 9, Page 279* **9**, 279.
176 Wadhwa N & Berg HC (2022) Bacterial motility: machinery and mechanisms. *Nat Rev Microbiol* **20**, 161–173.
177 Chidwick HS & Fascione MA (2020) Mechanistic and structural studies into the biosynthesis of the bacterial sugar pseudaminic acid (Pse5Ac7Ac). *Org Biomol Chem* **18**, 799–809.
178 Valiente E, Bouché L, Hitchen P, Faulds-Pain A, Songane M, Dawson LF, Donahue E, Stabler RA, Panico M, Morris HR, Bajaj-Elliott M, Logan SM, Dell A & Wren BW (2016) Role of glycosyltransferases modifying type B flagellin of emerging hypervirulent Clostridium difficile lineages and their impact on motility and biofilm formation. *Journal of Biological Chemistry* **291**, 25450–25461.
179 Bouché L, Panico M, Hitchen P, Binet D, Sastre F, Faulds-Pain A, Valiente E, Vinogradov E, Aubry A, Fulton K, Twine S, Logan SM, Wren BW, Dell A & Morris HR (2016) The type B flagellin of hypervirulent Clostridium difficile is modified with novel sulfonated peptidylamido-glycans. *Journal of Biological Chemistry* **291**, 25439–25449.
180 Heap JT, Pennington OJ, Cartman ST & Minton NP (2009) A modular system for Clostridium shuttle plasmids. *J Microbiol Methods* **78**, 79–85.
181 Cheung TK, Lee CY, Bayer FP, McCoy A, Kuster B & Rose CM (2021) Defining the carrier proteome limit for single-cell proteomics. *Nat Methods* **18**, 76–83.
182 Heap JT, Kuehne SA, Ehsaan M, Cartman ST, Cooksley CM, Scott JC & Minton NP (2010) The ClosTron: Mutagenesis in Clostridium refined and streamlined. *J Microbiol Methods* **80**, 49–55.
183 Halim A, Westerlind U, Pett C, Schorlemer M, Rüetschi U, Brinkmalm G, Sihlbom C, Lengqvist J, Larson G & Nilsson J (2014) Assignment of saccharide identities through analysis of oxonium ion fragmentation profiles in LC-MS/MS of glycopeptides. *J Proteome Res* **13**, 6024–6032.
184 Pirro M, Mohammed Y, de Ru AH, Janssen GMC, Tjokrodirijo RTN, Madunić K, Wuhrer M, van Veelen PA & Hensbergen PJ (2021) Oxonium ion guided analysis of quantitative proteomics data reveals site-specific o-glycosylation of anterior gradient protein 2 (Agr2). *Int J Mol Sci* **22**, 5369.
185 KENNEDY EP & WEISS SB (1956) THE FUNCTION OF CYTIDINE COENZYMES IN THE BIOSYNTHESIS OF PHOSPHOLIPIDES. *Journal of Biological Chemistry* **222**, 193–214.
186 Singh SK, Yang K, Karthikeyan S, Huynh T, Zhang X, Phillips MA & Zhang H (2004) The thrH Gene Product of Pseudomonas aeruginosa Is a Dual Activity Enzyme with a Novel Phosphoserine:Homoserine Phosphotransferase Activity. *Journal of Biological Chemistry* **279**, 13166–13173.
187 Tasteyre A, Barc MC, Collignon A, Boureau H & Karjalainen T (2001) Role of FliC and FliD flagellar proteins of Clostridium difficile in adherence and gut colonization. *Infect Immun* **69**, 7937–7940.

188 Dingle TC, Mulvey GL & Armstrong GD (2011) Mutagenic Analysis of the Clostridium difficile Flagellar Proteins, FliC and FliD, and Their Contribution to Virulence in Hamsters. *Infect Immun* **79**, 4061.

189 Baban ST, Kuehne SA, Barketi-Klai A, Cartman ST, Kelly ML, Hardie KR, Kansau I, Collignon A & Minton NP (2013) The Role of Flagella in Clostridium difficile Pathogenesis: Comparison between a Non-Epidemic and an Epidemic Strain. *PLoS One* **8**, e73026.

190 Batah J, Kobeissy H, Pham PTB, Denève-Larrazet C, Kuehne S, Collignon A, Janoir-Jouveshomme C, Marvaud JC & Kansau I (2017) Clostridium difficile flagella induce a pro-inflammatory response in intestinal epithelium of mice in cooperation with toxins. *Scientific Reports 2017 7:1* **7**, 1–10.

191 Taylor ZW & Raushel FM (2019) Manganese-Induced Substrate Promiscuity in the Reaction Catalyzed by Phosphoglutamine Cytidylyltransferase from Campylobacter jejuni. *Biochemistry*.

192 Hensbergen PJ, De Ru AH, Friggen AH, Corver J, Smits WK & Van Veelen PA (2022) New insights into the type A glycan modification of Clostridioides difficile flagellar protein flagellin C by phosphoproteomics analysis. *J Biol Chem* **298**.

193 Smits WK, Mohammed Y, de Ru AH, Cordo' V, Friggen AH, van Veelen PA & Hensbergen PJ (2022) Clostridioides difficile Phosphoproteomics Shows an Expansion of Phosphorylated Proteins in Stationary Growth Phase. *mSphere* **7**.

194 Pirro M, Mohammed Y, van Vliet SJ, Rombouts Y, Sciacca A, de Ru AH, Janssen GMC, Tjokrodirijo RTN, Wuhrer M, van Veelen PA & Hensbergen PJ (2020) N-Glycoproteins Have a Major Role in MGL Binding to Colorectal Cancer Cell Lines: Associations with Overall Proteome Diversity. *International Journal of Molecular Sciences 2020, Vol 21, Page 5522* **21**, 5522.

195 McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R, Haas W & Gygi SP (2014) MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**, 7150–7158.

196 Käll L, Canterbury JD, Weston J, Noble WS & MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923–925.

197 Ducarmon QR, van der Bruggen T, Harmanus C, Sanders IMJG, Daenen LGM, Fluit AC, Vossen RHAM, Kloet SL, Kuijper EJ & Smits WK (2023) Clostridioides difficile infection with isolates of cryptic clade C-II: a genomic analysis of polymerase chain reaction ribotype 151. *Clinical Microbiology and Infection* **29**, 538.e1-538.e6.

198 Kelley LA, Mezulis S, Yates CM, Wass MN & Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols 2015 10:6* **10**, 845–858.

199 Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S & Steinegger M (2022) ColabFold: making protein folding accessible to all. *Nature Methods 2022 19:6* **19**, 679–682.

200 Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, Kundu DJ, García-Seisdedos D, Jarnuczak AF, Hewapathirana S, Pullman BS, Wertz J, Sun Z, Kawano S, Okuda S, Watanabe Y, Hermjakob H, Maclean B, Maccoss MJ, Zhu Y, Ishihama Y & Vizcaíno JA (2020) The ProteomeXchange consortium in 2020: enabling "big data" approaches in proteomics. *Nucleic Acids Res* **48**, D1145–D1152.

201 Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A & Vizcaíno JA (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* **50**, D543.

202 Dubberke E (2012) Clostridium difficile infection: The scope of the problem. *J Hosp Med* **7**, S1–S4.
203 Miller BA, Chen LF, Sexton DJ & Anderson DJ (2011) Comparison of the burdens of hospital-onset, healthcare facility-associated Clostridium difficile Infection and of healthcare-associated infection due to methicillin-resistant Staphylococcus aureus in community hospitals. *Infect Control Hosp Epidemiol* **32**, 387–390.
204 Kordus SL, Thomas AK & Lacy DB (2021) Clostridioides difficile toxins: mechanisms of action and antitoxin therapeutics. *Nature Reviews Microbiology 2021 20:5* **20**, 285–298.
205 Jongeneel CV, Bouvier J & Bairoch A (1989) A unique signature identifies a family of zinc-dependent metallopeptidases. *FEBS Lett* **242**, 211–214.
206 Claushuis B, Cordfunke RA, De Ru AH, Van Angeren J, Baumann U, Van Veelen PA, Wuhrer M, Corver J, Drijfhout JW, Hensbergen PJ & Hensbergen PJ (2024) Non-prime- and prime-side profiling of Pro-Pro endopeptidase specificity using synthetic combinatorial peptide libraries and mass spectrometry. *FEBS J*.
207 Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A & Lopez R (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276–W279.
208 Hatakeyama T, Kohzaki H & Yamasaki N (1992) A microassay for proteases using succinylcasein as a substrate. *Anal Biochem* **204**, 181–184.
209 Heap JT, Pennington OJ, Cartman ST, Carter GP & Minton NP (2007) The ClosTron: A universal gene knock-out system for the genus Clostridium. *J Microbiol Methods* **70**, 452–464.
210 Cartman ST & Minton NP (2010) A mariner-Based transposon system for in vivo random mutagenesis of clostridium difficile. *Appl Environ Microbiol* **76**, 1103–1109.
211 Fuchs M, Lamm-Schmidt V, Sulzer J, Ponath F, Jenniches L, Kirk JA, Fagan RP, Barquist L, Vogel J & Faber F (2021) An RNA-centric global view of Clostridioides difficile reveals broad activity of Hfq in a clinically important gram-positive bacterium. *Proc Natl Acad Sci U S A* **118**, 1–11.
212 Oliveira Paiva AM, Friggen AH, Hossein-Javaheri S & Smits WK (2016) The Signal Sequence of the Abundant Extracellular Metalloprotease PPEP-1 Can Be Used to Secrete Synthetic Reporter Proteins in Clostridium difficile. *ACS Synth Biol*.
213 Jain S, Graham C, Graham RLJ, McMullan G & Ternan NG (2011) Quantitative proteomic analysis of the heat stress response in Clostridium difficile strain 630. *J Proteome Res* **10**, 3880–3890.
214 Maaß S, Bartel J, Mücke PA, Schlüter R, Sura T, Zaschke-Kriesche J, Smits SHJ & Becher D (2021) Proteomic Adaptation of Clostridioides difficile to Treatment with the Antimicrobial Peptide Nisin. *Cells* **10**, 1–21.
215 Abhyankar WR, Zheng L, Brul S, De Koster CG & De Koning LJ (2019) Vegetative Cell and Spore Proteomes of Clostridioides difficile Show Finite Differences and Reveal Potential Protein Markers. *J Proteome Res* **18**, 3967–3976.
216 Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, Becher I, Pachl F, Kuster B & Bantscheff M (2013) Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J Proteome Res* **12**, 3586–3598.
217 Diederich B, Wilkinson JF, Magnin T, Najafi SMA, Errington J & Yudkin MD (1994) Role of interactions between SpoIIAA and SpoIIAB in regulating cell-specific transcription factor sigma F of Bacillus subtilis. *Genes Dev* **8**, 2653–2663.
218 Saujet L, Pereira FC, Henriques AO & Martin-Verstraete I (2014) The regulatory network controlling spore formation in Clostridium difficile. *FEMS Microbiol Lett* **358**, 1–10.

219 Aubry A, Hussack G, Chen W, KuoLee R, Twine SM, Fulton KM, Foote S, Carrillo CD, Tanha J & Logan SM (2012) Modulation of Toxin Production by the Flagellar Regulon in Clostridium difficile. *Infect Immun* **80**, 3521.

220 Lee KK, Jang CS, Yoon JY, Kim SY, Kim TH, Ryu KH & Kim W (2008) Abnormal cell division caused by inclusion bodies in E. coli; increased resistance against external stress. *Microbiol Res* **163**, 394–402.

221 Reynolds SL & Fischer K (2015) Pseudoproteases: mechanisms and function. *Biochemical Journal* **468**, 17–24.

222 Adrain C & Freeman M (2012) New lives for old: evolution of pseudoenzyme function illustrated by iRhoms. *Nature Reviews Molecular Cell Biology 2012 13:8* **13**, 489–498.

223 Ting L, Rad R, Gygi SP & Haas W (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods 2011 8:11* **8**, 937–940.

224 Bakker D, Buckley AM, de Jong A, van Winden VJC, Verhoeks JPA, Kuipers OP, Douce GR, Kuijper EJ, Smits WK & Corver J (2014) The HtrA-like protease CD3284 modulates virulence of Clostridium difficile. *Infect Immun*.

225 Pettit LJ, Browne HP, Yu L, Smits WK, Fagan RP, Barquist L, Martin MJ, Goulding D, Duncan SH, Flint HJ, Dougan G, Choudhary JS & Lawley TD (2014) Functional genomics reveals that Clostridium difficile Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**, 160.

226 Willing SE, Richards EJ, Sempere L, Dale AG, Cutting SM & Fairweather NF (2015) Increased toxin expression in a Clostridium difficile mfd mutant Microbe-host interactions and microbial pathogenicity. *BMC Microbiol* **15**, 1–10.

227 Kevorkian Y & Shen A (2017) Revisiting the role of Csp family proteins in regulating Clostridium difficile spore germination. *J Bacteriol* **199**.

228 Francis MB, Allen CA, Shrestha R & Sorg JA (2013) Bile Acid Recognition by the Clostridium difficile Germinant Receptor, CspC, Is Important for Establishing Infection. *PLoS Pathog* **9**, e1003356.

229 Permpoonpattana P, Tolls EH, Nadem R, Tan S, Brisson A & Cutting SM (2011) Surface layers of Clostridium difficile endospores. *J Bacteriol* **193**, 6461–6470.

230 Claushuis B, Cordfunke RA, de Ru AH, Otte A, van Leeuwen HC, Klychnikov OI, van Veelen PA, Corver J, Drijfhout JW & Hensbergen PJ (2023) In-Depth Specificity Profiling of Endopeptidases Using Dedicated Mix-and-Split Synthetic Peptide Libraries and Mass Spectrometry. *Anal Chem* **95**, 11621–11631.

231 Heap JT, Cartman ST, Kuehne SA, Cooksley C & Minton NP (2010) ClosTron-targeted mutagenesis. *Methods in Molecular Biology* **646**, 165–182.

232 Fagan RP & Fairweather NF (2011) Clostridium difficile has two parallel and essential sec secretion systems. *Journal of Biological Chemistry*.

233 Claushuis B, de Ru AH, Rotman SA, van Veelen PA, Dawson LF, Wren BW, Corver J, Smits WK & Hensbergen PJ (2023) Revised Model for the Type A Glycan Biosynthetic Pathway in Clostridioides difficile Strain 630Δerm Based on Quantitative Proteomics of cd0241-cd0244 Mutant Strains. *ACS Infect Dis* **9**, 2665–2674.

234 Putnam EE, Nock AM, Lawley TD & Shen A (2013) SpoIVA and SipL Are Clostridium difficile Spore Morphogenetic Proteins. *J Bacteriol* **195**, 1214.

235 Oliveira Paiva AM, Friggen AH, Douwes R, Wittekoek B & Smits WK (2022) Practical observations on the use of fluorescent reporter systems in Clostridioides difficile. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* **115**, 297–323.

236 Anand K, Ziebuhr J, Wadhwani P, Mesters JR & Hilgenfeld R (2003) Coronavirus main proteinase (3CLpro) Structure: Basis for design of anti-SARS drugs. *Science (1979)* **300**, 1763–1767.

237 Coughlin SR (2000) Thrombin signalling and protease-activated receptors. *Nature 2000 407:6801* **407**, 258–264.

238 Barry M & Bleackley RC (2002) Cytotoxic T lymphocytes: all roads lead to death. *Nature Reviews Immunology 2002 2:6* **2**, 401–409.

239 Pop C & Salvesen GS (2009) Human caspases: Activation, specificity, and regulation. *Journal of Biological Chemistry* **284**, 21777–21781.

240 Poreba M, Salvesen GS & Drag M (2017) Synthesis of a HyCoSuL peptide substrate library to dissect protease substrate specificity. *Nature Protocols 2017 12:10* **12**, 2189–2214.

241 Vizovišek M, Vidmar R, Drag M, Fonović M, Salvesen GS & Turk B (2018) Protease Specificity: Towards In Vivo Imaging Applications and Biomarker Discovery. *Trends Biochem Sci* **43**, 829–844.

242 Galipeau HJ, Caminero A, Turpin W, Bermudez-Brito M, Santiago A, Libertucci J, Constante M, Raygoza Garay JA, Rueda G, Armstrong S, Clarizio A, Smith MI, Surette MG, Bercik P, Beck P, Bernstein C, Croitoru K, Dieleman L, Feagan B, Griffiths A, Guttman D, Jacobson K, Kaplan G, Krause DO, Madsen K, Marshall J, Moayyedi P, Ropeleski M, Seidman E, Silverberg M, Snapper S, Stadnyk A, Steinhart H, Surette M, Turner D, Walters T, Vallance B, Aumais G, Bitton A, Cino M, Critch J, Denson L, Deslandres C, El-Matary W, Herfarth H, Higgins P, Huynh H, Hyams J, Mack D, McGrath J, Otley A, Panancionne R & Verdu EF (2021) Novel Fecal Biomarkers That Precede Clinical Diagnosis of Ulcerative Colitis. *Gastroenterology* **160**, 1532–1545.

243 Corver J, Cordo' V, van Leeuwen HC, Klychnikov OI & Hensbergen PJ (2017) Covalent attachment and Pro-Pro endopeptidase (PPEP-1)-mediated release of Clostridium difficile cell surface proteins involved in adhesion. *Mol Microbiol* **105**, 663–673.

244 Diamond SL (2007) Methods for mapping protease specificity. *Curr Opin Chem Biol* **11**, 46–51.

245 Uzozie AC, Smith TG, Chen S & Lange PF (2022) Sensitive Identification of Known and Unknown Protease Activities by Unsupervised Linear Motif Deconvolution. *Anal Chem* **94**, 2244–2254.

246 Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN & Overall CM (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature Biotechnology 2010 28:3* **28**, 281–288.

247 Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A & Vandekerckhove J (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins. *Mol Cell Proteomics* **1**, 896–903.

248 Shin S, Hong JH, Na Y, Lee M, Qian WJ, Kim VN & Kim JS (2020) Development of Multiplexed Immuno-N-Terminomics to Reveal the Landscape of Proteolytic Processing in Early Embryogenesis of Drosophila melanogaster. *Anal Chem* **92**, 4926–4934.

249 Wood SE, Sinsinbar G, Gudlur S, Nallani M, Huang CF, Liedberg B & Mrksich M (2017) A Bottom-Up Proteomic Approach to Identify Substrate Specificity of Outer-Membrane Protease OmpT. *Angewandte Chemie - International Edition* **56**, 16531–16535.

250 Dai R, Ten AS & Mrksich M (2019) Profiling Protease Activity in Laundry Detergents with Peptide Arrays and SAMDI Mass Spectrometry. *Ind Eng Chem Res*.

251 Schilling O & Overall CM (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature Biotechnology 2008 26:6* **26**, 685–694.
252 Tanco S, Lorenzo J, Garcia-Pardo J, Degroeve S, Martens L, Aviles FX, Gevaert K & Van Damme P (2013) Proteome-derived peptide libraries to study the substrate specificity profiles of carboxypeptidases. *Mol Cell Proteomics* **12**, 2096–2110.
253 Nguyen MTN, Shema G, Zahedi RP & Verhelst SHL (2018) Protease Specificity Profiling in a Pipet Tip Using "charge-Synchronized" Proteome-Derived Peptide Libraries. *J Proteome Res* **17**, 1923–1933.
254 O'Donoghue AJ, Alegra Eroy-Reveles AA, Knudsen GM, Ingram J, Zhou M, Statnekov JB, Greninger AL, Hostetter DR, Qu G, Maltby DA, Anderson MO, Derisi JL, McKerrow JH, Burlingame AL & Craik CS (2012) Global Identification of Peptidase Specificity by Multiplex Substrate Profiling. *Nat Methods* **9**, 1095.
255 Lapek JD, Jiang Z, Wozniak JM, Arutyunova E, Wang SC, Joanne Lemieux M, Gonzalez DJ & O'Donoghue AJ (2019) Quantitative multiplex substrate profiling of peptidases by mass spectrometry. *Molecular and Cellular Proteomics* **18**, 968–981.
256 Lam KS, Salmon SE, Hersh EM, Hruby VJ, Kazmierskit WM & Knappt RJ (1991) A new type of synthetic peptide library for identifying ligand-binding activity. *Nature 1991 354:6348* **354**, 82–84.
257 Birkett AJ, Soler DF, Wolz RL, Bond JS, Wiseman J, Berman J & Harris RB (1991) Determination of enzyme specificity in a complex mixture of peptide substrates by N-terminal sequence analysis. *Anal Biochem* **196**, 137–143.
258 Petithory JR, Masiarzt# FR, Kirsch JF, Santi D V, Ii BAM, Chin M, Stratton-Thomas JR, Thudium KB, Ralston R, Rosenberg S & Jewell DA (1991) A rapid method for determination of endoproteinase substrate specificity: specificity of the 3C proteinase from hepatitis A virus. *Proceedings of the National Academy of Sciences* **88**, 11510–11514.
259 Turk BE, Huang LL, Piro ET & Cantley LC (2001) Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nature Biotechnology 2001 19:7* **19**, 661–667.
260 Ng NM, Pike RN & Boyd SE (2009) Subsite cooperativity in protease specificity. *Biol Chem* **390**, 401–407.
261 Cox J (2022) Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology 2022*, 1–11.
262 Mohammed Y & Palmblad M (2018) Visualization and application of amino acid retention coefficients obtained from modeling of peptide retention. *J Sep Sci* **41**, 3644–3653.
263 Winter D, Pipkorn R & Lehmann WD (2009) Separation of peptide isomers and conformers by ultra performance liquid chromatography. *J Sep Sci* **32**, 1111–1119.
264 Eckhard U, Huesgen PF, Schilling O, Bellac CL, Butler GS, Cox JH, Dufour A, Goebeler V, Kappelhoff R, auf dem Keller U, Klein T, Lange PF, Marino G, Morrison CJ, Prudova A, Rodriguez D, Starr AE, Wang Y & Overall CM (2016) Active site specificity profiling of the matrix metalloproteinase family: Proteomic identification of 4300 cleavage sites by nine MMPs explored with structural and synthetic peptide cleavage analyses. *Matrix Biology* **49**, 37–60.
265 Zelanis A, Oliveira AK, Prudova A, Huesgen PF, Tashima AK, Kizhakkedathu J, Overall CM & Serrano SMT (2019) Deep Profiling of the Cleavage Specificity and Human Substrates of Snake Venom Metalloprotease HF3 by Proteomic Identification of Cleavage Site Specificity (PICS) Using Proteome Derived Peptide Libraries and Terminal Amine Isotopic Labeling of Substrates (TAILS) N-Terminomics. *J Proteome Res* **18**, 3419–3428.

266 Liigand P, Kaupmees K & Kruve A (2019) Influence of the amino acid composition on the ionization efficiencies of small peptides. *Journal of Mass Spectrometry* **54**, 481–487.
267 Liao SM, Du QS, Meng JZ, Pang ZW & Huang RB (2013) The multiple roles of histidine in protein interactions. *Chem Cent J* **7**, 1–12.
268 Hallgren J, Tsirigos KD, Damgaard Pedersen M, Juan J, Armenteros A, Marcatili P, Nielsen H, Krogh A & Winther O (2022) DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*, 2022.04.08.487609.
269 Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G & Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology 2022 40:7* **40**, 1023–1025.
270 Hiemstra HS, Duinkerken G, Benckhuijsen WE, Amons R, de Vries RRP, Roep BO & Drijfhout JW (1997) The identification of CD4+ T cell epitopes with dedicated synthetic peptide libraries. *Proc Natl Acad Sci U S A* **94**, 10313.
271 Madunić K, Luijkx YMCA, Mayboroda OA, Janssen GMC, van Veelen PA, Strijbis K, Wennekes T, Lageveen-Kammeijer GSM & Wuhrer M (2023) O-Glycomic and Proteomic Signatures of Spontaneous and Butyrate-Stimulated Colorectal Cancer Cell Line Differentiation. *Molecular & Cellular Proteomics* **22**, 100501.
272 Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A & Lopez R (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, gkac240–gkac240.
273 Crooks GE, Hon G, Chandonia JM & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190.
274 Dalbey RE, Wang P & van Dijl JM (2012) Membrane Proteases in the Bacterial Protein Secretion and Quality Control Pathway. *Microbiology and Molecular Biology Reviews* **76**, 311–330.
275 Michel A, Agerer F, Hauck CR, Herrmann M, Ullrich J, Hacker J & Ohlsen K (2006) Global regulatory impact of ClpP protease of Staphylococcus aureus on regulons involved in virulence, oxidative stress response, autolysis, and DNA repair. *J Bacteriol* **188**, 5783–5796.
276 Benitez JA & Silva AJ (2016) Vibrio cholerae hemagglutinin(HA)/protease: An extracellular metalloprotease with multiple pathogenic activities. *Toxicon* **115**, 55–62.
277 Caminero A, Guzman M, Libertucci J & Lomax AE (2023) The emerging roles of bacterial proteases in intestinal diseases. *Gut Microbes* **15**.
278 Culp E & Wright GD (2016) Bacterial proteases, untapped antimicrobial drug targets. *The Journal of Antibiotics 2017 70:4* **70**, 366–377.
279 Razzaq A, Shamsi S, Ali A, Ali Q, Sajjad M, Malik A & Ashraf M (2019) Microbial proteases applications. *Front Bioeng Biotechnol* **7**, 451237.
280 Brömme D, Peters K, Fink S & Fittkau S (1986) Enzyme-substrate interactions in the hydrolysis of peptide substrates by thermitase, subtilisin BPN′, and proteinase K. *Arch Biochem Biophys* **244**, 439–446.
281 Kubota K, Tanokura M & Takahashi K (2005) Purification and characterization of a novel prolyl endopeptidase from Aspergillus niger. *Proceedings of the Japan Academy, Series B* **81**, 447–453.
282 Lee L, Zhang Y, Ozar B, Sensen CW & Schriemer DC (2016) Carnivorous Nutrition in Pitcher Plants (Nepenthes spp.) via an Unusual Complement of Endogenous Enzymes. *J Proteome Res* **15**, 3108–3117.

283 Lambeir AM, Durinx C, Scharpé S & De Meester I (2003) Dipeptidyl-peptidase IV from bench to bedside: An update on structural properties, functions, and clinical aspects of the enzyme DPP IV. *Crit Rev Clin Lab Sci* **40**, 209–294.
284 Potempa M, Lee SK, Kurt Yilmaz N, Nalivaika EA, Rogers A, Spielvogel E, Carter CW, Schiffer CA & Swanstrom R (2018) HIV-1 Protease Uses Bi-Specific S2/S2′ Subsites to Optimize Cleavage of Two Classes of Target Sites. *J Mol Biol* **430**, 5182–5195.
285 Bredemeyer AJ, Lewis RM, Malone JP, Davis AE, Gross J, Townsend RR & Ley TJ (2004) A proteomic approach for the discovery of protease substrates. *Proc Natl Acad Sci U S A* **101**, 11785–11790.
286 Backes BJ, Harris JL, Leonetti F, Craik CS & Ellman JA (2000) Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. *Nat Biotechnol* **18**.
287 Coorevits A, Dinsdale AE, Halket G, Lebbe L, de Vos P, van Landschoot A & Logan NA (2012) Taxonomic revision of the genus Geobacillus: Emendation of Geobacillus, G. stearothermophilus, G. jurassicus, G. toebii, G. thermodenitrificans and G. thermoglucosidans (nom. corrig., formerly 'thermoglucosidasius'); transfer of Bacillus thermantarcticus to the genus as G. thermantarcticus comb. nov.; proposal of Caldibacillus debilis gen. nov., comb. nov. *Int J Syst Evol Microbiol* **62**, 1470–1485.
288 Kataeva IA, Seidel RD, Shah A, West LT, Li XL & Ljungdahl LG (2002) The fibronectin type 3-like repeat from the Clostridium thermocellum cellobiohydrolase CbHa promotes hydrolysis of cellulose by modifying its surface. *Appl Environ Microbiol* **68**, 4292–4300.
289 Schwarzbauer JE & DeSimone DW (2011) Fibronectins, Their Fibrillogenesis, and In Vivo Functions. *Cold Spring Harb Perspect Biol* **3**, a005041.
290 Fagerquist CK & Zaragoza WJ (2018) Proteolytic Surface-Shaving and Serotype-Dependent Expression of SPI-1 Invasion Proteins in Salmonella enterica Subspecies enterica. *Front Nutr* **5**, 424179.
291 van Balen P, Kester MGD, de Klerk W, Crivello P, Arrieta-Bolaños E, de Ru AH, Jedema I, Mohammed Y, Heemskerk MHM, Fleischhauer K, van Veelen PA & Falkenburg JHF (2020) Immunopeptidome Analysis of HLA-DPB1 Allelic Variants Reveals New Functional Hierarchies. *The Journal of Immunology* **204**, 3273–3282.
292 Schilling B, Rardin MJ, MacLean BX, Zawadzka AM, Frewen BE, Cusack MP, Sorensen DJ, Bereman MS, Jing E, Wu CC, Verdin E, Kahn CR, MacCoss MJ & Gibson BW (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol Cell Proteomics* **11**, 202–214.
293 Crooks GE, Hon G, Chandonia JM & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190.
294 Claushuis B, de Ru AH, van Veelen PA, Hensbergen PJ & Corver J (2024) Characterization of the Clostridioides difficile 630Δerm putative Pro-Pro endopeptidase CD1597. .
295 Zondlo NJ (2013) Aromatic-proline interactions: Electronically tunable CH/π interactions. *Acc Chem Res* **46**, 1039–1049.
296 Brandl M, Weiss MS, Jabs A, Sühnel J & Hilgenfeld R (2001) C-H···π-interactions in proteins. *J Mol Biol* **307**, 357–377.
297 Perrin CL (2010) Are short, low-barrier hydrogen bonds unusually strong? *Acc Chem Res* **43**, 1550–1557.

298 Arnau J, Lauritzen C, Petersen GE & Pedersen J (2006) Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. *Protein Expr Purif* **48**, 1–13.
299 Rizzello CG, De Angelis M, Di Cagno R, Camarca A, Silano M, Losito I, De Vincenzi M, De Bari MD, Palmisano F, Maurano F, Gianfrani C & Gobbetti M (2007) Highly efficient gluten degradation by lactobacilli and fungal proteases during food processing: New perspectives for celiac disease. *Appl Environ Microbiol* **73**, 4499–4507.
300 Lopez M & Edens L (2005) Effective prevention of chill-haze in beer using an acid proline-specific endoprotease from Aspergillus niger. *J Agric Food Chem* **53**, 7944–7949.
301 Mohd Azmi SI, Kumar P, Sharma N, Sazili AQ, Lee SJ & Ismail-Fitry MR (2023) Application of Plant Proteases in Meat Tenderization: Recent Trends and Future Prospects. *Foods* **12**.
302 Zheng L, Baumann U & Reymond JL (2004) An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res* **32**.
303 Adams PD, Afonine P V., Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC & Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *urn:issn:0907-4449* **66**, 213–221.
304 Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501.
305 Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH & Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70.
306 Haiko J & Westerlund-Wikström B (2013) The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology 2013, Vol 2, Pages 1242-1267* **2**, 1242–1267.
307 Arora SK, Ritchings BW, Almira EC, Lory S & Ramphal R (1998) The Pseudomonas aeruginosa Flagellar Cap Protein, FliD, Is Responsible for Mucin Adhesion. *Infect Immun* **66**, 1000.
308 Lillehoj EP, Kim BT & Kim KC (2002) Identification of Pseudomonas aeruginosa flagellin as an adhesin for Muc1 mucin. *Am J Physiol Lung Cell Mol Physiol* **282**.
309 Guerry P (2007) Campylobacter flagella: not just for motility. *Trends Microbiol* **15**, 456–461.
310 Chaban B, Hughes HV & Beeby M (2015) The flagellum in bacterial pathogens: For motility and a whole lot more. *Semin Cell Dev Biol* **46**, 91–103.
311 Lemon KP, Higgins DE & Kolter R (2007) Flagellar motility is critical for Listeria monocytogenes biofilm formation. *J Bacteriol* **189**, 4418–4424.
312 Andersen-Nissen E, Smith KD, Strobe KL, Rassoulian Barrett SL, Cookson BT, Logan SM & Aderem A (2005) Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc Natl Acad Sci U S A* **102**, 9247–9252.
313 Taguchi F, Takeuchi K, Katoh E, Murata K, Suzuki T, Marutani M, Kawasaki T, Eguchi M, Katoh S, Kaku H, Yasuda C, Inagaki Y, Toyoda K, Shiraishi T & Ichinose Y (2006) Identification of glycosylation genes and glycosylated amino acids of flagellin in Pseudomonas syringae pv. tabaci. *Cell Microbiol* **8**, 923–938.
314 Taguchi F, Shibata S, Suzuki T, Ogawa Y, Aizawa SI, Takeuchi K & Ichinose Y (2008) Effects of Glycosylation on Swimming Ability and Flagellar Polymorphic Transformation in Pseudomonas syringae pv. tabaci 6605. *J Bacteriol* **190**, 764.
315 Taguchi F, Suzuki T, Takeuchi K, Inagaki Y, Toyoda K, Shiraishi T & Ichinose Y (2009) Glycosylation of flagellin from Pseudomonas syringae pv. tabaci 6605 contributes to evasion of host tobacco plant surveillance system. *Physiol Mol Plant Pathol* **74**, 11–17.

316 Kreutzberger MAB, Ewing C, Poly F, Wang F & Egelman EH (2020) Atomic structure of the Campylobacter jejuni flagellar filament reveals how e Proteobacteria escaped Toll-like receptor 5 surveillance. *Proc Natl Acad Sci U S A* **117**, 16985–16991.
317 Song WS, Jeon YJ, Namgung B, Hong M & Yoon S Il (2017) A conserved TLR5 binding and activation hot spot on flagellin. *Sci Rep* **7**.
318 Batah J, Denève-Larrazet C, Jolivot PA, Kuehne S, Collignon A, Marvaud JC & Kansau I (2016) Clostridium difficile flagella predominantly activate TLR5-linked NF-κB pathway in epithelial cells. *Anaerobe* **38**, 116–124.
319 Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM & Aderem A (2001) The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103.
320 Gong YN & Shao F (2012) Sensing bacterial infections by NAIP receptors in NLRC4 inflammasome activation. *Protein Cell* **3**, 98.
321 Halff EF, Diebolder CA, Versteeg M, Schouten A, Brondijk THC & Huizinga EG (2012) Formation and Structure of a NAIP5-NLRC4 Inflammasome Induced by Direct Interactions with Conserved N- and C-terminal Regions of Flagellin. *J Biol Chem* **287**, 38460.
322 Ghose C, Eugenis I, Sun X, Edwards AN, McBride SM, Pride DT, Kelly CP & Ho DD (2016) Immunogenicity and protective efficacy of recombinant Clostridium difficile flagellar protein FliC. *Emerg Microbes Infect* **5**, e8.
323 Bruxelle JF, Mizrahi A, Hoÿs S, Collignon A, Janoir C & Péchiné S (2017) Clostridium difficile flagellin FliC: Evaluation as adjuvant and use in a mucosal vaccine against Clostridium difficile. *PLoS One* **12**, e0187212.
324 Bruxelle JF, Tsapis N, Hoys S, Collignon A, Janoir C, Fattal E & Péchiné S (2018) Protection against Clostridium difficile infection in a hamster model by oral vaccination using flagellin FliC-loaded pectin beads. *Vaccine* **36**, 6017–6021.
325 Wang S, Ju X, Heuler J, Zhang K, Duan Z, Warnakulasuriya Patabendige HML, Zhao S & Sun X (2023) Recombinant Fusion Protein Vaccine Containing Clostridioides difficile FliC and FliD Protects Mice against C. difficile Infection. *Infect Immun* **91**.
326 Song WS & Yoon S Il (2014) Crystal structure of FliC flagellin from Pseudomonas aeruginosa and its implication in TLR5 binding and formation of the flagellar filament. *Biochem Biophys Res Commun* **444**, 109–115.
327 Yamaguchi T, Toma S, Terahara N, Miyata T, Ashihara M, Minamino T, Namba K & Kato T (2020) Structural and Functional Comparison of Salmonella Flagellar Filaments Composed of FljB and FliC. *Biomolecules 2020, Vol 10, Page 246* **10**, 246.
328 Kirby JM, Ahern H, Roberts AK, Kumar V, Freeman Z, Acharya KR & Shone CC (2009) Cwp84, a surface-associated cysteine protease, plays a role in the maturation of the surface layer of Clostridium difficile. *Journal of Biological Chemistry* **284**, 34666–34673.
329 Karlsson S, Burman LG & Åkerlund T (1999) Suppression of toxin production in Clostridium difficile VPI 10463 by amino acids. *Microbiology (N Y)* **145**, 1683–1693.
330 Jenny RJ, Mann KG & Lundblad RL (2003) A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expr Purif* **31**, 1–11.
331 Takeya H, Nishida S, Miyata T, Kawada SI, Saisaka Y, Morita T & Iwanaga S (1992) Coagulation factor X activating enzyme from Russell's viper venom (RVV-X). A novel metalloproteinase with disintegrin (platelet aggregation inhibitor)-like and C-type lectin-like domains. *Journal of Biological Chemistry* **267**, 14109–14117.

332 Krishnaswamy S, Church WR, Nesheim ME & Mann KG (1987) Activation of human prothrombin by human prothrombinase. Influence of factor Va on the reaction mechanism. *Journal of Biological Chemistry* **262**, 3291–3299.
333 Hosfield T & Lu Q (1999) Influence of the Amino Acid Residue Downstream of (Asp)4Lys on Enterokinase Cleavage of a Fusion Protein. *Anal Biochem* **269**, 10–16.
334 Boulware KT & Daugherty PS (2006) Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc Natl Acad Sci U S A* **103**, 7583.
335 Oi WL, Chong JPC, Yandle TG & Brennan SO (2005) Preparation of recombinant thioredoxin fused N-terminal proCNP: Analysis of enterokinase cleavage products reveals new enterokinase cleavage sites. *Protein Expr Purif* **41**, 332–340.
336 Waugh DS (2011) An overview of enzymatic reagents for the removal of affinity tags. *Protein Expr Purif* **80**, 283.
337 Dougherty WG, Cary SM & Dawn Parks T (1989) Molecular genetic analysis of a plant virus polyprotein cleavage site: a model. *Virology* **171**, 356–364.
338 Tözsér J, Tropea JE, Cherry S, Bagossi P, Copeland TD, Wlodawer A & Waugh DS (2005) Comparison of the substrate specificity of two potyvirus proteases. *FEBS J* **272**, 514–523.
339 Kapust RB, Toözseór J, Copeland TD & Waugh DS (2002) The P1′ specificity of tobacco etch virus protease. *Biochem Biophys Res Commun* **294**, 949–955.
340 Ng KK, Reinert ZE, Corver J, Resurreccion D, Hensbergen PJ & Prescher JA (2021) A bioluminescent sensor for rapid detection of PPEP-1, a clostridioides difficile biomarker. *Sensors* **21**.
341 Mir Khan U & Selamoglu Z (2020) Use of Enzymes in Dairy Industry: A Review of Current Progress. *Arch Razi Inst* **75**, 131.
342 Hasan MJ, Haque P & Rahman MM (2022) Protease enzyme based cleaner leather processing: A review. *J Clean Prod* **365**, 132826.
343 Sharma PE & Fatima N (2014) Quality Improvement of Wool Fabric Using Protease Enzyme. *Environment and Ecology Research* **2**, 301–310.
344 Gomaa SK, Zaki RA, Wahba MI, Taleb MA, El-Refai HA, El-Fiky AF & El-Sayed H (2022) Green method for improving performance attributes of wool fibres using immobilized proteolytic thermozyme. *3 Biotech* **12**, 254.
345 Niyonzima FN & More S (2015) Detergent-Compatible Proteases: Microbial Production, Properties, and Stain Removal Analysis. *Prep Biochem Biotechnol* **45**, 233–258.
346 Solanki P, Putatunda C, Kumar A, Bhatia R & Walia A (2021) Microbial proteases: ubiquitous enzymes with innumerable uses. *3 Biotech 2021 11:10* **11**, 1–25.
347 Shouket HA, Ameen I, Tursunov O, Kholikova K, Pirimov O, Kurbonov N, Ibragimov I & Mukimov B (2020) Study on industrial applications of papain: A succinct review. *IOP Conf Ser Earth Environ Sci* **614**, 012171.
348 Sanchez MI & Ting AY (2019) Directed evolution improves the catalytic efficiency of TEV protease. *Nature Methods 2019 17:2* **17**, 167–174.
349 Zhou J, Li S, Leung KK, O'Donovan B, Zou JY, DeRisi JL & Wells JA (2020) Deep profiling of protease substrate specificity enabled by dual random and scanned human proteome substrate phage libraries. *Proc Natl Acad Sci U S A* **117**, 25464–25475.
350 Ratnikov B, Cieplak P & Smith JW (2009) High throughput substrate phage display for protease profiling. *Methods in Molecular Biology* **539**, 93–114.
351 Rodi DJ, Soares AS & Makowski L (2002) Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries. *J Mol Biol* **322**, 1039–1052.

352 Krumpe LRH, Atkinson AJ, Smythers GW, Kandel A, Schumacher KM, McMahon JB, Makowski L & Mori T (2006) T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries. *Proteomics* **6**, 4210–4222.
353 Abdul-Khalek N, Wimmer R, Overgaard MT & Gregersen Echers S (2023) Insight on physicochemical properties governing peptide MS1 response in HPLC-ESI-MS/MS: A deep learning approach. *Comput Struct Biotechnol J* **21**, 3715.
354 Meek JL (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition (lipophilicity/separation techniques). *Medical Sciences* **77**, 1632–1636.
355 Mant CT, Burke TWL, Black JA & Hodges RS (1988) Effect of peptide chain length on peptide retention behaviour in reversed-phase chromatogrphy. *J Chromatogr A* **458**, 193–205.
356 Reimer J, Spicer V & Krokhin O V. (2012) Application of modern reversed-phase peptide retention prediction algorithms to the Houghten and DeGraw dataset: peptide helicity and its effect on prediction accuracy. *J Chromatogr A* **1256**, 160–168.
357 Dorfer V, Maltsev S, Winkler S & Mechtler K (2018) CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *J Proteome Res* **17**, 2581–2589.
358 Skiadopoulou D, Vašíček J, Kuznetsova K, Bouyssié D, Käll L & Vaudel M (2023) Retention Time and Fragmentation Predictors Increase Confidence in Identification of Common Variant Peptides. *J Proteome Res* **22**, 3190–3199.
359 Moruz L, Hoopmann MR, Rosenlund M, Granholm V, Moritz RL & Käll L (2013) Mass fingerprinting of complex mixtures: protein inference from high-resolution peptide masses and predicted retention times. *J Proteome Res* **12**, 5730–5741.
360 Moruz L & Käll L (2017) Peptide retention time prediction. *Mass Spectrom Rev* **36**, 615–623.
361 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P & Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589.
362 Zhu W, Shenoy A, Kundrotas P & Elofsson A (2023) Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* **39**.
363 Scardino V, Di Filippo JI & Cavasotto CN (2023) How good are AlphaFold models for docking-based virtual screening? *iScience* **26**, 105920.
364 Borkakoti N & Thornton JM (2023) Alphafold2 protein structure prediction : Implications for drug discovery. *Curr Opin Struct Biol* **78**, 102526.
365 Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M & Baker D (2023) De novo design of protein structure and function with RFdiffusion. *Nature 2023 620:7976* **620**, 1089–1100.
366 Karl-Heinz Leitner, Jeff Butler, Giovanni Cerulli, Theo Dunnewijk, Franziska Kampik, Andrea Kasztler, Huub Meijers, Bianca Poti, Meirion Thomas, Eva Trier, Stig Slipersæter, Rene Wintjes & Jan Youtie (2011) *Analysis of the evolution of the costs of research - trends, drivers and impacts*.

English summary

The group of biomolecules called proteins is essential for all life on Earth, including bacteria. Bacterial proteins serve as structural elements, regulators of gene expression, defense mechanisms, and govern processes such as signaling, communication, adhesion, motility, and pathogenesis. The vast diversity of proteins stems from the countless possible combinations of the building blocks that make up these molecules, i.e., the 20 proteogenic amino acids. However, protein diversity is further increased by the different forms of post-translational modifications (PTMs). Two PTMs, proteolysis and glycosylation, play a central role in this thesis.

Glycosylation is a biochemical process where enzymes facilitate the attachment of carbohydrates to proteins, lipids, or other organic molecules through covalent bonds. This PTM is essential for several bacterial functions, such as constructing the cell wall, forming biofilms, and interacting with host organisms. In pathogenic bacteria, glycosylation plays a pivotal role in virulence by altering surface structures like pili and flagella, which helps the bacteria evade the immune system and adhere to host tissues. For example, flagella of the bacterium *C. difficile* are glycosylated with a unique glycan structure that is synthesized through a cascade of enzymatic reactions. In the case of *C. difficile*, glycosylation of the flagella is essential for flagella-mediated motility.

Proteolysis is the hydrolytic breakdown of proteins into smaller polypeptides or single amino acids by a group of enzymes called proteases. In bacteria, proteases are involved in the breakdown, activation, and maturation of their substrates and thereby ensure the correct functioning of the cells. Proteases exhibit a specificity for certain amino acids surrounding the cleavage site. This specificity can be broad, allowing the protease to cleave a wide range of substrates, or it can be highly specific, recognizing only a single or a few proteins.

A group of highly specific bacterial proteases, the Pro-Pro endopeptidases (PPEPs), possess the unique specificity to hydrolyze their substrates between two proline residues. The cyclic structure of prolines imposes conformational constraints that prevent hydrolysis by other proteases. Besides their preference for two proline residues surrounding the cleavage site, their specificity extends to other surrounding residues as well. Another shared characteristic of PPEPs is their secretion from the cell, indicating an extracellular function. PPEP homologs have been identified in over 130 bacterial species spread over 9 genera.

The first PPEP that was characterized, PPEP-1, was identified in a secretome analysis of *C. difficile. C. difficile* is an anaerobic, endospore-forming, gram-positive bacterium and the leading causative agent of antibiotic-associated diarrhea and colitis. *C. difficile*

infection (CDI) starts with the ingestion of spores by the host. After travelling the gastrointestinal (GI) tract, the opportunistic pathogen is able to colonize the colonic epithelium, especially in patients with a disturbed microbiome. After germination of the spores, the vegetative cells adhere to the gut mucosa and start to produce toxins that damage the host's colonic epithelium and cause disease.

C. difficile is able to bind to the host's intestinal mucosa through the cell wall-tethered adhesion proteins CD2831 and CD3246. However, environmental stimuli such as nutrient stress can prompt the bacteria to release from the gut mucosa and travel to other parts of the colon. To do this, *C. difficile* produces PPEP-1, which hydrolyzes its substrates CD2831 and CD3246 at several cleavage sites, thereby releasing the cells. Interestingly, another putative PPEP exists in *C. difficile* which has not been characterized. This protein, CD1597, possesses a domain similar to other PPEPs, but also displays distinct characteristics.

The research described in this thesis aims to uncover the roles of bacterial enzymes involved in the processes of adhesion and motility in bacteria, with an emphasis on the enzymes' substrate specificities. A summary of the current scientific knowledge that is relevant to the topics discussed in this thesis is provided in **Chapter 1.**

Flagella of the *C. difficile* strain 630Δ*erm* are post-translationally modified with the Type A glycan. While earlier research had established the importance of the Type A glycan for motility, it did not identify a role for one of the biosynthetic proteins, CD0244, which was considered non-essential. In **Chapter 2**, we shed more light on the Type A biosynthetic pathway using quantitative mass spectrometry-based proteomics. We show that CD0244 is essential for the biosynthesis of the Type A glycan. In addition, bioinformatic and structural analyses allowed for the prediction of precise enzymatic functions of the other enzymes involved in the synthesis of the Type A glycan. Based on our results, a new model for the Type A biosynthetic pathway was proposed that provides a basis for future studies.

In **Chapter 3**, we explored the activity and function of CD1597 from *C. difficile*. This protein possesses a PPEP-like domain, but also displays unique features distinguishing it from typical PPEPs. We evaluated the proteolytic activity of CD1597 against various substrates and created a *cd1597* insertional mutant to assess the impact of its absence. Through a series of phenotypic assays, microscopy, and comparative proteomics, we aimed to elucidate the role of this protein in *C. difficile*.

Chapter 4 introduced a new approach for detailed profiling of PPEP specificity. By integrating a synthetic combinatorial peptide library—offering a high diversity of peptides and equimolar concentrations—with the sensitivity and specificity of mass spectrometry, we were able to explore the prime-side specificity of multiple PPEPs,

including PPEP-1, PPEP-2, and a newly identified PPEP from *Geobacillus thermodenitrificans*, now named PPEP-3. This novel PPEP exhibited a distinct prime-side specificity compared to PPEP-1 and PPEP-2.

The newly developed method for profiling PPEP specificity was expanded by creating a complementary combinatorial peptide library to examine non-prime-side specificity in **Chapter 5**. This dual-library approach allowed for the characterization of the full PPEP specificity in a single experiment. We applied this expanded method to analyze known PPEPs, PPEP mutants, and a new PPEP from *Anoxybacillus tepidamans*, and also tested CD1597. By integrating specificity data with structural insights, we enhanced our understanding of the structure-function relationships of PPEPs.

To gain more insight into the structural determinants that govern PPEP specificity, the atomic cocrystal structure of PPEP-3 from *G. thermodenitrificans* was determined by X-ray crystallography in **Chapter 6**. We examined the protease-substrate complex, comparing it with other PPEPs with known structures. Additionally, the substrate specificity was characterized in detail using the newly developed approach that combines synthetic combinatorial peptide libraries with MS detection. Together, these data were used to identify the molecular determinants that explain the substrate specificity of PPEP-3.

Finally, the research presented in this thesis is reflected upon in **Chapter 7**. In addition, we discuss the implications of our findings for future research and potential applications in industry, research, and healthcare settings. Lastly, we provide a framework for future studies that can aid in the understanding of PPEPs and FliC glycosylation.

Nederlandse samenvatting

De groep biomoleculen die eiwitten wordt genoemd, is essentieel voor al het leven op aarde, inclusief dat van bacteriën. Bacteriële eiwitten dienen als structurele elementen, regelaars van genexpressie, verdedigingsmechanismen, en regelen processen zoals signalering, communicatie, adhesie, motiliteit en pathogenese. De enorme diversiteit aan eiwitten komt voort uit de ontelbare mogelijke combinaties van de bouwstenen waaruit deze moleculen zijn opgebouwd, namelijk de 20 proteogene aminozuren. De eiwitdiversiteit wordt nog verder vergroot door de verschillende vormen van post-translationele modificaties (PTM's). Twee PTM's, proteolyse en glycosylering, spelen een centrale rol in dit proefschrift.

Glycosylering is een biochemisch proces waarbij enzymen de koppeling van koolhydraten aan eiwitten, lipiden of andere organische moleculen bewerkstelligen. Deze PTM is essentieel voor verschillende bacteriële functies, zoals het vormen van de celwand, het vormen van biofilms en de interactie met de gastheer. Bij pathogene bacteriën speelt glycosylering een cruciale rol in virulentie door het veranderen van oppervlaktestructuren zoals pili en flagellen, wat de bacteriën helpt het immuunsysteem te ontwijken en zich aan weefsels in de gastheer te hechten. Zo zijn de flagellen van de bacterie *Clostridioides difficile* geglycosyleerd met een unieke glycaanstructuur die wordt gesynthetiseerd door een reeks van enzymatische reacties. In het geval van *C. difficile* is glycosylering van de flagellen essentieel voor flagella-gemedieerde motiliteit.

Proteolyse is de hydrolytische afbraak van eiwitten tot kleinere polypeptiden of losse aminozuren door een groep enzymen genaamd proteases. In bacteriën zijn proteases betrokken bij de afbraak, activering en maturatie van hun substraten en zorgen zo voor de juiste werking van de cellen. Proteases vertonen specificiteit voor bepaalde aminozuren rond de knipplaats. Deze specificiteit kan breed zijn, waardoor de protease een breed scala aan substraten kan knippen, of het kan zeer specifiek zijn, waarbij slechts één of enkele eiwitten worden herkend.

Een groep zeer specifieke bacteriële proteases, de Pro-Pro endopeptidasen (PPEP's), heeft de unieke specificiteit om hun substraten tussen twee prolines te knippen. De cyclische structuur van proline zorgt voor conformationele beperkingen die hydrolyse door andere proteases vaak verhinderen. Naast hun voorkeur voor twee prolines direct rondom de knipplaats, strekt hun specificiteit zich ook uit tot omliggende residuen. Een andere gedeelde eigenschap van PPEPs is de secretie uit de cel, wat wijst op een extracellulaire functie. PPEP-homologen zijn geïdentificeerd in meer dan 130 bacteriesoorten verspreid over 9 genera.

De eerste PPEP die werd gekarakteriseerd, PPEP-1, werd geïdentificeerd in een op massaspectrometrie gebaseerde proteomics analyse van het secretoom van *C. difficile*. *C. difficile* is een anaërobe, endosporenvormende, gram-positieve bacterie en de belangrijkste veroorzaker van met antibiotica geassocieerde diarree en colitis. *C. difficile* infectie (CDI) begint met de opname van sporen door de gastheer. Na ingestie is het opportunistische pathogeen in staat het darmepitheel te koloniseren, vooral bij patiënten met een verstoord microbioom. Na de kieming van de sporen hechten de vegetatieve cellen zich aan de darmmucosa en beginnen ze toxines te produceren die het darmepitheel van de gastheer beschadigen en ziekte veroorzaken.

C. difficile is in staat zich te binden aan de darm mucosa van de gastheer via de aan de celwand gebonden adhesie-eiwitten, waaronder CD2831 en CD3246. Externe stimuli zoals nutriëntenstress kunnen de bacteriën echter aanzetten om zich los te maken van de darmmucosa en zich naar andere delen van de dikke darm te verplaatsen. Om dit te doen, produceert *C. difficile* PPEP-1, dat zijn substraten CD2831 en CD3246 op meerdere plaatsen knipt, waardoor de cellen loskomen. Interessant genoeg bestaat er een andere vermeende PPEP in *C. difficile* die nog niet is gekarakteriseerd. Dit eiwit, CD1597, bezit een domein dat lijkt op andere PPEPs, maar vertoont ook onderscheidende kenmerken.

Het onderzoek beschreven in dit proefschrift heeft tot doel de rollen van bacteriële enzymen die betrokken zijn bij de processen van adhesie en motiliteit in bacteriën te onthullen, met de nadruk op de specificiteit van de substraten van deze enzymen. Een samenvatting van de huidige wetenschappelijke kennis die relevant is voor de in dit proefschrift besproken onderwerpen, wordt gegeven in **Hoofdstuk 1**.

De flagellen van de *C. difficile* stam 630Δ*erm* zijn post-translationeel gemodificeerd met de Type A glycaan structuur. Terwijl eerder onderzoek het belang van de Type A glycaan structuur voor motiliteit had vastgesteld, identificeerde het geen rol voor één van de biosynthetische enzymen, CD0244, dat als niet essentieel werd beschouwd. In **Hoofdstuk 2** werpen we meer licht op de biosynthetische route van de Type A glycaan structuur met behulp van kwantitatieve en op massaspectrometrie-gebaseerde proteomics methoden. Hiermee konden we laten zien dat CD0244 wel degelijk essentieel is voor de biosynthese van Type A. Bovendien maakten bioinformatische en structurele analyses het mogelijk om de precieze enzymatische functies van de andere enzymen die betrokken zijn bij de synthese van de Type A glycaan structuur te voorspellen. Op basis van onze resultaten werd een nieuw model voor de biosynthetische route voorgesteld dat een basis biedt voor toekomstig onderzoek.

In **Hoofdstuk 3** onderzochten we de activiteit en functie van CD1597 uit *C. difficile*. Dit eiwit bezit een PPEP-achtig domein, maar vertoont ook unieke kenmerken die het onderscheidt van andere PPEPs. We evalueerden de proteolytische activiteit van CD1597

tegen verschillende substraten en creëerden een insertiemutant van het *cd1597* gen om de impact van de afwezigheid ervan te beoordelen. Door een reeks fenotypische assays, microscopie en vergelijkende proteomics experimenten probeerden we de rol van dit eiwit in *C. difficile* te verduidelijken.

Hoofdstuk 4 introduceerde een nieuwe benadering voor gedetailleerde karakterisatie van PPEP-specificiteit. Door een synthetische combinatorische peptidebibliotheek met een hoge diversiteit aan peptiden in equimolaire concentraties te integreren met de gevoeligheid en specificiteit van massaspectrometrie, konden we de prime-side specificiteit van meerdere PPEP's onderzoeken. We deden dat voor PPEP-1, PPEP-2 en een nieuw geïdentificeerde PPEP uit *Geobacillus thermodenitrificans*, nu genaamd PPEP-3. Deze nieuwe PPEP vertoonde een andere prime-side specificiteit in vergelijking met PPEP-1 en PPEP-2.

De nieuw ontwikkelde methode voor het karakteriseren van PPEP-specificiteit werd uitgebreid door het creëren van een aanvullende combinatorische peptidebibliotheek om de non-prime-side specificiteit te onderzoeken in **Hoofdstuk 5**. Deze benadering met twee peptidebibliotheken stelde ons in staat om de volledige PPEP-specificiteit in een enkel experiment te karakteriseren. We pasten deze uitgebreide methode toe om bekende PPEPs, PPEP-mutanten, en een nieuwe PPEP uit *Anoxybacillus tepidamans* te analyseren, en testten daarnaast ook CD1597. Door data over specificiteit te integreren met structurele analyses, hebben we ons begrip van de structuur-functie relaties van PPEPs vergroot.

Om meer inzicht te krijgen in de structurele determinanten die de PPEP-specificiteit beïnvloeden, werd de atomaire kristalstructuur van PPEP-3 uit *G. thermodenitrificans* bepaald door middel van röntgenkristallografie in **Hoofdstuk 6**. We onderzochten het protease-substraat complex, en vergeleken het met andere PPEPs met bekende structuren. Daarnaast werd de substraatspecificiteit in detail gekarakteriseerd met behulp van de nieuw ontwikkelde methode die synthetische combinatorische peptidebibliotheken combineert met MS detectie. Samen werden deze gegevens gebruikt om de moleculaire determinanten te identificeren die de substraatspecificiteit van PPEP-3 verklaren.

Ten slotte wordt in **Hoofdstuk 7** op het in dit proefschrift beschreven onderzoek gereflecteerd. Bovendien bespreken we de implicaties van onze bevindingen voor toekomstig onderzoek en potentiële toepassingen in de industrie, onderzoek en gezondheidszorg. Tot slot bieden we een kader voor toekomstige studies die kunnen bijdragen aan onze kennis over PPEPs en FliC-glycosylering.

Curriculum vitae

Bart Claushuis was born on the 24th of July in 1996 in The Hague, the Netherlands. After obtaining his VWO/Gymnasium diploma from the 'Segbroek College' in 2014, he started the bachelor's program Biology at Leiden University. During his last year internship, he studied the genetic instability in *Streptomyces coelicolor* under the supervision of dr. Daniel Rozen. After obtaining his bachelor's degree, Bart enrolled for the master's program Molecular Genetics and Biotechnology at Leiden University and graduated *cum laude* in 2020. During his Msc, he studied bacterial cell-cell fusion using L-form cells at the Microbial Sciences department of the Institute of Biology Leiden under the supervision of prof. dr. Dennis Claessen. During his second internship at the department of Medical Microbiology at the Leiden University Medical Center, Bart studied the regulation of *htrA* transcription in *Clostridioides difficile* under the supervision of dr. ir. Jeroen Corver.

After obtaining his MSc degree, Bart started his PhD research at the Center for Proteomics and Metabolomics at the Leiden University Medical Center under the supervision of dr. Paul Hensbergen, dr. ir. Jeroen Corver, and prof. dr. Manfred Wuhrer. During his PhD research, he studied Pro-Pro endopeptidases (PPEPs), a group of bacterial proteases with a unique cleavage specificity, using molecular biology techniques and mass spectrometry-based proteomics approaches. He helped to develop a new assay that allowed for a detailed characterization of PPEP specificity and developed a personal interest in structural biology. Upon completing his PhD, Bart will continue his career as a researcher outside of academia.

List of publications

In chronological order

Mutational meltdown of putative microbial altruists in *Streptomyces coelicolor* colonies

Z. Zhang, S. Shitut, B. Claushuis, D. Claessen & D.E. Rozen

Nature Communications, 2022, 13:1, 1-9 (not part of this thesis)

Generating Heterokaryotic Cells via Bacterial Cell-Cell Fusion

S. Shitut, M. Shen, B. Claushuis, R.J.E. Derks, M. Giera, D.E. Rozen, D. Claessen, A. Kros

Microbiology Spectrum, 2022, 10, 4 (not part of this thesis)

In-Depth Specificity Profiling of Endopeptidases Using Dedicated Mix-and-Split Synthetic Peptide Libraries and Mass Spectrometry

B. Claushuis, R.A. Cordfunke, A.H. de Ru, A. Otte, H.C. van Leeuwen, O.I. Klychnikov, P.A. van Veelen, J. Corver, J.W. Drijfhout, P.J. Hensbergen

Analytical Chemistry, 2023, 95, 31, 11621 – 11631 (Chapter 4)

Revised Model for the Type A Glycan Biosynthetic Pathway in *Clostridioides difficile* Strain 630Δ*erm* Based on Quantitative Proteomics of *cd0241–cd0244* Mutant Strains

B. Claushuis, A.H. de Ru, S.A. Rotman, P.A. van Veelen, L.F. Dawson, B.W. Wren, J. Corver, W.K. Smits, P.J. Hensbergen

ACS Infectious Diseases, 2023, 9, 12, 2665 – 2674 (Chapter 2)

Non-prime- and Prime-side Profiling of Pro-Pro Endopeptidase Specificity Using Synthetic Combinatorial Peptide Libraries and Mass Spectrometry

B. Claushuis, R.A. Cordfunke, A.H. de Ru, J. van Angeren, U. Baumann, P.A. van Veelen, M. Wuhrer, J. Corver, J.W. Drijfhout, P.J. Hensbergen

The FEBS journal, 2024, 17160 (Chapter 5)

Characterization of the *Clostridioides difficile* 630Δ*erm* putative Pro-Pro endopeptidase CD1597

B. Claushuis, A.H. de Ru, P.A. van Veelen, P.J. Hensbergen, and J. Corver

Accepted for publication in Access Microbiology (Chapter 3)

Biochemical and structural characterization of PPEP-3 from *Geobacillus thermodenitrificans*

B. Claushuis[§], F. Wojtalla[§], H.C. van Leeuwen, J. Corver, U. Baumann[¶], Paul J. Hensbergen[¶]

In preparation

[§] and [¶]: Authors contributed equally to this work (Chapter 6)

Protease specificity profiling using synthetic combinatorial peptide libraries and mass spectrometry

B. Claushuis, R.A. Cordfunke, A.H. de Ru, P.A. van Veelen, J. Corver, J.W. Drijfhout, P.J. Hensbergen

In preparation for publishing in Methods in Molecular Biology book series (Springer) (not part of this thesis)

PhD portfolio

PhD student	Bart Claushuis
PhD period	2020-2024
Promotor	prof. dr. Manfred Wuhrer
Copromotors	dr. Paul J. Hensbergen and dr. ir. Jeroen Corver

Leiden University Medical Center, Center for Proteomics and Metabolomics

Courses

- Basiscursus Regelgeving en Organisatie voor Klinisch onderzoekers (BROK)
- Basic Methods and Reasoning in Biostatistics
- PhD Introductory Meeting
- Academic Writing for PHDs
- Using R for Data Analysis
- Use Your Brain
- Job Orientation
- Job Search skills

Conferences and presentations

- International Clostridium Difficile Symposium, Online, 2020. A*ttendance*
- ClostPath 2021, Online, 2021. *Poster presentation*
- Netherlands Proteomics Platform, Utrecht, The Netherlands, 2022. *Oral presentation*
- GRC Proteolytic Enzymes and Their Inhibitors, Lucca, Italy, 2022. *Poster presentation*
- Netherlands Proteomics Platform, Utrecht, The Netherlands, 2032. *Attendance*
- ClostPath 2023, Banff, Canada, 2023. *Poster presentation*

Supervision of students

2023 Erawan van der Veere Leiden University Medical Center

Teaching activities

2021 Lecture on proteomics, proteases and N-terminomics (LUMC)

2022 Young Scientist 4 a Day (Stedelijk Gymnasium Leiden, the Netherlands)

2022 Workgroup Infectious Agents and Immunity (LUMC)

2023 Science4U (Stedelijk Gymnasium Leiden, the Netherlands)

2023 Lecture on N-terminomics and combinatorial peptide libraries (LUMC)

2024 Science4U (Stedelijk Gymnasium Leiden, the Netherlands)

Acknowledgements

This thesis is the culmination of four years of work and would not have been possible without the contributions and support of many individuals, to whom I am deeply grateful.

Jeroen, I want to thank you for your show of confidence in me by suggesting to apply for a PhD position at the CPM, not knowing you would be one of two people who would read my application. Over the years, you have taught me so much about molecular biology and made me a far better researcher than I was before I met you.

Paul, due to your endless time and energy, I went from a mass spec rookie to a mass spec expert in no time and we have spent many hours together behind a PC, looking at peaks, peptides, and proteins. I greatly appreciate that your door was always open and that I was always welcome to annoy you with yet another question.

I would like to thank Manfred for being my promotor and for the opportunity to work at the Center for Proteomics and Metabolomics. Your guidance and advice throughout my years at CPM and your support as I completed my PhD are greatly appreciated.

My thanks go to all my colleagues from both the CPM and LUCID for all their assistance in the lab, for sharing their knowledge, and for the fun times together. I especially want to thank Peter, Arnoud, and Rayman, who were there from the start of my PhD and had a great part in realizing this thesis. Rayman, I'm also grateful to you for being my paranymph and your support during and around my defense. I also want to thank Sarah, Jordy, Annemarie, and George for all their contributions, as well as Wiep Klaas and Annemieke for their help with experiments and for all their valuable insights.

I'm also thankful to Uli and Fabian, our collaborators from Cologne, for our fruitful discussions via Skype and for solving the PPEP-3 structure. My interest in structural biology grew significantly during my PhD, and I am grateful to you for sharing your knowledge and the assistance you provided in my analyses.

Also, a big thank you to all my friends for supporting me during my PhD. You guys have helped me a lot with all the great times we had. Luc, I'm glad that you, as my oldest friend, can be part of this thesis by designing the cover. Justin, I want to thank you for being my paranymph. Kirsten, I owe you a great amount of thanks as well. Over the past four years, you have always believed in me and provided a lot of support. The final months of writing this thesis were a little stressful at times, and I am grateful for your patience and understanding during those moments.

Pap, ook al zal jij het waarschijnlijk niet beseffen, dit proefschrift was nooit tot stand gekomen zonder jou. Vanaf mijn geboorte heb jij mij weten te interesseren voor de

natuur en de biologie. Ondanks dat ik voornamelijk in de stad ben opgegroeid, nam jij mij altijd mee naar waar de dieren waren. De vele dieren bij jou thuis, ieder weekend naar de boerderij, het samen jagen in het buitenland, het leren van de Latijnse namen van dieren en vele andere activiteiten hebben ervoor gezorgd dat biologie mijn passie is geworden. Hiervoor ben ik jou eeuwig dankbaar.

Mama, van kleins af aan heb jij mij altijd gesteund in al mijn keuzes en dat heeft ervoor gezorgd dat ik op dit punt ben gekomen. Jij bent altijd mijn steun en toeverlaat geweest, zeker tijdens één van onze vele reisjes die wij de afgelopen jaren hebben gemaakt. Er zijn niet genoeg woorden om jou te bedanken voor alles wat je voor mij hebt betekend, maar ik hoop dat je weet hoe zeer ik alles waardeer.