Title

Engineering peptide ligase specificity by proteomic identification of ligation sites

Journal

Nature Chemical Biology, 14(1)

ISSN

1552-4450

Authors

Weeks, Amy M

Wells, James A

Publication Date

2018

DOI

10.1038/nchembio.2521

Peer reviewed

HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2018 May 20.

Published in final edited form as:

Nat Chem Biol. 2018 January ; 14(1): 50–57. doi:10.1038/nchembio.2521.

Engineering peptide ligase specificity by proteomic identification of ligation sites

Amy M. Weeks[1] and **James A. Wells**[1,2,*]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California, USA

[2]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California, USA

Abstract

Enzyme-catalyzed peptide ligation is a powerful tool for site-specific protein bioconjugation, but stringent enzyme–substrate specificity limits its utility. Here, we present an approach for comprehensive characterization of peptide ligase specificity for N termini using proteome-derived peptide libraries. We used this strategy to characterize the ligation efficiency for >25,000 enzyme–substrate pairs in the context of the engineered peptide ligase subtiligase and identified a family of 72 mutant subtiligases with activity toward N-terminal sequences that were previously recalcitrant to modification. We applied these mutants individually for site-specific bioconjugation of purified proteins including antibodies, and in algorithmically selected combinations for sequencing of the cellular N terminome with reduced sequence bias. We also developed a web application to enable algorithmic selection of the most efficient subtiligase variant(s) for bioconjugation to user-defined sequences. These studies provide a new toolbox of enzymes for site-specific protein modification and a general approach for rapidly defining and engineering peptide ligase specificity.

Introduction

Site-specific protein modification strategies have enabled a wide array of advances in the biological sciences, including development of probes of enzyme function[1–3], discovery of enzyme inhibitors and drugs[4–6], synthesis of antibody–drug conjugates[7,8], and implementation of advanced imaging techniques[9]. Site-specific strategies include modification of engineered cysteine or methionine residues[2,7,10], enzymatic ligation to genetically encoded sequence epitopes[11–13], introduction of unnatural amino acids[14,15], and native chemical ligation[16–18]. Although these methods have proven powerful for a number

*To whom correspondence should be addressed: Prof. James A. Wells, Jim.Wells@ucsf.edu.

Author contributions

A.M.W. and J.A.W. designed the research. A.M.W. performed all experiments, wrote data analysis scripts, and built the ALPINE web application. A.M.W. and J.A.W. analyzed data and interpreted results. A.M.W. and J.A.W. wrote the manuscript.

Competing financial interests.

A.M.W., J.A.W., and the Regents of the University of California have filed a patent application (U.S. Provisional Patent Application No. 62/398,898) related to engineered subtiligase variants.

of applications, they require genetic engineering of the protein of interest, which may disrupt biological function, reduce expression yield, and limit their utility as chemoproteomic probes.

The N terminus is a universal feature of all proteins that is an attractive handle for site-specific protein modification based on its uniqueness within each polypeptide chain[19,20]. Although a number of chemical strategies target the N terminus[20,21], they are commonly limited by poor selectivity, the requirement for particular N-terminal residues, or the inability to form a native peptide bond. Based on their ability to target the N terminus and to generate a native peptide bond, peptide ligases including *Staphylococcus aureus* sortase[11] and *Clitoria ternatea* butelase 1 (ref. 22), which catalyze transpeptidation reactions, have been applied as an alternative strategy for site-specific protein bioconjugation. However, these enzymes retain strict sequence specificity, often requiring genetic engineering of the target ligation site. In contrast, the engineered peptide ligase subtiligase, which catalyzes a ligation reaction between a peptide ester or thioester and the N-terminal α-amine of a protein or peptide[23] (Fig. 1a), has broader specificity and higher catalytic efficiency (>10^5 M^{-1} s^{-1}) compared to sortase[11,24] or butelase 1 (ref. 22). However, qualitative specificity studies show that this enzyme harbors sequence specificity that limits its utility for N-terminal bioconjugation. Furthermore, incomplete characterization of its prime-side (P'; C-terminal to the newly formed peptide bond)[25] specificity makes the suitability of subtiligase for any particular application difficult to predict.

Here, we present a strategy for comprehensive characterization of peptide ligase prime-side specificity that utilizes database-searchable, proteome-derived peptide libraries as a platform for rapid characterization and engineering of ligase specificity. Inspired by a method for mapping protease sequence specificity[26], this approach, termed Proteomic Identification of Ligation Sites (PILS), enables selective isolation of ligated peptides and sequencing by liquid chromatography-tandem mass spectrometry (LC-MS/MS) for rapid determination of positional enrichment or de-enrichment of each amino acid at each P' site. We deployed PILS to engineer a family of peptide ligases with defined specificities, greatly expanding the enzymatic toolbox for site-specific modification of protein N termini.

Results

Proteome-derived peptide libraries to map specificity

To enable comprehensive characterization of subtiligase prime-side specificity, we developed a mass spectrometry-based assay, PILS, inspired by the Proteomic Identification of protease Cleavage Sites (PICS) method for determining protease specificity[26]. We generated diverse α-amine acceptor peptide libraries by digesting the *E. coli* proteome with two proteases of orthogonal P1 specificity: trypsin (P1 = K or R) or Glu-C (P1 = E or D). This produced two libraries that, in combination, represent every possible single amino acid from P1'–P6', and nearly all 400 P1'–P2' dipeptide combinations, with only proline and cysteine underrepresented (Supplementary Fig. 1). We incubated each acceptor peptide library with subtiligase (1 μM) and a limiting amount of the donor peptide ester 1 (200 μM, Supplementary Fig. 2) that contains an N-terminal biotin for avidin capture, a TEV protease cleavage site for unbiased[27] proteolytic release, and an aminobutyric acid (Abu) mass tag for

unequivocal confirmation that ligation occurred[28,29] (Fig. 1b). Both the unenriched input libraries and the enriched, eluted peptides were analyzed by LC-MS/MS to quantify the frequency with which each amino acid appeared in each position. In each enriched sample, >2,000 Abu-tagged subtiligase substrate peptides were identified (Supplementary Dataset 1).

To determine subtiligase sequence specificity, we evaluated the position-specific differences in the frequencies of each amino acid in enriched samples compared to the input libraries by calculating an enrichment score (Online Methods). We observed no substantial sequence specificity beyond P2' of the ligated peptide, consistent with previous structural studies of subtilisin (Fig. 1c, Supplementary Table 1). However, we observed substantial sequence preferences in both the P1' and P2' positions. Small amino acids (Ala, Gly, and Ser), Met, and Arg were enriched at P1', while acidic residues (Asp and Glu), branched-chain amino acids (Ile, Leu, Thr, and Val), Pro, and Gln were de-enriched. At P2', aromatic (Phe, Trp, and Tyr) and large hydrophobic (Ile, Leu, and Val) residues were enriched, while acidic, basic, and polar residues (Asn, Gln, and Ser), Gly, and Pro were de-enriched. These results are in general agreement with previous qualitative studies of subtiligase specificity[30,23], demonstrating the validity of the PILS method for determining ligase specificity.

Substrate amino acid subsites in proteases often exhibit cooperativity that cannot be assessed by evaluating one position at time[31]. To examine the role of subsite (S) cooperativity in sequence recognition at P1' and P2' of subtiligase, we measured the enrichment or de-enrichment of each dipeptide sequence relative to the input library. We performed hierarchical clustering on the enrichment scores and identified five clusters of dipeptide sequences that behave similarly to one another (Fig. 2a, Supplementary Dataset 1). Sequences in cluster 1 are good substrates for subtiligase and have a small amino acid, Met, or Arg in the P1' position and an aromatic or large hydrophobic residue, His, Met, Cys, or Ala in the P2' position. Sequences in cluster 2 are also good substrates for subtiligase, and have any amino acid except Asp or Glu at P1' and an aromatic amino acid at P2'. Sequences in cluster 3 are poor substrates for subtiligase and contain aromatic, large hydrophobic, polar, or acidic residues at P1', and acidic, basic, or polar residues at P2'. Sequences in clusters 4 and 5 are also poor substrates for subtiligase and contain acidic residues at P1' and P2', respectively. Subsite cooperativity is apparent based on examination of the dipeptide heatmap (Fig. 2a). For example, when P2' is a favorable aromatic residue, any amino acid except Asp or Glu is accepted in the P1' position. Similarly, when a favorable amino acid is present in the P1' position, a broader set of amino acids are accepted at P2'. These results suggest that energetically favorable interactions at one subsite can help to overcome weak or unfavorable interactions at the other subsite, expanding the total number of sequences that can be efficiently ligated by subtiligase.

Defining the S1' and S2' subsites by alanine scanning

While the S1–S4 pockets of subtilisin have been defined through extensive structural, biochemical, and mutagenesis studies[32–34], the S1' and S2' pockets are poorly understood. Because protease specificity is sometimes determined by elements outside of the substrate binding site[35], we chose a functional approach to determine key residues for substrate recognition. As a starting point, we chose wild-type subtiligase (subtilisin-S221C/P225A),

as mutations introduced to improve subtiligase stability have not been characterized with respect to their impact on substrate specificity. To determine which residues contribute to substrate binding, we purified alanine mutants of twenty residues within 7 Å of the catalytic triad (Supplementary Table 2) and quantified the resultant changes in ligation specificity using the PILS method (Fig. 2b,c, Supplementary Figure 3, Supplementary Datasets 2–21). Residues were interpreted to contribute substantially to prime-side specificity if the corresponding alanine mutation led to a change in the distribution of Abu-labeled peptides in any of the five previously defined sequence clusters by at least two standard deviations compared to the wild-type enzyme. Of the twenty alanine mutants studied, eight mutants showed substantial changes in specificity and could be divided into four classes based on similarities in their patterns of specificity change.

Class I mutants (D60A, N62A, S63A, and S125A) are characterized by enhanced labeling of poor substrates (clusters 3, 4, and 5) and decreased labeling of good substrates (clusters 1 and 2), representing apparent broadened specificity (Fig. 2c, Supplementary Fig. 3). Conversely, class II mutants (H76A and L126A) show the opposite pattern. Kinetic analysis of these mutants revealed that class I mutations increase k_{cat}/K_M by 4–8-fold compared to the wild-type enzyme, while class II mutations decrease k_{cat}/K_M by 4–16-fold (Supplementary Fig. 4). The apparent changes in specificity therefore likely result from catalytic differences rather than increased or decreased molecular recognition of particular substrate sequences. Consistent with this conclusion, a structure of subtilisin bound to the *Streptomyces* subtilisin inhibitor (SSI)[34] suggests that these residues are far from the site in which the prime side of an acceptor peptide substrate is expected to bind (Fig. 2b, Supplementary Fig. 5). Therefore, although some of these mutants are useful for enhancing subtiligase's catalytic efficiency, class I and class II residues do not functionally contribute to binding of P1' and P2' substrate residues.

Class III includes only one variant, F189A, which is characterized by a decrease in modification of cluster 2 sequences and an increase in modification of cluster 3 and cluster 5 sequences (Fig. 2c, Supplementary Fig. 3). The identity of the P2' residue defines both cluster 2 (P2' = aromatic) and cluster 5 (P2' = acidic), and contributes to cluster 3 (P2' = charged, polar, Pro, or Gly), suggesting that Phe 189 is a major determinant of P2' specificity and makes up part of the S2' pocket. Similarly, class IV only includes one variant, Y217A, and is characterized by an increase in modification of cluster 4 sequences and a decrease in labeling of cluster 3 sequences (Fig. 2c). Cluster 4 is defined by the identity of the P1' residue (P1' = acidic), suggesting that Tyr217 is an important determinant of P1' specificity. Consistent with these conclusions, both Phe189 and Tyr217 lie near the predicted prime-side interaction site based on the subtilisin–SSI structure[34].

Engineering subtiligase specificity

Based on the functional analysis above, we targeted 'hot spot' positions 189 (S2' pocket) and 217 (S1' pocket) for saturation mutagenesis to fully explore how these positions impact specificity. We purified the 36 additional single mutants (expression yields varied between 20–100 mg/L, Supplementary Table 2) and analyzed their specificity profiles using PILS (Fig. 2d, Supplementary Fig. 6, Supplementary Datasets 22–52). At position 217, the

Y217K and Y217R mutations led to substantial improvements in modification of sequences with an acidic P1' residue relative to both the wild-type enzyme and the Y217A mutant. Providing validation of our screening strategy, the Y217K mutation was previously rationally incorporated into subtiligase to enable Cys-free, enzyme-catalyzed expressed protein ligation in the context of acidic P1' sequences[18]. The Y217D and Y217E variants showed improved modification of peptides containing His, Lys, Ser, or Arg at P1'. At position 189, the F189S, F189Q, F189K, and F189R mutations led to substantial improvement in modification of sequences with an acidic P2' residue and diminished labeling of sequences with an aromatic residue at P2'. Additionally, although they didn't shift specificity on average across an entire cluster, a number of other mutations led to improved or diminished modification of specific sequences (Supplementary Fig. 6). Together, these results indicate that introduction of a charged or polar residue in either the S1' or S2' pocket creates the opportunity for new, favorable electrostatic or hydrogen-bonding interactions with charged or polar peptide substrates, expanding the number of N-terminal dipeptide sequences that can be efficiently labeled with subtiligase. These mutants represent a new toolbox of peptide ligases that can be deployed for a variety of applications based on their specific N-terminal specificity requirements.

While a number of F189 variants showed useful changes in sequence specificity, many of these variants expressed at much lower levels compared to wild-type subtiligase. Furthermore, when we characterized these variants by LC-MS, we observed a 16-Da mass modification consistent with methionine oxidation (Supplementary Table 2). Previous studies of subtilisin demonstrated that Met222 is prone to an oxidation event that impacts enzyme activity, and protein engineering work determined that substitution of alanine or glycine at this position improves subtilisin activity[36] and enhances the aminolysis-to-hydrolysis ratio in the context of subtiligase[37]. We therefore introduced the F189 and Y217 mutations with substantial specificity changes into the more stable heptamutant of subtiligase (M50F, N76D, N109S, M124L, S125A, K213R, N218S), termed stabiligase[30], or an octamutant that also incorporated the M222A mutation. These variants expressed at levels comparable to wild-type subtiligase and maintained specificity profiles indistinguishable from the mutants in the subtiligase background (Supplementary Fig. 6, Supplementary Datasets 53–72). We found that introduction of the M222A mutation both eliminated the observed oxidation event and improved the subtiligase peptide ligation-to-hydrolysis ratio (Supplementary Table 2, Supplementary Fig. 7), consistent with previous studies. To further enable adoption of subtiligase-catalyzed bioconjugation, we also constructed a pro-domain- and Ca^{2+}-independent variant suitable for expression in *E. coli* and one-step affinity purification. This variant was purified in high yield (~20 mg/L) and, in the context of the Y217K mutant, exhibited a specificity profile similar to subtiligase-Y217K and stabiligase-Y217K (Supplementary Fig. 6, Supplementary Dataset 73).

In the context of stabiligase, we also examined the specificity of a number of F189/Y217 double mutants, including F189K/Y217K, F189R/Y217K, F189R/Y217D, F189S/Y217K, and F189Q/Y217D (Supplementary Fig. 8, Supplementary Datasets 61–72). The double mutants impacted sequence specificity in a predictable way based on PILS analysis of the single mutants, suggesting that our PILS specificity datasets can be leveraged for the design of tailor-made mutants to label specific N-terminal sequences.

Protein scope for subtiligase-catalyzed bioconjugation

To examine the scope of N-terminal protein sequences that can be targeted for biocojugation using our new panel of subtiligase mutants, we generated *E. coli* lysate under native conditions and incubated it with wild-type stabiligase, stabiligase-M222A, stabiligase-Y217K/M222A, or stabiligase-F189R/M222A and biotinylated peptide ester **1** (Fig. 3a, Supplementary Fig. 9). Following labeling, we isolated the biotinylated proteins on immobilized neutravidin, digested with trypsin to remove internal peptides, selectively eluted the Abu-tagged N-terminal peptides with TEV protease, and sequenced them by LC-MS/MS. Each subtiligase variant labeled >200 native proteins at their translational N-termini or at annotated signal peptide cleavage sites (Fig. 3a). Compared to stabiligase alone, the mutants increased the number of native proteins that could be labeled by 50%, from 250 to 374, with the N-terminal sequences of the additional proteins reflecting the altered specificity of the mutants as measured by PILS (Fig. 3a, Supplementary Figs. 9 and 10, Supplementary Dataset 74).

To examine the utility of engineered subtiligase mutants for high-yield protein bioconjugation, we tested the ability of the engineered mutants to modify recombinant antibodies, an important class of therapeutic proteins (Fig. 3b). Our group, as part of the Recombinant Antibody Network (RAN), has produced an automation platform for producing thousands of recombinant antibodies to more than 500 protein targets using a synthetic Fab library displayed on filamentous phage[38]. Because all of these antibodies are built on a single scaffold derived from Trastuzamab[39], they have a common N-terminal light-chain sequence of Ser-Asp and a common N-terminal heavy-chain sequence of Glu-Ile. Based on PILS analysis of wild-type subtiligase, both of these N-terminal sequences are predicted to be poor ligation substrates. To test this prediction, we attempted to ligate azide-bearing peptide ester **2** onto the N-terminus of an anti-GFP antibody (αGFP) that we constructed[40]. Indeed, αGFP was completely refractory to modification with wild-type subtiligase (Supplementary Figs. 11 and 12). Based on the PILS specificity maps, we predicted that αGFP could be labeled on the heavy chain by the Y217K mutant, and indeed, we observed quantitative labeling on the heavy chain by subtiligase-Y217K within 1 h (Fig. 3b), demonstrating that high-yield bioconjugation can be achieved by judicious matching of enzyme and substrate. PILS also predicted that the Ser-Asp N terminus of the light chain would be labeled specifically by the stabiligase-F189R/M222A mutant having favorable specificity for P2' Asp. However, we did not observe measurable labeling within 1 h in this context, while overnight incubation produced the peptide-antibody bioconjugate in 11.4% yield (Supplementary Fig. 11). We hypothesized that the inefficiency of light-chain modification was due to inaccessibility of the N terminus, a limitation that has been demonstrated to impact yield from other N-terminal modification methods[21]. Indeed, we observed increased modification yields when we extended the N terminus by one (21%, Gly), two (32%, Gly-Gly), three (53%, Gly-Gly-Gly), or four (62%, Gly-Gly-Gly-Ser) amino acids while maintaining the native N-terminal sequence (Supplementary Fig. 11). These results suggest that inefficiency due to N-terminal inaccessibility can be overcome by multiple rounds of labeling when genetic modification is not possible, or by modification of the N terminus to enhance its accessibility.

We next set out to test whether orthogonality could be achieved in the context of the αGFP heterodimer. We extended the N terminus by three residues (Ala-Phe-Ala) having a favorable sequence for wild-type subtiligase. Within 1 h, we observed specific and quantitative labeling of only the light chain with wild-type subtiligase and labeling of both the heavy and light chains with subtiligase-Y217K (Supplementary Fig. 12). These results demonstrate that careful selection of subtiligase mutants matched to their optimal substrates by PILS can produce orthogonal labeling in the context of heterodimers or protein mixtures.

We next asked whether a small panel of subtiligase mutants covering a broad swath of sequence space could label a protein without knowledge of the N-terminal target sequence. To test this, we purchased engineered recombinant protein A, whose N-terminal sequence was unknown to us. We tested a panel of five stabiligase mutants and discovered that one mutant, Y217K, could indeed label protein A quantitatively (Fig. 3b, Supplementary Fig. 13), demonstrating the feasibility of subtiligase modification even in the absence of sequence information.

We next tested the labeling yield produced by engineered subtiligases using green fluorescent protein (GFP) with its native N terminus (Met-Val), or engineered N termini that were either good (Ala-Phe) or poor (Glu-Phe, Asp-Phe, Ala-Glu, and Ala-Asp) substrates for wild-type subtiligase (Fig. 3b, Supplementary Table 3). These GFP variants were modified by the azide-bearing peptide ester **2** and a panel of ten subtiligase mutants: Y217K, F189K, and F189R in the context of wild-type subtiligase, stabiligase, or stabiligase-M222A (Fig. 3c). We also included a fourth set of reaction conditions in which the GFP N-terminal variants were labeled a second time following desalting of the reaction mixture (Fig. 3c). As predicted by PILS, Ala-Phe-GFP could be labeled nearly quantitatively with wild-type subtiligase, stabiligase, and stabiligase-M222A (Fig. 3c). In contrast, the labeling yield for the remaining sequences was poor (<25%) when subtiligases retaining wild-type specificity were used. However, Glu-Phe-GFP and Asp-Phe-GFP could be labeled much more efficiently with variants harboring the Y217K mutation (Fig. 3c). For Glu-Phe-GFP, ≥90% bioconjugation yield was achieved by subtiligase-Y217K, stabiligase-Y217K, and stabiligase-M222A/Y217K, while for Asp-Phe-GFP, 95% yield was achieved after two rounds of labeling with stabiligase-M222A/Y217K. In contrast, all other subtiligase variants gave <10% yield with this sequence. Similarly, for Ala-Glu-GFP and Ala-Asp-GFP, ≥90% yield was achieved after two rounds of labeling with stabiligase-M222A/F189K and stabiligase-M222A/F189R. Although the native GFP N terminus is predicted to be a good substrate for wild-type subtiligase, labeling yields were poor with subtiligases retaining wild-type specificity (Fig. 3c). However, >95% yield could be achieved when two rounds of labeling with stabiligase-M222A were performed. Because native GFP is two residues shorter than the other variants tested, this suggested that N-terminal accessibility could be a limiting factor in subtiligase labeling yields. Our data suggest that poor N-terminal accessibility can be overcome with multiple rounds of labeling or introduction of a short N-terminal extension.

One-step and modular protein modification strategies

We next set out to develop reagents and protocols for both one-step and modular bioconjugation of diverse payloads to protein N termini using subtiligase. To enable one-step protein modification, we designed an N-terminally capped (succinylated) peptide containing a single free lysine at the subtiligase P5 position, outside the substrate recognition sequence (**3**; Supplementary Fig. 14). This free amine was readily acylated with a commercially available biotin N-hydroxysuccinimide ester (NHS ester), enabling site-specific biotinylation with a typically non-specific reagent that would normally target all surface-exposed lysines in a protein. Numerous NHS ester reagents are commercially available and can be converted to site-specific reagents using this strategy, making this a versatile approach for one-step modification of proteins with diverse payloads.

We also developed a modular bioconjugation protocol by using subtiligase to incorporate a bioorthogonal azide group at the protein N terminus (**2**; Fig. 4a). This azide can be modified after incorporation into the protein by copper-catalyzed or copper-free azide–alkyne click chemistry with commercially available alkynes or dibenzyocyclooctynes (DBCOs) or by Bertozzi–Staudinger ligation[41] with commercially available phosphine reagents. Using this modular strategy, we incorporated an azide into αGFP and then used this as a starting material for modification with a number of DBCO reagents to produce biotinylated αGFP (DBCO-biotin) (Fig. 4a), fluorescent αGFP (DBCO-Cy3), a αGFP–drug conjugate (DBCO-monomethyl auristatin E), an oligonucleotide-modified αGFP (5'-DBCO-oligonucleotide), and a PEGylated αGFP (DBCO-PEG 5000) (Fig. 4b). Importantly, these modifications led to only small decreases in affinity of the αGFP for GFP, demonstrating that protein function is maintained upon modification (Supplementary Table 4).

To test the utility of these conjugates in a biological context, we employed a HEK-293T cell line modified for doxycycline (Dox)-inducible expression of cell-surface GFP in combination with Cy3-αGFP (Fig. 4c). In Dox-induced cells, we observed binding of the Cy3-αGFP and co-localization of the Cy3 and GFP signals. In contrast, no Cy3-αGFP binding was observed in un-induced cells. These results demonstrate the utility of subtiligase-catalyzed N-terminal modification for incorporating probes into proteins while maintaining their biological functions.

Cellular N-terminomics with subtiligase cocktails

Previously, our lab had applied subtiligase as a tool for enrichment of proteolytic neo-N termini in the context of apoptotic proteolysis catalyzed by caspases[28,29,42]. Caspase P1'–P2' specificity fortuitously encompasses the most preferred sequences for subtiligase ligation based on the results of our PILS studies[43] (Fig. 5a). However, many other proteases have different P1'–P2' specificity. We hypothesized that applying cocktails of subtiligase mutants for enrichment of neo-N termini generated by these proteases would more comprehensively capture their prime-side sequence specificity. To test this hypothesis, we analyzed the substrates of two proteases, methionine aminopeptidase (MetAP) and signal peptide peptidase (SPP), which have divergent specificities[44,45]. For comparison, we also analyzed apoptotic proteolysis (Fig. 5b). Using our PILS datasets, we algorithmically selected the three stabiligase mutants, F189S/Y217K, F189D, and Y217D/M222A, that are capable of

modifying the maximum number of N-terminal dipeptide sequences with an enrichment score ≥0 in combination with wild-type stabiligase. We labeled N termini in Jurkat cell lysate with 2.5 mM biotinylated ester **1** and either a mixture of 1 μM of each mutant (4 μM total enzyme) or 4 μM stabiligase. Following sequencing of the N-terminal peptides by LC-MS/MS, we identified >1300 unique N termini from >650 unique proteins in each sample (Supplementary Dataset 75).

MetAP substrates within our datasets were identified by the presence of an Abu tag at residue 2 within a protein. Biochemical studies of the human MetAPs have demonstrated that they prefer small amino acids in the P1' position[44], similar to the sequence preference of wild-type stabiligase. Because of this preference, we predicted that both stabiligase and the stabiligase cocktail would capture MetAP selectivity equally well. Indeed, both stabiligase and the cocktail captured cleavage events at Met-Ala, Met-Gly, Met-Ser, Met-Val, and Met-Thr sequences at similar frequencies (Fig. 5c), accurately reflecting MetAP specificity.

SPP substrates in our datasets were identified by the presence of an Abu tag at a predicted SPP cleavage site as annotated in the Uniprot Knowledgebase[46]. In contrast to MetAP, SPP has less P1' specificity based on a proteome-wide survey of its predicted cleavage sites in Uniprot (n = 3,449) (Supplementary Dataset 76), and we predicted that the stabiligase cocktail would more accurately capture this based on its more efficient labeling of polar and charged sequences. For stabiligase, we observed that P1' Asp, Glu, His, and Gln were under-represented compared to the true SPP specificity, while Ala, Gly, Leu, and Ser were over-represented (Fig. 5d, Supplementary Table 5). The cocktail, in contrast, exhibited lower sequence capture bias, with only P1' Gln under-represented and P1' Gly and Ser over-represented (Fig. 5d, Supplementary Table 6). The stabiligase cocktail therefore broadens sequence coverage, enabling capture of neo-N termini that more accurately reflect protease specificity. Although we optimized this cocktail for broad sequence coverage, it is also possible to design custom subtiligase cocktails for the study or proteases of known prime-side specificity using the toolbox of mutants and PILS specificity maps that we have generated.

A web-based tool for subtiligase variant selection

We have developed a web-based tool, ALPINE (α-Amine Ligation Profiling Informing N-terminal modification Enzyme selection), to enable the chemical biology community to leverage the 72 PILS datasets encompassing >25,000 enzyme–substrate pairs that we have collected for selection of optimal subtiligase variants for protein and peptide N-terminal modification applications. ALPINE enables exploration of PILS datasets, identification of the most efficient subtiligase variant for modifying a particular N-terminal sequence, algorithmic selection of customized cocktails for modifying user-defined groups of sequences, and analysis of user-generated PILS datasets. This tool, along with tutorials and sample data, is freely accessible on the web at https://wellslab.ucsf.edu/alpine.

Discussion

In recent years, the pace of discovery and development of peptide ligases has accelerated, opening up new avenues for site-specific protein bioconjugation. However, the usefulness of

these enzymes depends on well-defined, predictable sequence specificity. The PILS strategy enables comprehensive characterization of prime-side ligase specificity to meet this challenge. PILS is generalizable to other peptide ligase enzymes, as well as to the study of chemical reactions that target the N terminus, providing a platform for the development of new N-terminal modification strategies to enable site-specific protein modification.

The ALPINE web tool that we have developed (https://wellslab.ucsf.edu/alpine) enables users to identify subtiligase mutants optimized for a particular target of interest, thus eliminating the need for genetic engineering to achieve site-specific protein modification in many cases. In cases that do require genetic modification, such as those in which the native protein N terminus is inaccessible or has a sequence that remains resistant to modification, the large number of subtiligase-compatible sequence epitopes enables selection of a sequence that minimizes impact on protein expression level, solubility, and function. We anticipate that PILS could be applied to re-engineer the specificity of peptide ligases that harbor strict sequence requirements, such as sortase and butelase 1, further augmenting the toolbox of peptide ligases available to protein engineers and chemical biologists.

In combination with the versatile substrates that we have developed for one-step and modular protein modification, the mutants that we have engineered should be widely applicable for N-terminal modification with a variety of payloads. Furthermore, subtiligase exhibits broad specificity on the non-prime side and can be used with a wide array of user-designed substrates. Based on its broad sequence compatibility, site selectivity, fast reaction times, mild reaction conditions, and ease of use, we anticipate that subtiligase-catalyzed protein modification using fit-to-purpose mutants can be widely adopted to advance a variety of scientific fields.

Online Methods

Plasmid construction

Plasmids were constructed using standard Gibson cloning methods[47] with E. coli XL10 as the cloning host. PCR amplifications were performed using KOD Hot Start Polymerase (EMD Millipore) using the oligonucleotides listed in Supplementary Table 7. All plasmids were verified by Sanger sequencing (Quintara Biosciences).

pBS42-pre-pro-Subtiligase-His_6—A codon-optimized synthetic gene encoding pre-pro-subtiligase-His_6 was purchased from Integrated DNA Technologies. The DNA sequence and mature protein sequence are given in Supplementary Fig. 15. The synthetic gene was PCR amplified using primers Subtiligase F1 and Subtiligase R1 (Supplementary Table 7) and inserted between the EcoRI and BamHI sites of pBS42[48] using Gibson assembly. Plasmids encoding subtiligase variants were constructed using oligonucleotide primers encoding the desired mutation (Supplementary Table 7). Site-directed mutagenesis reactions[49] contained forward and reverse primers (0.5 μM each), pBS42-Subtiligase-His_6 template (100 ng), dNTPs (0.2 mM each), $MgSO_4$ (2.5 mM), KOD Hot Start DNA polymerase buffer (1×), and KOD Hot Start DNA polymerase (0.02 U/μL). The reaction mixture was subjected to the following thermocycling conditions: 95 °C for 2 min; 16 cycles of 95 °C for 20 s, 55 °C for

10 s, 72 °C for 3 min 30 s; a final extension at 72 °C for 7 min. Reaction mixtures were digested with DpnI (0.8 U/μL) for 1 h at 37°C and transformed into *E. coli* XL10.

pET28b-His6-Smt3-eGFP and variants. S. cerevisiae—Smt3 was amplified from pET28b-Smt3 (a gift from L. Pack) using primers pET28b SUMO NheI F1 and pET28b no linker GFP SUMO R1 and eGFP was amplified from pBH4-eGFP (a gift from S. Coyle) using primers pET28b no linker GFP F1 and pET28b GFP HindIII R1 (Supplementary Table 7). Both PCR products were inserted between the NheI and HindIII sites of pET28b using Gibson assembly. To construct vectors for expression of eGFP variants with N-terminal dipeptide extensions, Smt3 was amplified with the universal primer pET28b SUMO NheI F1 and the appropriate reverse primer listed in Supplementary Table 7, and by amplifying eGFP with the appropriate forward primer listed in Supplementary Table 7 and the universal primer pET28b GFP HindIII R1. PCR products were then inserted between the NheI and HindIII sites of pET28b.

Expression and purification of subtiligase and mutants

Subtiligase and variants were expressed as C-terminal His_6-tag fusions and secreted from *B. subtilis* BG2864. *E. coli* ER1821 were transformed with each subtiligase expression plasmid and concatameric DNA was prepared using a QIAprep Spin Miniprep Kit (Qiagen), omitting the Buffer PB wash. *B. subtilis* BG2864 were transformed with the concatameric DNA and grown on LB agar supplemented with 5 μg/mL chloramphenicol. 2×YT (5 mL) containing 12.5 μg/mL chloramphenicol was then inoculated with a single colony and grown overnight at 37°C with shaking at 200 rpm. 2×YT (50 mL) supplemented with 12.5 μg/mL chloramphenicol and 5 mM $CaCl_2$ was inoculated with the saturated overnight culture to an OD_{600} of 0.03 and grown in a baffled flask at 37°C with shaking at 200 rpm for 20–24 h. Cells were then removed by centrifugation at 4,000 × g for 15 min at 4°C. Secreted subtiligase was precipitated out of the media by addition of 3 volumes of cold ethanol and pelleted by centrifugation at 4,000 × g for 15 min at 4°C. Pellets were resuspended in 10 mL Ni-NTA wash buffer (50 mM sodium phosphate, pH 8.0, 300 mM NaCl, 20 mM imidazole) and insoluble material was removed by centrifugation at 4,000 × g for 15 min at 4°C. The supernatant was allowed to bind to HisPur Ni-NTA resin from a HisPur Ni-NTA spin column for 1 h at 4°C. The resin was collected by centrifugation at 500 × g for 5 min, resuspended in 400 μL Ni-NTA wash buffer, and loaded into the HisPur spin column. The column was washed with 4 × 400 μL Ni-NTA wash buffer by centrifugation at 700 × g for 2 min. Subtiligase variants were eluted with 3 × 400 μL Ni-NTA elution buffer (50 mM sodium phosphate, pH 8, 300 mM NaCl, 250 mM imidazole) and quantified by absorbance at 280 nm. HisPur spin columns were discarded after purification and a new spin column was used to purify each mutant to avoid the possibility of cross-contamination. The purified protein was buffer exchanged into 100 mM tricine, pH 8, 5 mM DTT, 10% glycerol by five cycles of 10-fold concentration and dilution in an Amicon Ultra centrifugal filter unit (0.5 mL, 3,000 MWCO). Single-use aliquots were flash frozen and stored at −80°C. Protein molecular weights were verified by LC-MS (Supplementary Table 2).

Preparation of proteome-derived peptide libraries

2×YT (50 mL) was inoculated with a single colony of *E. coli* XL10 and incubated overnight at 37°C with shaking at 200 rpm. Cells were harvested by centrifugation at 4,000 × g for 15 min at 4°C and resuspended in 50 mL of lysis buffer (10 mM HEPES, pH 7.5, 1 mM PMSF, 10 mM EDTA). Cells were lysed by three passes through a microfluidizer at 15,000 psi. Insoluble material was removed by centrifugation at 10,000 × g for 20 min at 4°C. DNA was precipitated by dropwise addition of 10% (w/v) streptomycin sulfate to a final concentration of 1% (w/v) and removed by centrifugation at 10,000 × g for 20 min at 4°C. Protein concentration was determined by BCA assay and subsequent steps were carried out on a total of 10 mg of protein at 2 mg/mL. The lysate was adjusted to 100 mM HEPES, pH 7.5 and DTT (1 M) was added to 5 mM. Following a 1 h incubation at room temperature, iodoacetamide (500 mM) was added to 10 mM and the sample was incubated in the dark for 1 h at 37°C. Protein was precipitated by addition of 15% (w/v) trichloroacetic acid (TCA) followed by an overnight incubation at −20°C. The sample was centrifuged at 20,000 × g for 10 min and the pellet was washed twice with ice-cold methanol. The pellet was solubilized by ultrasonication in 5 mL 20 mM NaOH (20% amplitude, 5 s on/1 s off) and adjusted to 200 mM HEPES, pH 7.5. Insoluble material was removed by centrifugation at 20,000 × g for 20 min at 4°C. The protein concentration of the supernatant was determined by BCA assay and protein was digested overnight at 37°C with a 1:100 (w/w) ratio of mass-spectrometry grade trypsin or Glu-C. After digestion, 1 mM PMSF (for trypsin) or 1 mM PMSF and 0.5 mM diisopropylfluorophosphate (for Glu-C) was added to inhibit the digest protease. Reduction and alkylation were repeated and peptide libraries were purified by C18 solid-phase extraction and eluted in 80% acetonitrile/20% water. Libraries were concentrated in a vacuum centrifuge and diluted three times with water to remove acetonitrile, diluted to 2 mg/mL in water, and stored at −80°C until further use.

Estimation of peptide library molarity

Libraries were stored as 2 mg/mL stock solutions by determining the concentration with a bicinchoninic acid (BCA) assay. The molarity of trypsin and Glu-C peptide libraries was estimated by taking into account the average length of an *E. coli* protein (300 amino acids (a. a.))[50], the average length of a tryptic peptide (10 a. a.)[51], and the average molecular weight (MW) of an amino acid (110 Da). Calculations are shown below.

$$\text{Average protein MW} = 300\,\text{a.a.} \times \frac{110\ \text{Da}}{\text{a.a.}} = 33,000\ \text{Da}$$

$$\text{Average number of peptides per protein} = 300\,\text{a.a.} \times \frac{1\ \text{peptide}}{10\,\text{a.a}} = 30\ \text{peptides}$$

$$\text{Peptide library molarity} = \frac{2\ \text{mg}}{\text{mL}} \times \frac{1\ \text{mmol protein}}{33,000\ \text{mg}} \times \frac{30\ \text{peptides}}{\text{protein}} \times \frac{1000\ \text{mL}}{\text{L}} = 1.8\ \text{mM peptide}$$

Peptide synthesis

Peptides were synthesized using fluorenylmethyloxycarbonyl (Fmoc) chemistry on Rink Amide AM resin (EMD Millipore)[52]. The following side chain protecting groups were used: Arg(Pbf), Gln(Trt), Tyr(tBu), Asn(Trt), Glu(OtBu). Coupling reactions were performed using 5 equiv. of the appropriate Fmoc amino acid, 5 equiv. of diisopropylcarbodiimide (DIC), and 5 equiv. of 1-hydroxy-benzotriazole (HOBt) in *N, N*-dimethylformamide (DMF) for 1 h at room temperature, except where noted. Fmoc groups were deprotected using a 30 min incubation in 20% (v/v) 4-methylpiperadine in DMF. The glycolic acid moiety was incorporated by coupling the amine of the resin-bound peptide to acetoxyacetic acid (5 equiv.) in the presence of DIC (5 equiv.) and HOBt (5 equiv.) for 1 h, followed by deprotection with 2.5 M hydrazine monohydrate in DMF for 16 h. The amino acid immediately N-terminal to the glycolic acid group (5 equiv.) was coupled to the peptide in the presence of 1 M DIC and 1 mol % *N, N*-dimethylaminopyridine for 1 h[53]. Biotin (5 equiv.) was dissolved in dimethylsulfoxide (DMSO) and coupled to the peptide using 5 equiv. DIC and 5 equiv. HOBt. Peptides were cleaved and side chains were deprotected by incubating the resin with 95:2.5:2.5 ratio of trifluoroacetic acid (TFA), water, and triisopropylsilane (TIPS). The solution was concentrated to 5 mL on a rotary evaporator and the peptide was precipitated by addition of 9 volumes of diethyl ether and washed twice with diethyl ether. Peptides were purified by C18 reverse-phase HPLC using a gradient from 0.1% TFA in water to 0.1% TFA in acetonitrile. Acetonitrile was removed using a vacuum centrifuge and peptides were lyophilized. Lyophilized peptides were dissolved in DMSO and stored at −80°C until use. LC-MS characterization data for each peptide is shown in Supplementary Fig. 16.

Subtiligase specificity profiling using Proteomic Identification of Ligation Sites (PILS)

Specificity profiling reactions were initiated by addition of subtiligase or variant (1 μM) to a reaction mixture containing peptide library (1 mM), biotinylated peptide ester **1** (0.2 mM), and 100 mM tricine, pH 8.0. After 1 h, reactions were quenched by addition of 1 volume of 8 M guanidine hydrochloride. Biotinylated peptides were enriched on High-Capacity Neutravidin resin (Thermo Fisher Scientific) (0.25 mL of 50% resin slurry). The resin was washed five times with 0.5 mL 4 M guanidine hydrochloride and five times with TEV elution buffer (100 mM ammonium bicarbonate, 2 mM DTT). The resin was resuspended in 0.25 mL TEV elution buffer and incubated with TEV protease (10 μg) for 2 h to selectively elute biotinylated peptides. Resin was removed from the eluted peptides using a spin filter. The solution containing the eluted peptides was adjusted to 5% TFA, incubated at room temperature for 10 min, and spun at 20,000 × g to remove precipitated TEV protease. Peptides were then desalted on C18 OMIX tips, dried, dissolved in 10 μL 0.1% formic acid, and analyzed by LC-MS/MS.

LC-MS/MS data collection

LC-MS/MS analysis was performed on an Acclaim PepMap RSLC column (75 μm × 15 cm, 2 μm particle size, 100 Å pore size, Thermo Scientific) using a Thermo Dionex UltiMate 3000 RSLCnano liquid chromatography system coupled to a Thermo Q-Exactive Plus hybrid quadrupole-Orbitrap mass spectrometer. Mobile phase A was 0.1% formic acid and

mobile phase B was 0.1% formic acid, 80% acetonitrile. Samples (5 μL) were loaded over 15 min at 0.5 μL/min in mobile phase A and peptides were eluted at 0.3 μL/min with a linear gradient from mobile phase A to 40% mobile phase B over either 30 min (for PILS experiments) or 125 min (for N terminomics experiments). Data-dependent acquisition of MS data was performed using Thermo Xcalibur software scanning a mass range from 300–1,500 m/z.

Mass spectrometry data analysis

Peak lists from Thermo RAW files were generated using MSConvert (Proteowizard). Peptides were identified from the *E. coli* or human SwissProt database using Protein Prospector (UCSF) with a false discovery rate of <1%. The parent ion tolerance was set at 6 ppm and the fragment ion tolerance was set at 20 ppm and two missed cleavages were allowed. Search parameters included carbamidomethylation at Cys as a constant modification and aminobutyric acid (Abu) at peptide N termini, acetylation at protein N termini, oxidation at Met, pyroglutamate formation at N-terminal Gln, and Met excision at protein N termini as variable modifications. Trypsin specificity was defined to include cleavage C-terminal to Arg or Lys, and Glu-C specificity was defined to include cleavage C-terminal to Glu or Asp. For analysis of PILS data, the appropriate specificity was required at both the N- and C-terminal ends of the peptide. For analysis of N terminomics datasets, the appropriate specificity was required at only the C terminus of the peptide to enable identification of protease cleavage events of different specificity. For reference trypsin and Glu-C datasets used in PILS analysis, data were analyzed similarly, omitting the Abu variable modification.

PILS specificity data analysis

PILS analysis was implemented using custom Python scripts (Supplementary Dataset 78). Lists of identified peptides were filtered for bona fide subtiligase substrates based on the presence of an Abu modification at the peptide N terminus. Peptides from trypsin and Glu-C datasets were combined and the frequency with which each amino acid appeared in each position was compared to the frequency in the combined trypsin and Glu-C reference sets. An enrichment score (z) was calculated according to the following formula:

$$z = \frac{X - \mu}{\sigma}$$

where X is the frequency of the amino acid in the enriched, Abu-tagged sample, μ is the frequency of the amino acid in the reference sample, and σ is the standard deviation. A positive enrichment score indicates that an amino acid is enriched compared to the input libraries, while a negative enrichment score indicates that an amino acid is de-enriched compared to the input libraries. For analysis of dipeptide sequences, the same approach was used, except the frequency of the dipeptide sequence at the N termini of peptides in the sample and reference sets was compared. The tryptic reference set contained 5,720 peptides and the Glu-C reference set contained 4,278 peptides (Supplementary Dataset 77). Individual enriched datasets (listed in Supplementary Table 8) generally contained 1,000–4,000 peptides (exact numbers of peptides identified in each experiment are given in

Supplementary Table 8, Supplementary Datasets 1–73). Hierarchical clustering of enrichment scores was performed using the 'heatmap' function in R (www.r-project.org).

Kinetic analysis of subtiligase mutants

The peptide ligation activity of subtiligase was measured using the FRET-based assay shown schematically in Supplementary Figure 4a. Assays were performed in 96-well plates in 200 μL total volume containing 100 mM tricine, pH 8.0, 5 mM DTT, 20 μM Pacific Blue-GAAPF-glc-RK(Dabcyl) (a subtiligase ester substrate) and 0, 6.25, 12.5, 25, 50, 100, or 200 μM AFAK(FAM). Reactions were initiated by the addition of subtiligase or subtiligase variant to a final concentration of 25 nM. Fluorescence was monitored over time in a Molecular Devices SpectraMax M5 plate reader with excitation at 405 nm and emission at 450 nm (hydrolysis product) and 520 nm (ligation product). A standard curve of Pacific Blue-GAAPFAFAK(FAM) was constructed to correlate fluorescence intensity with ligation product concentration. A plot of AFAK(FAM) concentration vs. observed rate was fit to a line to determine the relative k_{cat}/K_M for each mutant enzyme.

Mammalian cell culture

Cell lines used in this study were tested annually for mycoplasma contamination. Jurkat E6.1 (a gift from K. Roybal, Lim lab, UCSF) cells were grown in RPMI-1640 media supplemented with 10% fetal bovine serum, 2 mM l-glutamine, and 1% penicillin-streptomycin to a density of 1×10^6 cells per mL. The day before harvest, cells were split by two-fold and treated with either 50 μM etoposide for 12 h or an equal volume of DMSO. Cell death was assessed using the CellTiterGlo assay (Promega) according to the manufacturer's instructions. Cells were harvested at 300 × g for 5 min, washed twice with PBS, and stored at −80°C until use.

N terminomics analysis in *E. coli* and Jurkat cell lysate

E. coli XL10 were lysed by three passes through a microfluidizer at 15,000 psi in 100 mM tricine, pH 8, 150 mM NaCl, 100 μM PMSF, 100 μM AEBSF, 2.5 mM EDTA. Insoluble material was removed by centrifugation at 20,000 × g for 20 min at 4°C. Biotinylated subtiligase substrate **1** was added to the supernatant at a final concentration of 2.5 mM. The reaction was initiated by addition of the appropriate subtiligase variant (1 μM) and allowed to proceed for 1 h at room temperature on an end-over-end mixer. After labeling, biotinylated N-terminal peptides were enriched as described previously[28,29,42] and analyzed by LC-MS/MS. Two cell culture replicates were performed for all experiments.

For N terminomics studies of Jurkat lysate, cells were lysed by ultrasonication (20% amplitude, 5 s/1 s on/off) in 400 mM tricine, pH 8, 4% (w/v) SDS, 100 μM PMSF, 100 μM AEBSF, 2.5 mM EDTA. Insoluble material was removed by centrifugation at 20,000 × g for 20 min at room temperature. The sample was reduced by boiling for 15 min in the presence of 5 mM TCEP and alkylated by 1 h incubation at room temperature in the presence of 10 mM iodoacetamide. DTT (25 mM) was added to quench the remaining iodoacetamide and Triton X-100 was added to a final concentration of 2.5% (v/v). The sample was diluted four-fold with water and biotinylated subtiligase substrate **1** was added to a final concentration of 2.5 mM. The reaction was initiated by addition of stabiligase or the stabiligase cocktail (4

μM) and allowed to proceed at room temperature for 1 h. After labeling, biotinylated N-terminal peptides were enriched as described previously[28,29,54]. Two cell culture replicates were performed for all experiments.

Purification of GFP variants for protein bioconjugation

GFP and N-terminal variants were expressed as His_6-SUMO tag fusion proteins and purified using Ni-NTA affinity chromatography. After affinity purification, the His_6-SUMO tag was cleaved using Senp1 protease as previously described[55] and the cleaved His_6-SUMO was removed using Ni-NTA affinity chromatography.

Purification of recombinant antibodies

Recombinant antibodies were expressed in *E. coli* from a single vector with the light chain fused to a PelB leader sequence and the heavy chain fused to an STII signal sequence for secretion to the periplasm[38]. Cells were lysed with B-PER (ThermoFisher Scientific) and the lysate was heated to 60°C for 20 min. After removal of insoluble material by centrifugation, antibodies were purified on a HiTrap Protein A sepharose column and exchanged into PBS for storage.

Protein bioconjugation

Purified protein was diluted to 50 μM in 100 mM tricine, pH 8.0 containing 5 mM of the subtiligase substrate to be conjugated. The reaction was initiated by addition of subtiligase or the appropriate variant (1 μM) and allowed to proceed for 1 h at room temperature. Protein was then exchanged into PBS using a 0.5 mL Zeba desalting spin column (ThermoFisher Scientific) and the completeness of the reaction was analyzed on a Xevo G2-XS mass spectrometer equipped with a LockSpray (ESI) source and Acquity Protein BEH C4 column (2.1 mm inner diameter, 50 mm length, 300 Å pore size, 1.7 μm particle size) connected to an Acquity I-class liquid chromatography system (Waters). Deconvolution of mass spectra was performed using the maximum entropy (MaxEnt) algorithm in MassLynx 4.1 (Waters). All mass spectra shown are representative of at least two independent bioconjugation experiments.

Measurement of rAb affinities

The affinities of αGFP rAbs for GFP were measured using biolayer interferometry on an Octet RED 384 system (ForteBio). The modified or unmodified αGFP rAbs were diluted to 300 nM in PBS containing 0.05% Tween 20, 0.2% BSA, and 10 μM biotin. The rAbs were immobilized on 2nd Generation Dip and Read Anti-Human-Fab-CH1 sensor tips (ForteBio). Binding of GFP to the immobilized rAbs was assessed by loading serial dilutions of recombinant GFP onto the sensors. Results were fit using the Data Analysis 9.0 software provided with the Octet RED 384 to determine dissociation constants (K_Ds).

Modification of peptide 3 with biotin

Peptide **3** (100 mM in DMF, 20 μL) was mixed with EZ-Link NHS-biotin (110 mM in DMF, 20 μL) and incubated for 1 h at room temperature. Excess NHS-biotin was quenched by

addition of 20 μL of water followed by an overnight incubation at room temperature. The reaction mixture was used without further purification in protein bioconjugation reactions.

Modification of peptide 2-modified proteins with DBCO reagents

Following protein bioconjugation with peptide **2**, proteins were desalted into PBS three times using 0.5 mL Zeba desalting spin columns (ThermoFisher Scientific). Proteins (50 μM) were then modified with the appropriate DBCO reagent (130 μM) by incubating for 2–16 h at room temperature. Excess DBCO reagent was removed by exchanging into PBS using a 0.5 mL Zeba desalting spin column.

GFP-αGFP co-localization experiments

HEK293T cells modified with a doxycycline-inducible cell surface GFP expression system were plated at 10,000 cells per well in a 96-well flat-bottom tissue culture plate. GFP expression was induced at 50% confluency by addition of 1 μg/mL doxycycline to the culture medium, or an equal volume of water as a negative control. After 18 h, cells were washed three time with PBS containing 3% BSA and stained with 0.1 μg/mL αGFP-rAb in PBS + 3% BSA for 30 min at room temperature. Cells were washed three times with PBS + 3% BSA and imaged using a Zeiss AxioObserver Z1 inverted fluorescence microscope. Fluorescence microscopy images shown in Figure 3c are representative of experiments performed for three cell culture replicates.

Statistical analysis

Enrichment scores for PILS experiments were calculated using custom Python scripts (Supplementary Data 78) and correspond to the standard score (z-score) comparing the enriched, subtiligase-labeled population to the input peptide library. The exact numbers of peptides analyzed in each experiment and replicate are listed in Supplementary Table 8. Other statistical tests were performed using Prism 6 (GraphPad). For comparison of two samples, p-values were calculated using a two-tailed, unpaired t-test. P-values were corrected for multiple comparisons using the Holm-Sidak method available in the Prism 6 software. Unless otherwise indicated, two cell culture replicates of all mass spectrometry experiments were performed. Error bars indicate mean ± standard deviation. Fluorescence microscopy images are representative of experiments performed for three cell culture replicates. All mass spectra shown are representative of at least two independent bioconjugation experiments.

Computer code availability

Python and R scripts used for data analysis are included as Supplementary Dataset 78. Additionally, the ALPINE web application (https://wellslab.ucsf.edu/alpine) includes web interfaces for many of these scripts.

Data availability

All data generated or analyzed for this study are available within the paper and its associated supplementary information files, or from the corresponding author upon reasonable request.

Additionally, raw mass spectrometry data and search results have been deposited in the Proteome×change repository under the accession numbers listed in Supplementary Table 8.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank S. Coyle, Z. Hill, M. Hornsby, H. Huang, O. Julien, P. Lee, D. Sashital, L. Pack, A. Stewart, H. Tran, K. Wypysniak, and members of the Wells laboratory for helpful discussions. We thank P. Lee and M. Hornsby (Department of Pharmaceutical Chemistry, University of California, San Francisco) for the αGFP rAb expression vector, S. Pollock (Department of Pharmaceutical Chemistry, University of California, San Francisco) for the HEK-293T-GFP cell line, and H. Tran (Department of Pharmaceutical Chemistry, University of California, San Francisco) for the subtiligase *E. coli* expression vector. This work was supported by NIH grant 5R01GM081051-09 and the Harry and Dianna Hind Professorship in Pharmaceutical Sciences (to J.A.W.). A.M.W. is a Merck Fellow of the Helen Hay Whitney Foundation (F-1112).

References

1. Liu Y, Patricelli MP, Cravatt BF. Activity-based protein profiling: the serine hydrolases. Proc. Natl. Acad. Sci. U.S. A. 1999; 96:14694–14699. [PubMed: 10611275]
2. Banghart M, Borges K, Isacoff E, Trauner D, Kramer RH. Light-activated ion channels for remote control of neuronal firing. Nat. Neurosci. 2004; 7:1381–1386. [PubMed: 15558062]
3. Weerapana E, et al. Quantitative reactivity profiling predicts functional cysteines in proteomes. Nature. 2010; 468:790–795. [PubMed: 21085121]
4. Erlanson DA, Wells JA, Braisted AC. Tethering: fragment-based drug discovery. Annu. Rev. Biophys. Biomol. Struct. 2004; 33:199–223. [PubMed: 15139811]
5. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature. 2013; 503:548–551. [PubMed: 24256730]
6. Serafimova IM, et al. Reversible targeting of noncatalytic cysteines with chemically tuned electrophiles. Nat. Chem. Biol. 2012; 8:471–476. [PubMed: 22466421]
7. Junutula JR, et al. Site-specific conjugation of a cytotoxic drug to an antibody improves the therapeutic index. Nat. Biotechnol. 2008; 26:925–932. [PubMed: 18641636]
8. Lyon RP, et al. Self-hydrolyzing maleimides improve the stability and pharmacological properties of antibody-drug conjugates. Nat. Biotechnol. 2014; 32:1059–1062. [PubMed: 25194818]
9. Pleiner T, et al. Nanobodies: site-specific labeling for super-resolution imaging, rapid epitope-mapping and native protein complex isolation. Elife. 2015; 4:e11349. [PubMed: 26633879]
10. Lin S, et al. Redox-based reagents for chemoselective methionine bioconjugation. Science. 2017; 355:597–602. [PubMed: 28183972]
11. Guimaraes CP, et al. Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. Nat. Protoc. 2013; 8:1787–1799. [PubMed: 23989673]
12. Fernández-Suárez M, et al. Redirecting lipoic acid ligase for cell surface protein labeling with small-molecule probes. Nat. Biotechnol. 2007; 25:1483–1487. [PubMed: 18059260]
13. Wu P, et al. Site-specific chemical modification of recombinant proteins produced in mammalian cells by using the genetically encoded aldehyde tag. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:3000–3005. [PubMed: 19202059]
14. Hooker JM, Esser-Kahn AP, Francis MB. Modification of aniline containing proteins using an oxidative coupling strategy. J. Am. Chem. Soc. 2006; 128:15558–15559. [PubMed: 17147343]
15. Lang K, et al. Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. Nat. Chem. 2012; 4:298–304. [PubMed: 22437715]
16. Dawson P, Muir T, Clark-Lewis I, Kent S. Synthesis of proteins by native chemical ligation. Science. 1994; 266:776–779. [PubMed: 7973629]

17. Muir TW, Sondhi D, Cole PA. Expressed protein ligation: a general method for protein engineering. Proc. Natl. Acad. Sci. U.S.A. 1998; 95:6705–6710. [PubMed: 9618476]
18. Henager SH, et al. Enzyme-catalyzed expressed protein ligation. Nat. Methods. 2016; 13:925–927. [PubMed: 27669326]
19. Jacob E, Unger R. A tale of two tails: why are terminal residues of proteins exposed? Bioinformatics. 2007; 23:e225–e230. [PubMed: 17237096]
20. Rosen CB, Francis MB. Targeting the N terminus for site-selective protein modification. Nat. Chem. Biol. 2017; 13:697–705. [PubMed: 28632705]
21. MacDonald JI, Munch HK, Moore T, Francis MB. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. Nat. Chem. Biol. 2015; 11:326–331. [PubMed: 25822913]
22. Nguyen GKT, et al. Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. Nat. Chem. Biol. 2014; 10:732–738. [PubMed: 25038786]
23. Abrahmsen L, et al. Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. Biochemistry. 1991; 30:4151–4159. [PubMed: 2021606]
24. Frankel BA, Kruger RG, Robinson DE, Kelleher NL, McCafferty DG. *Staphylococcus aureus* sortase transpeptidase SrtA: insight into the kinetic mechanism and evidence for a reverse protonation catalytic mechanism. Biochemistry. 2005; 44:11188–11200. [PubMed: 16101303]
25. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Comm. 1967; 425:497–502.
26. Schilling O, Overall CM. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat. Biotechnol. 2008; 26:685–694. [PubMed: 18500335]
27. Renicke C, Spadaccini R, Taxis C. A tobacco etch virus protease with increased substrate tolerance at the P1' position. PLoS ONE. 2013; 8:e67915. [PubMed: 23826349]
28. Mahrus S, et al. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008; 134:866–876. [PubMed: 18722006]
29. Shimbo K, et al. Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. Proc. Natl. Acad. Sci. U.S.A. 2012; 109:12432–12437. [PubMed: 22802652]
30. Chang TK, Jackson DY, Burnier JP, Wells JA. Subtiligase: a tool for semisynthesis of proteins. Proc. Natl. Acad. Sci. U.S.A. 1994; 91:12544–12548. [PubMed: 7809074]
31. Schilling O, Overall CM. Proteomic discovery of protease substrates. Curr. Opin. Chem. Biol. 2007; 11:36–45. [PubMed: 17194619]
32. Wells JA, Powers DB, Bott RR, Graycar TP, Estell DA. Designing substrate specificity by protein engineering of electrostatic interactions. Proc. Natl. Acad. Sci. U.S.A. 1987; 84:1219–1223. [PubMed: 3547407]
33. Estell DA, et al. Probing steric and hydrophobic effects on enzyme-substrate interactions by protein engineering. Science. 1986; 233:659–663. [PubMed: 17835820]
34. Takeuchi Y, et al. Molecular recognition at the active site of subtilisin BPN': crystallographic studies using genetically engineered proteinaceous inhibitor SSI (*Streptomyces* subtilisin inhibitor). Protein Eng. 1991; 4:501–508. [PubMed: 1891457]
35. Hedstrom L, Szilagyi L, Rutter WJ. Converting trypsin to chymotrypsin: the role of surface loops. Science. 1992; 255:1249–1253. [PubMed: 1546324]
36. Estell DA, Graycar TP, Wells JA. Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. J. Biol. Chem. 1985; 260:6518–6521. [PubMed: 3922976]
37. Nuijens T, et al. Engineering a diverse ligase toolbox for peptide segment condensation. Adv. Synth. Catal. 2016; 358:4041–4048.
38. Hornsby M, et al. A high through-put platform for recombinant antibodies to folded proteins. Mol. Cell. Proteomics. 2015; 14:2833–2847. [PubMed: 26290498]
39. Fellouse FA, et al. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. J. Mol. Biol. 2007; 373:924–940. [PubMed: 17825836]
40. Pan Y, et al. Determination of equilibrium dissociation constants for recombinant antibodies by high-throughput affinity electrophoresis. Sci. Rep. 2016; 6:27. [PubMed: 28442707]

41. Saxon E, Bertozzi CR. Cell surface engineering by a modified Staudinger reaction. Science. 2000; 287:2007–2010. [PubMed: 10720325]
42. Wiita AP, Hsu GW, Lu CM, Esensten JH, Wells JA. Circulating proteolytic signatures of chemotherapy-induced cell death in humans discovered by N-terminal labeling. Proc. Natl. Acad. Sci. U.S.A. 2014; 111:7594–7599. [PubMed: 24821784]
43. Crawford ED, Wells JA. Caspase substrates and cellular remodeling. Annu. Rev. Biochem. 2011; 80:1055–1087. [PubMed: 21456965]
44. Xiao Q, Zhang F, Nacev BA, Liu JO, Pei D. Protein N-terminal processing: substrate specificity of *Escherichia coli* d human methionine aminopeptidases. Biochemistry. 2010; 49:5588–5599. [PubMed: 20521764]
45. Voss M, Schröder B, Fluhrer R. Mechanism, specificity, and physiology of signal peptide peptidase (SPP) and SPP-like proteases. Biochim. Biophys. Acta. 2013; 1828:2828–2839. [PubMed: 24099004]
46. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45:D158–D169. [PubMed: 27899622]

Methods References

47. Gibson DG, et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods. 2009; 6:343–345. [PubMed: 19363495]
48. Wells JA, Ferrari E, Henner DJ, Estell DA, Chen EY. Cloning, sequencing, and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*. Nucleic Acids Res. 1983; 11:7911–7925. [PubMed: 6316278]
49. Zheng L, Baumann U, Reymond J-L. An efficient one-step site-directed and site-saturation mutagenesis protocol. Nucleic Acids Res. 2004; 32:e115–e115. [PubMed: 15304544]
50. Milo, R., Phillips, R. "How Big Is the "Average" Protein?". Cell Biology by the Numbers. http://book.bionumbers.org/how-big-is-the-average-protein/
51. Giansanti P, Tsiatsiani L, Low TY, Heck AJR. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. Nat. Protoc. 2016; 11:993–1006. [PubMed: 27123950]
52. Coin I, Beyermann M, Bienert M. Solid-phase peptide synthesis: from standard procedures to the synthesis of difficult sequences. Nat. Protoc. 2007; 2:3247–3256. [PubMed: 18079725]
53. Braisted AC, Judice JK, Wells JA. Synthesis of proteins by subtiligase. Methods Enzymol. 1997; 289:298–313. [PubMed: 9353727]
54. Wiita AP, Seaman JE, Wells JA. Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini. Methods Enzymol. 2014; 544:327–358. [PubMed: 24974296]
55. Reverter D, Lima CD. Preparation of SUMO proteases and kinetic analysis using endogenous substrates. Methods Mol. Biol. 2009; 497:225–239. [PubMed: 19107421]

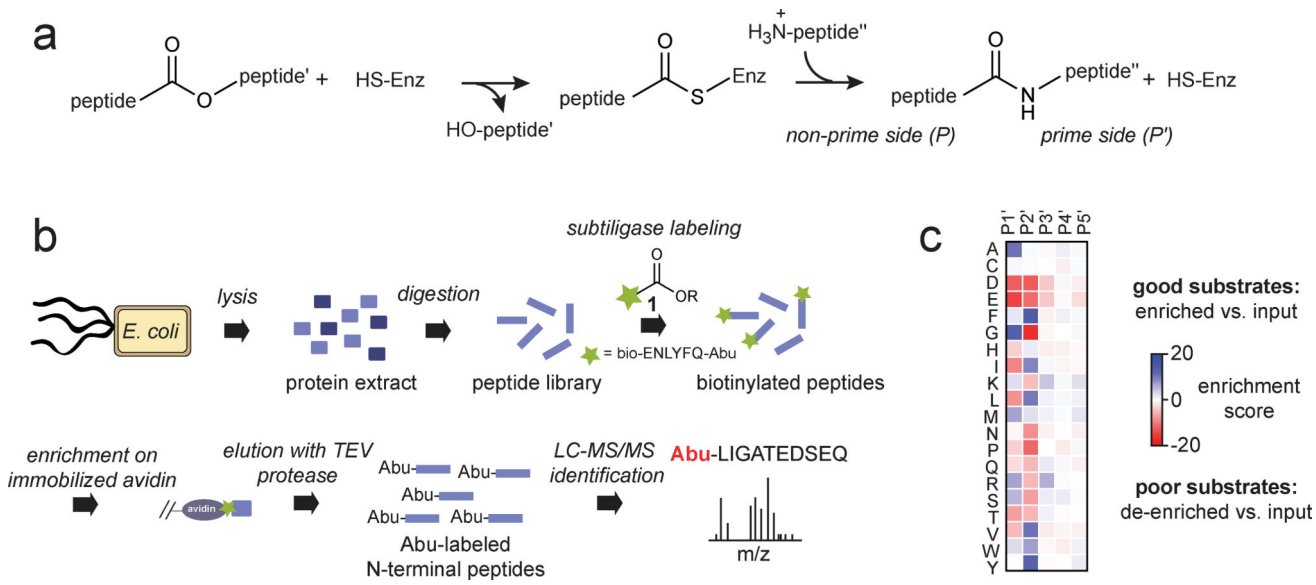

Figure 1. Proteomic identification of ligation sites (PILS) applied to comprehensive characterization of subtiligase prime-side specificity

(**a**) The ligation reaction catalyzed by subtiligase accepts a peptide ester substrate, forms a thioester intermediate, and then transfers the peptide to an α-amine-containing acceptor peptide. (**b**) A schematic representation of the PILS strategy for comprehensive characterization of prime-side subtiligase specificity. Proteome-derived peptide libraries are generated by protease digestion of *E. coli* protein extract. The peptide libraries are used as substrates for modification by subtiligase and biotinylated peptide ester **1** (biotin-EEENLYFQ-Abu-glycolate-R). Biotinylated peptides are enriched on immobilized neutravidin and selectively eluted by cleavage with TEV protease, leaving an Abu mass tag on the N termini of subtiligase substrates for positive identification. (**c**) Heatmap showing positional enrichment or de-enrichment of each amino acid at P1'–P5' compared to the input peptide libraries.

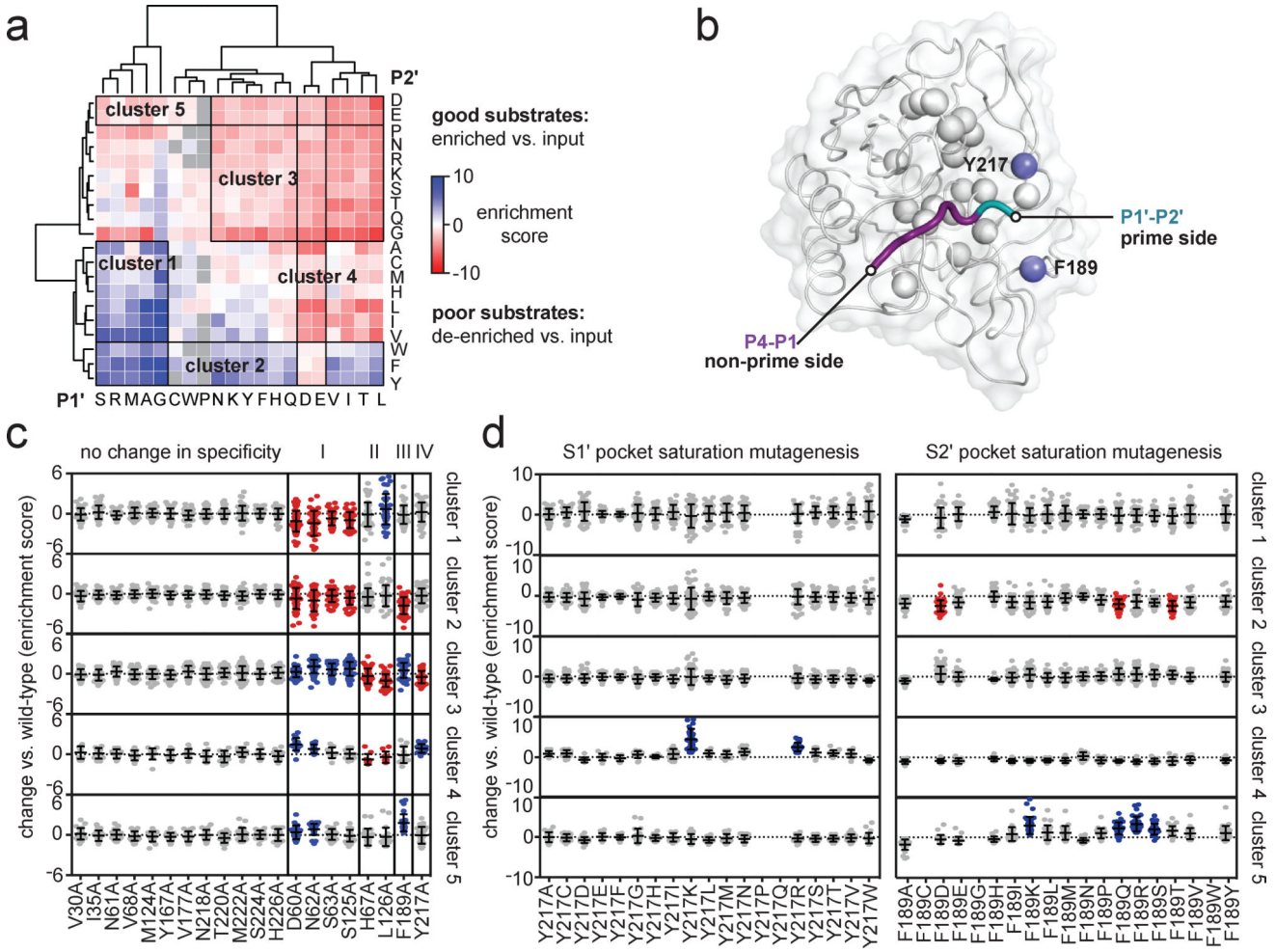

Figure 2. Defining and re-engineering subtiligase specificity for P1' and P2' residues

(**a**) PILS specificity map for the 400 possible N-terminal dipeptide sequences. Sequences that were not observed in the input library (18/400; 4.5%) are colored in grey. (**b**) Structure of the subtilisin–SSI complex showing the P4–P1 residues of SSI in purple and the P1'–P2' residues of SSI in teal. Sites targeted for alanine-scanning mutagenesis within 7 Å of the catalytic triad are shown as grey spheres. Sites targeted for subsequent saturation mutagenesis are shown as light blue spheres. (c) Results of alanine-scanning mutagenesis. Each dot represents one dipeptide sequence. Dots are colored grey if there was no change in the mean enrichment score for the cluster; red if the sequences in the cluster were worse substrates for the mutant than the wild-type enzyme with a change in mean enrichment score of >2; and blue if the sequences in the cluster were better substrates for the mutant compared to the wild-type enzyme with a change in mean enrichment score of >2. Black bars show the mean and standard deviation of enrichment scores for each cluster. (**d**) Results of saturation mutagenesis studies. Left, position 217 (S1' pocket); right, position 189 (S2' pocket). Colors, same as in **c**.

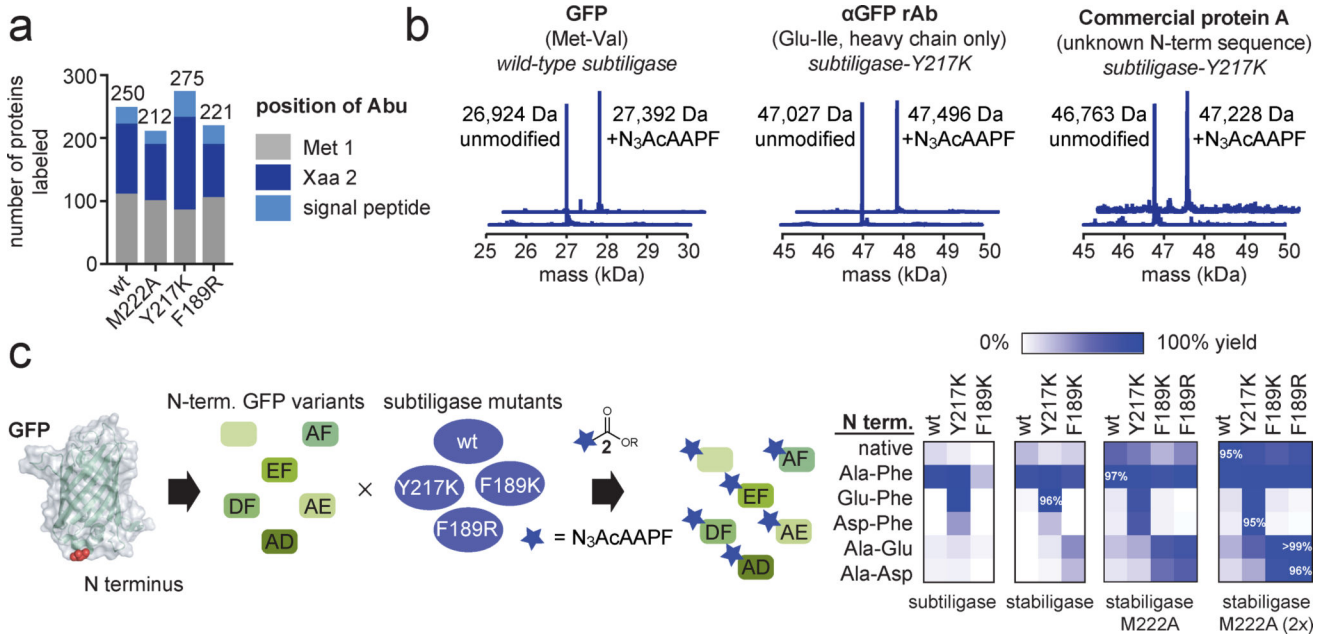

Figure 3. Scope of subtiligase-catalyzed N-terminal modification of folded proteins

(**a**) Native proteins from *E. coli* lysate labeled at translational N termini (initiator Met or residue 2 if the initiator Met is removed) and annotated signal peptide cleavage sites. (**b**) ESI mass spectra showing modification of GFP, a recombinant antibody, and commercial protein A by subtiligase or variants. (**c**) Ligation of azide-bearing peptide **2** onto GFP containing different N-terminal sequences tested with optimal subtiligase mutants. The heatmap shows the bioconjugation yield given by each mutant for each subtiligase variant tested. Sequence context of the mutant is shown at the bottom of the heatmap. ‘2x’ indicates that the labeling procedure was carried out a second time following desalting of the reaction mixture. The numerical value for the highest yield achieved for a particular N-terminal sequence is indicated in white.

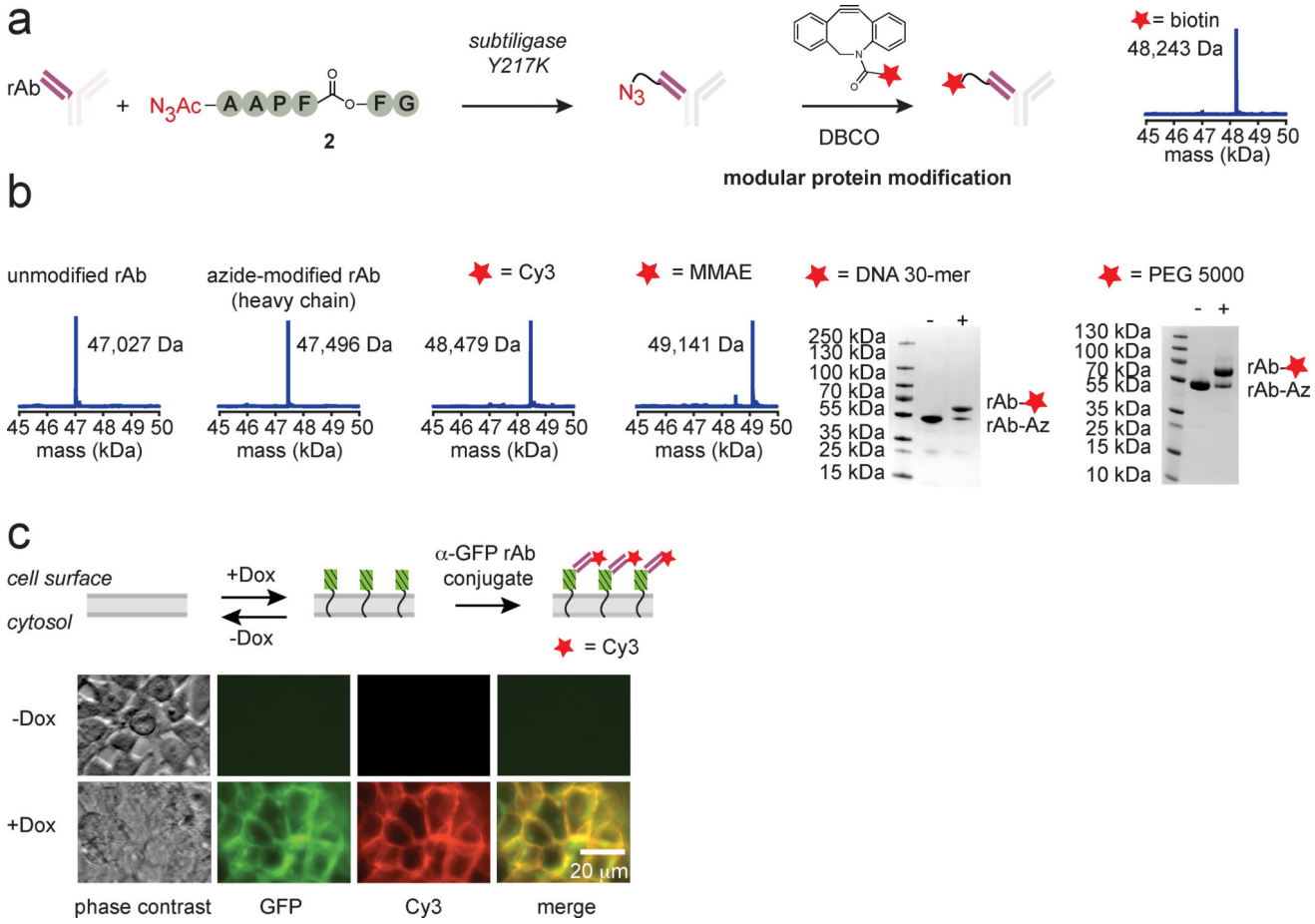

Figure 4. Modular strategy for subtiligase-catalyzed protein bioconjugation

(**a**) Azide-bearing peptide ester **3** reacts with commercially available dibenzocyclooctynes (DBCOs), providing a convenient route for modular protein labeling. (**b**) ESI mass spectra or SDS-PAGE gels for an anti-GFP rAb (αGFP) modified with a variety of different payloads using DBCO chemistry. MMAE, monomethyl auristatin E. (**c**) Cy3-α-GFP rAb staining of a HEK293T cell line modified for doxycycline-inducible expression of cell surface GFP.

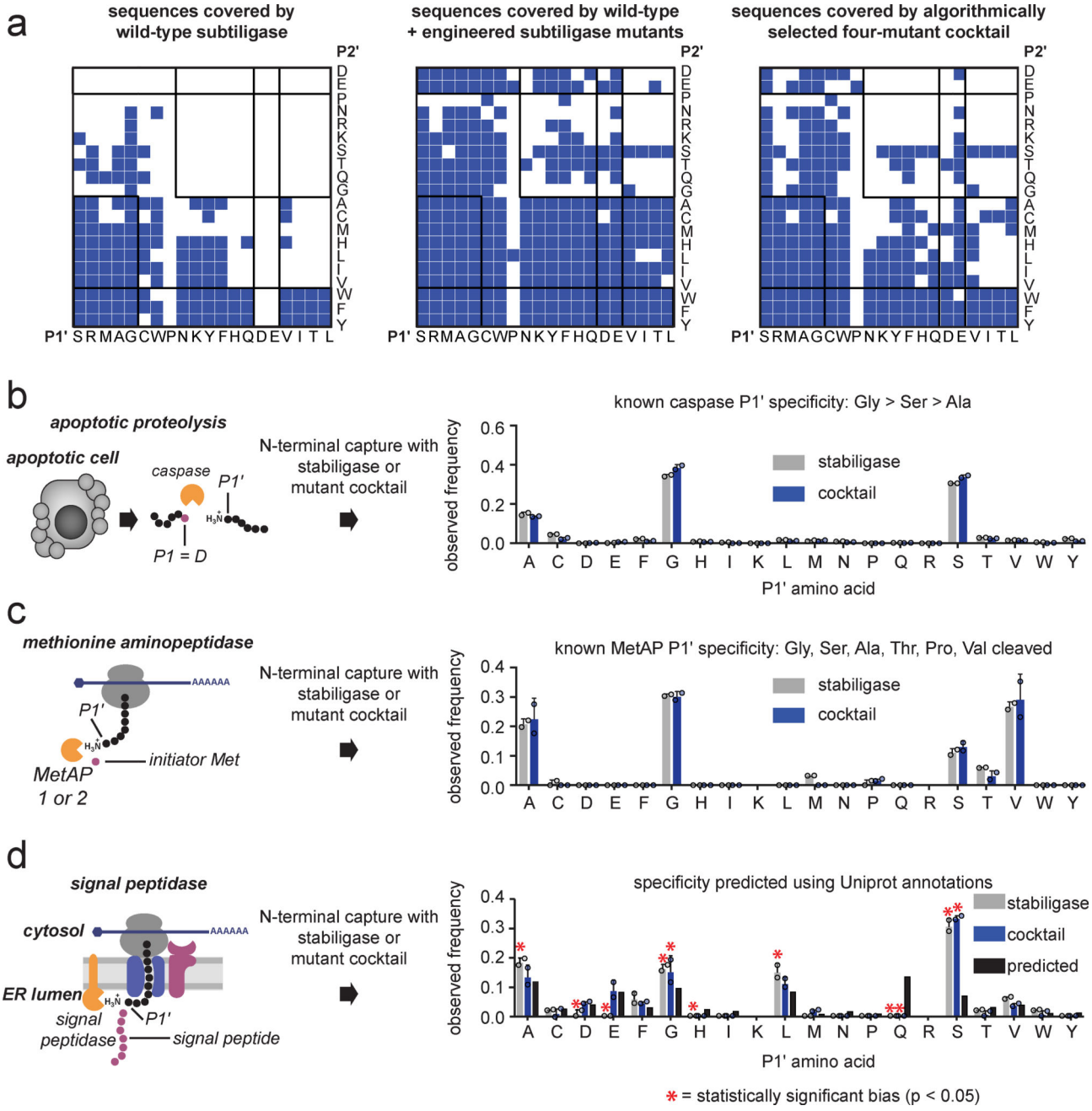

Figure 5. Algorithmically selected subtiligase cocktails for cellular N terminomics

(**a**) Sequences covered by wild-type subtiligase (left), the 72 subtiligase variants that we characterized (center), and an algorithmically selected four-mutant cocktail (right). (**b**) Frequency of amino acids at the P1' position of apoptotic protease substrates (P1 = D) captured by stabiligase (757 P1 = D peptides in replicate 1, 896 in replicate 2) or the stabiligase cocktail (754 P1 = D peptides in replicate 1, 775 in replicate 2). (**c**) Frequency of amino acids at the P1' position of native proteins cleaved by methionine aminopeptidase (Abu tag at position 2), and labeled by stabiligase (60 MetAP peptides in replicate 1, 58 in replicate 2) or the stabiligase cocktail (63 MetAP peptides in replicate 1, 44 in replicate 2).

(**d**) Comparison of P1' amino acids of signal peptidase substrates captured by stabiligase (55 SPP peptides in replicate 1, 57 in replicate 2) or the stabiligase cocktail (50 SPP peptides in replicate 1, 69 in replicate 2) compared to the predicted frequency of P1' amino acids in predicted signal peptide cleavage sites. Asterisks indicate values that are significantly different (two-tailed, unpaired t-test, Holm-Sidak corrected p-value < 0.05) from the predicted distribution of P1' amino acids generated by signal peptidase cleavage. Blue and grey bars indicate the mean frequency (n = 2 cell culture replicates), dots indicate the frequencies observed in the individual replicates, and error bars are ± s.d.