COMMUNICATION

Efficient Multiple Domain Ligation for Proteins using Asparaginyl Endopeptidase by Selection of Appropriate Ligation Sites Based on Steric Hindrance

Aya Okuda[ɫ], Masahiro Shimizu[ɫ], Rintaro Inoue, Reiko Urade* and Masaaki Sugiyama*

Institute for Integrated Radiation and Nuclear Science, Kyoto University, Kumatori, Sennan-gun, Osaka 590-0494, Japan
E-mail: sugiyama.masaaki.5n@kyoto-u.ac.jp, urade.reiko.8w@kyoto-u.ac.jp
ɫ   These authors contributed equally

Supporting information for this article is given via a link at the end of the document.

Abstract: Three domain fragments of a multi-domain protein, ER-60, were ligated in two short linker regions using asparaginyl endopeptidase not involving denaturation. To identify appropriate ligation sites, by selecting several potential ligation sites with fewer mutations around two short linker regions, their ligation efficiencies and the functions of the ligated ER-60s were examined experimentally. To evaluate the dependence of ligation efficiencies on the ligation sites computationally, steric hinderances around the sites for the ligation were calculated through molecular dynamics simulations. Utilizing the steric hindrance, a site-dependent ligation potential index was introduced as reproducing the experimental ligation efficiency. Referring to this index, the reconstruction of ER-60 was succeeded by the ligation of the three domains for the first time. In addition, the new ligation potential index well-worked for application to other domain ligations. Therefore, the index may serve as a more time-effective tool for multi-site ligations.

Domain ligation is an essential technique for providing a multi-domain protein (MDP) with differently labeled domains. With the ligated MDP, domain conformations and dynamics can be measured using neutron scattering, nuclear magnetic resonance, and fluorescence resonance energy transfer[1-11]. Progress in these studies has involved MDPs with two domains, as well as three or more domains, making multi-site ligation techniques an important research tool.

There are several techniques used for protein ligation, which are mainly classified into three methods (Figure S1 and Table S1): protein trans-splicing with split inteins[12-18], enzymatic methods[19-22], and native chemical ligation[3,12,18,23,24]. Several single-site ligations were performed using these methods[2,3,5,6,8-11]. Regarding the ligation at two or more sites, there are only a few reports due to the difficulty that all ligation sites should have reasonably high efficiencies. The succeeded studies utilized a site in a long unstructured region, which was induced by denaturation[1,7,25,26] or its intrinsic nature[4], meaning that the long unstructured region has the advantage of high ligation efficiency. However, protein denaturation carries the risk of poor refolding after ligation. Furthermore, addition of a large fragment, such as a split intein in a protein trans-splicing method, may result in the insolubilization of the domain with the fragment[19,27] Therefore, it is crucially demanded for multi-site ligation on any MDP to develop a method with high ligation efficiency even at a site in a short flexible region but not involving denaturation.

To this end, we focused on an enzymatic method that does not require denaturation in a ligation reaction and has a low risk of insolubilization. In this study, we first targeted the multi-domain reconstruction of ER-60: ER-60, which is an oxidative protein-folding enzyme, is one of typical MDPs with short flexible regions between domains (Figures S2 and S3)[28-31]. In this study, multi-site ligations of ER-60 not involving denaturation were carried out by enzymatic ligations using *Oldenlandia affinis* asparaginyl endopeptidase (*Oa*AEP)[8,10,27,32-36] (Figures 1). The structures and functions of ER-60 and *Oa*AEP are described in Supporting Information (Functional and structural information of ER-60 and *Oa*AEP, Figures S2-S4).
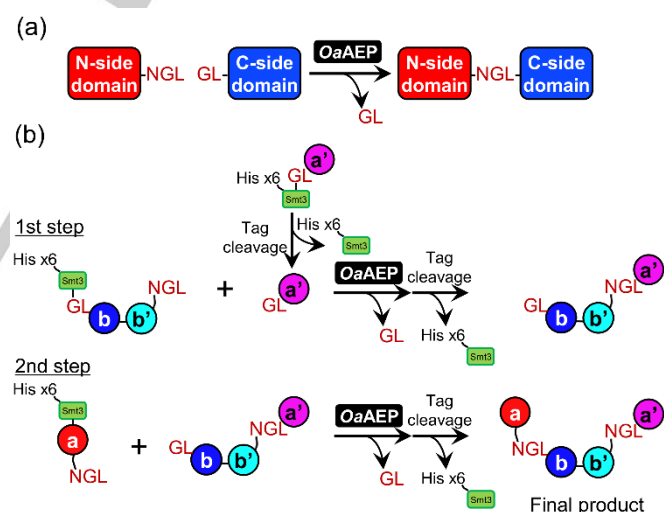

Figure 1. Schematic representation of *Oa*AEP-mediated protein ligation. (a) Ligation reaction with *Oa*AEP. *Oa*AEP recognizes the N-G-L motif in the C-terminus of the N-side domain, resulting in the cleavage of G-L, and then *Oa*AEP connects the C-terminus of the N-side domain with the G-L motif in the N-terminus of the C-side domain. (b) Process for multi-site ligation for three domains, **a**, **bb**', and **a**', of ER-60. The His x6-smt3 tag was added to the domains to protect for the unintentional ligation in each reaction step. In the first step, the ligation reaction of the **bb**' domain and **a**' domain of which the tag was priorly cleavage was performed, followed by the purification and cleavage of the tag on the **bb**' domain. In the second step, the ligation reaction of the **a** domain and the ligated **bb**'-**a**' domain was conducted, followed by the purification and cleavage of the tag on the **a** domain.

To achieve multi-site ligation for MDP, appropriate ligation sites should be identified under the following three conditions. The first is that the ligation sites are set near the domain boundary,

retaining the enzyme recognition motif with few mutations without structural disruptions. The second is that the selected sites have high ligation efficiencies. The third is that the ligated products have the same function as the wild-type. Generally, it is possible to identify several ligation sites which satisfy the first condition for the sequences. Therefore, in this study, by selecting several potential ligation sites, their ligation efficiencies were determined, followed by the experimental determination of the functions of the ligated proteins.

ER-60 reconstitution was designed by ligation of three fragments: the **a**, **bb'**, and **a'** domains (Figure 1b). The potential ligation sites were selected as follows: *Oa*AEP recognition requires an N-G-L motif in the C- terminus of the N-side domain and a G-L motif in the N-terminus of the C-side domain (Figure 1a). To minimize mutations in the recognition sites, two ligation sites were selected near the boundary between the **a** and **b** domains (S(**a**:**b**)-1 and S(**a**:**b**)-2) and four sites were between the **b'** and **a'** domains (S(**b'**:**a'**)-1, S(**b'**:**a'**)-2, S(**b'**:**a'**)-3, and S(**b'**:**a'**)-4) (Figure 2 and Figure S3).
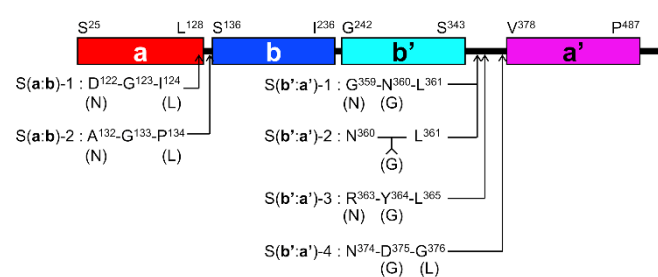

Figure 2. Schematic representation of the ER-60 domain structure and locations of the ligation sites. S(**a**:**b**)-1 and S(**a**:**b**)-2 were set between the **a** and **b** domains. S(**b'**:**a'**)-1, S(**b'**:**a'**)-2, S(**b'**:**a'**)-3, and S(**b'**:**a'**)-4 were between the **b'** and **a'** domains. As shown in parentheses, the *Oa*AEP recognition motif, N-G-L, was created by mutation, except for S(**b'**:**a'**)-2 of which the motif was created by insertion. S(**b'**:**a'**)-1 and S(**b'**:**a'**)-2 were designed at the overlapping positions.

To identify appropriate ligation sites between the six selected sites, the single-site ligations were performed at them. Domain fragments with ligation motifs of the six selected sites were individually prepared, followed by ligation reactions (Figure S5 and Methods in Supporting Information). Figures 3 and S6 show the ligation results at the six selected sites and Table S2 summarizes their ligation efficiencies. The result clearly indicates a considerable difference in ligation efficiencies between the ligation sites. Between the **a** and **b** domains, S(**a**:**b**)-1 achieved a ligation efficiently with 66% under the optimal conditions while the reaction did not proceed in S(**a**:**b**)-2. Between the **b'** and **a'** domains, the reaction in S(**b'**:**a'**)-3 achieved the most efficient ligation with a yield of 69%. For S(**b'**:**a'**)-4 and S(**b'**:**a'**)-2, the reaction efficiencies were 67% and 36%, respectively. In S(**b'**:**a'**)-1, the efficiency was extremely low.

The quality of the ligation products was checked by SDS-PAGE and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Figure S7a and S7b). The measured molecular weights agreed with the calculated values within experimental error (Table S3). Their functions were evaluated based on their oxidative refolding activities. No difference was observed in terms of the activities between the ligated products and wild-type ER-60 (Table 1 and Figure S7c), indicating that the enzymatic function of ER-60 was not affected by our ligation methodology.
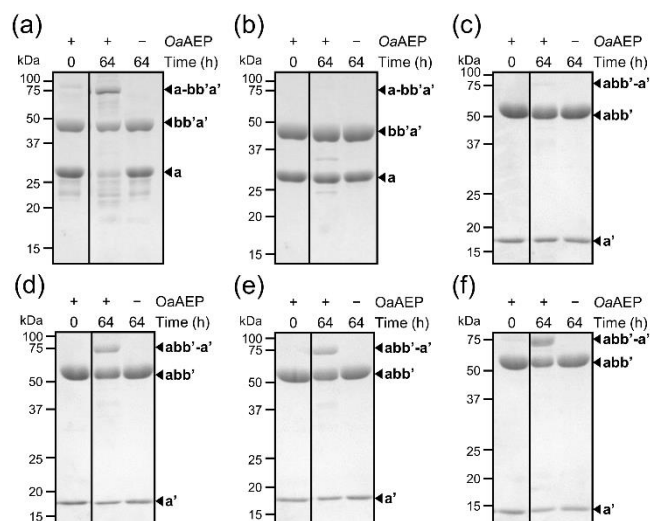

Figure 3. SDS-PAGE of ligation products from two domain fragments of ER-60 by *Oa*AEP (reaction times of 64 h at 30°C). The ligation sites are (a) S(**a**:**b**)-1 and (b) S(**a**:**b**)-2 between the **a** and **b** domains, and (c) S(**b'**:**a'**)-1, (d) S(**b'**:**a'**)-2, (e) S(**b'**:**a'**)-3, and (f) S(**b'**:**a'**)-4 between the **b'** and **a'** domains.

Table 1. Oxidative refolding activities of the wild-type (WT) and purified ligation products of ER-60.

	WT	S(**a**:**b**)-1	S(**b'**:**a'**)-2	S(**b'**:**a'**)-3	S(**b'**:**a'**)-4
Activity (mmol/mol/ min)	199.1 ± 8.9	176.6 ± 8.9	197.1 ± 44.5	175.8 ± 39.3	206.4 ± 2.6

The errors indicate standard deviations (N = 3).

As above-mentioned experiments, the evaluation of the ligation efficiencies for many potential ligation sites is necessary to achieve multi-site ligation. However, this requires extensive time and labor. Therefore, information on the ligation efficiency for each site, such as a site-dependent ligation-potential index, is helpful in identifying appropriate ligation sites. Prior to the two-site ligation of ER-60, we conducted molecular dynamics simulations to investigate the differences in ligation efficiencies between the selected sites, followed by a consideration of the site-dependent ligation-potential index.

It is possible that one of the dominant factors for ligation efficiency involves steric hindrance around the recognition motif for an enzyme-substrate intermediate. If the domain fragment directly linking to the recognition motif has a bulky structure, there is less free space around the recognition motif due to the steric hindrance. Accordingly, such a bulky structure prevents the recognition motif from accessing the binding region of the enzyme. On the other hand, if the recognition motif was directly linked to the flexible chain, the recognition motif could have sufficient free space and easily accesses to the binding region of the enzyme. To estimate the degree of steric hindrance around a recognition motif, N_{inv} was defined as the invasion number of amino acids in the vicinity of the recognition motif within the domain fragment; therefore, N_{inv} reflects the shape of the domain fragment to which the recognition motif directly links. Next, the threshold number $N_{inv}*$ was introduced. When N_{inv} of a ligation fragment is less than $N_{inv}*$ ($N_{inv} \leq N_{inv}*$), the recognition motif can easily access the binding region of the enzyme because of sufficient free space. In this case, the ligation fragment is categorized into a "sterically favorable state" (F-state). Conversely, in the case of $N_{inv} > N_{inv}*$,

the ligation fragment is categorized as a "sterically unfavorable state" (U-state). Furthermore, the ligation fragment temporally fluctuates between the two states owing to its dynamics in solution. Consequently, the appearance probability of the F-state (P_F) is regarded as the site-dependent ligation-potential index for the fragment of the ligation site.
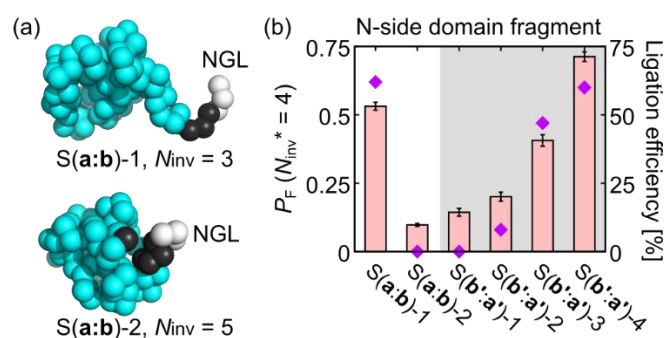

Figure 4. (a) Two representative simulation snapshots of N-side fragments, (up) N_{inv}=3 and (bottom) N_{inv}=5. The NGL motif, the residues within 10 Å around N and the other residues are expressed with white, black, and cyan, respectively. (b) Appearance probabilities of the sterically favorable states with N_{inv}^* = 4 (P_F) based on coarse-grained molecular dynamics simulations and experimental ligation efficiencies at 4ºC for 16 h for the N-side domain fragments. P_F and experimental ligation efficiencies are shown with bars and rhombuses, respectively.

In this study, the long-term structural fluctuations of all domain fragments were sampled using coarse-grained molecular dynamics (CG-MD) simulations[37,38]. An area of 10 Å around N in the N-G-L motif of the N-side fragment was set as the vicinity, and that around L in the G-L motif of the C-side fragment was set likewise. This size was sufficiently large to evaluate the bulkiness of the structure linked to the motif. The number of amino acids in the area was monitored as N_{inv} in the simulation (Figure S8). Based on the N_{inv} value, each snapshot was apportioned to the F-state ($N_{inv} \leq N_{inv}^*$) or U-state ($N_{inv} > N_{inv}^*$). The numbers of F-states and U-states, N (F-state) and N (U-state), were calculated over the simulation time. The appearance probability of the F-state, P_F, was defined as N (F-state)/ [N (F-state) + N (U-state)]. The P_F values for the domain fragments should correspond to the experimental ligation efficiencies in our enzymatic system. Therefore, to find the N_{inv}^* value which provides the appropriate P_F values, this calculation was repeated by changing N_{inv}^* as shown in Figure S9. In the simulation snapshots, the N-G-L motifs of the N-side fragments with N_{inv} = 3 or 4 were mainly linked to the disordered loop and had sufficient free space. In contrast, the motifs were often directly linked to the folded structure and had less free space when N_{inv} = 5 or 6 (Figures 4a and S10). When N_{inv}^* was set to 3 or 4, the P_F values for the N-side domain fragments were found to well-correlate with the experimental ligation efficiencies (Figures 4b, S11, S12, and S13, and Table S4). Therefore, it was reasonable to set the threshold at N_{inv}^* = 4 or lower to evaluate steric hindrance.

As indicating the ligation experiment, even though S(**b':a'**)-1 and S(**b':a'**)-2 were overlapping sites, the former showed a much lower reaction efficiency than the latter. The difference between the two sites was the method to create the enzyme recognition motif, involving mutation (G-N-L=>N-G-L) for S(**b':a'**)-1 or insertion (N-L=>N-G-L) for S(**b':a'**)-2 (Figure 2). Therefore, the ligation efficiency could be affected by the slight difference in the ligation motif and P_F (N_{inv}^* = 4) well predicted the effect on the ligation efficiency.

The measured ligation efficiency of *Oa*AEP could be related to the steric hindrance around the ligation motif of the N-side fragments, whereas the appropriate N_{inv}^* was not provided in the C-side fragment (Figure S12b). It has been reported that accessibility to the motif of the N-side fragment is the more dominant factor for reaction efficiency compared with that of the C-side fragment in an enzymatic reaction[39]. Accordingly, P_F with N_{inv}^* = 4 at the N-side fragment well-correlated with these experimental ligation efficiencies, suggesting that the difference in ligation efficiencies should be interpreted as the difference in the accessibility of the N-side fragment to the binding region of the enzyme. It has also been reported that a flexible and exposed structure close to the C-terminus is required for ligation with sortase[40, 41]. Our index of steric hindrance has semi-quantified the degree of the flexibility and exposedness the C-terminus region of N-side domain for *Oa*AEP. The P_F in the N-side fragment can be regarded as a ligation-potential index.

The final step was to challenge the ligation of three fragments, involving the **a**, **bb'**, and **a'** domains (Figure 1b), with the appropriate ligation sites derived from experiments and calculations. Considering the efficiencies of ~~both~~ the reaction and purification, S(**a:b**)-1 and S(**b':a'**)-2 were selected. First, the **bb'** and **a'** domain fragments were ligated at S(**b':a'**)-2. After purification, the ligation product **bb'-a'** and **a** domain fragment were ligated at S(**a:b**)-1. The first and second ligations were successfully carried out (Figure 5 and Figure S14). The yields of the first ligation product, **bb'-a'**, and the final product, **a-bb'-a'**, were 46% and 71%, respectively (Table S5).
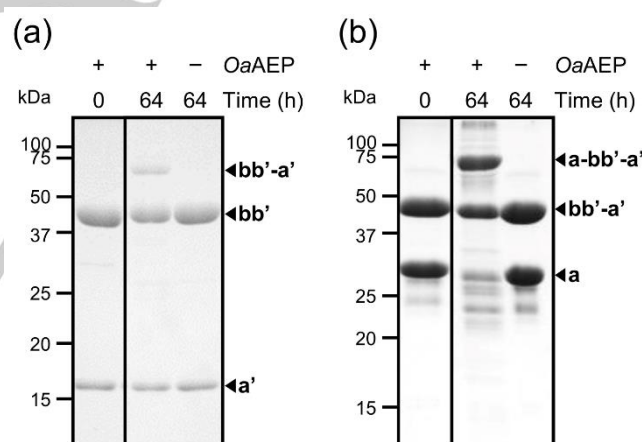

Figure 5. SDS-PAGE of ligation products in the two-sites ligation of ER-60 by *Oa*AEP (reaction times of 64 h at 30°C). (a) The first step: the **bb'** and **a'** domains were ligated at S(**b':a'**)-2. (b) The second step: the **a** domain and the ligated **bb'-a'** domain were ligated at S(**a:b**)-1.

In summary, this study experimentally verified that ligation efficiency is remarkably dependent on ligation sites. This means that identifying efficient ligation sites is important for multi-site ligation. By exploring the ligation sites and reaction conditions, we identified those with an efficiency of 69%. This value is approximately 1.4-fold higher than that of the former record with *Oa*AEP[10]. Utilizing these results, two-step domain ligation was succeeded not involving denaturation. To the best of our knowledge, this is the first report of the ligation reaction of more than three domains with an enzyme that does not involve denaturation.

In this study, the appearance probability of a sterically favorable state of a ligating domain fragment were calculated through CG-MD simulations. The calculated probability was experimentally confirmed to be regarded as a ligation-potential index. The feasibility of the proactive application of our index was examined with human PDI, which is another MDP (Figure S15). In this examination, firstly the indices at the potential ligation sites were calculated (Figures S16 and S17, and Table S6) and then examined them experimentally (Figure S18). As a result, the ligation reactions were successful at the site with high indexes ($P_F \geq 0.5$) and unsuccessful at those with low ones ($P_F \leq 0.2$) (Table S7). To further examine the index, chimera ligation reactions of ER-60 and PDI were performed (Figure S19), and succeeded and failed by the sites of N-side domain with a high index ($P_F \geq 0.5$) and a low one ($P_F \leq 0.2$), respectively (Figure S20 and Table S8 for ligation between various ER-60 N-side domain fragments and PDI C-side domain fragments, and Figure S21 and Table S9 for those between of various PDI N-side domain fragments and ER-60 C-side domain fragments, respectively). All ligation experiments indicated that P_F with $N_{inv}{}^* = 4$ at the N-side fragment well worked as a site-dependent ligation potential index. The detailed experiments and results are described in Supporting Information, Application of the ligation-potential index to ligation of PDI. In addition, as another practical application, partially deuterated and hydrogenated domains were successfully ligated for neutron scattering experiments (Figure S22).

In conclusion, it is crucially important to identify appropriate ligation sites for multi-domain ligations of a wide variety of proteins. The newly developed site-dependent ligation-potential index has a potential to be a useful parameter for identifying the feasible ligation sites with lesser time and labor. We are therefore conducting the refinement and improvement of the index by using different enzymes, enzyme recognition motifs, and other parameters.

[1]   U. K. Blaschke, G. J. Cotton, T. W. Muir, *Tetrahedron* **2000**, *56*, 9461–9470.

[2]   F. Vitali, A. Henning, F. C. Oberstrass, Y. Hargous, S. D. Auweter, M. Erat, F. H.-T. Allain, *EMBO J.* **2006**, *25*, 150–162.

[3]   L. Skrisovska, F. H.-T. Allain, *J. Mol. Biol.* **2008**, *375*, 151–164.

[4]   A. E. L. Busche, A. S. Aranko, M. Talebzadeh-Farooji, F. Bernhard, V. Dötsch, H. Iwaï, *Angew. Chem. Int. Ed.* **2009**, *48*, 6128–6131.

[5]   Y. Minato, T. Ueda, A. Machiyama, I. Shimada, H. Iwaï, *J. Biomol. NMR* **2012**, *53*, 191–207.

[6]   L. Freiburger, M. Sonntag, J. Hennig, J. Li, P. Zou, M. Sattler, *J. Biomol. NMR* **2015**, *63*, 1–8.

[7]   C. Gallagher, F. Burlina, J. Offer, A. Ramos, *Sci. Rep.* **2017**, *7*, 14083.

[8]   K. M. Mikula, I. Tascón, J. J. Tommila, H. Iwaï, *FEBS Lett.* **2017**, *591*, 1285–1294.

[9]   M. Sonntag, P. K. A. Jagtap, B. Simon, M.-S. Appavou, A. Geerlof, R. Stehle, F. Gabel, J. Hennig, M. Sattler, *Angew. Chem. Int. Ed.* **2017**, *56*, 9322–9325.

[10] K. M. Mikula, L. Krumwiede, A. Plückthun, H. Iwaï, *J. Biomol. NMR* **2018**, *71*, 225–235.

[11] T. Aizu, T. Suzuki, A. Kido, K. Nagai, A. Kobayashi, R. Sugiura, Y. Ito, M. Mishima, *Biochim. Biophys. Acta BBA - Gen. Subj.* **2020**, *1864*, 129419.

[12] V. Muralidharan, T. W. Muir, *Nat. Methods* **2006**, *3*, 429–438.

[13] G. Volkmann, H. Iwaï, *Mol. Biosyst.* **2010**, *6*, 2110–2121.

[14] A. S. Aranko, G. Volkmann, **2011**, *2*, 183–198.

[15] A. S. Aranko, A. Wlodawer, H. Iwaï, *Protein Eng. Des. Sel.* **2014**, *27*, 263–271.

[16] A. S. Aranko, J. S. Oeemig, D. Zhou, T. Kajander, A. Wlodawer, H. Iwaï, *Mol. Biosyst.* **2014**, *10*, 1023–1034.

[17] J. A. Gramespacher, A. J. Stevens, R. E. Thompson, T. W. Muir, *Protein Sci.* **2018**, *27*, 614–619.

[18] Z. A. Wang, P. A. Cole, *Methods Mol. Biol. Clifton NJ* **2020**, *2133*, 1–13.

[19] T. Proft, *Biotechnol. Lett.* **2009**, *32*, 1.

[20] M. Schmidt, A. Toplak, P. J. Quaedflieg, T. Nuijens, *Curr. Opin. Chem. Biol.* **2017**, *38*, 1–7.

[21] S. Xu, Z. Zhao, J. Zhao, *Chin. Chem. Lett.* **2018**, *29*, 1009–1016.

[22] T. Nuijens, A. Toplak, M. Schmidt, A. Ricci, W. Cabri, *Front. Chem.* **2019**, *7*, DOI 10.3389/fchem.2019.00829.

[23] P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. Kent, *Science* **1994**, *266*, 776–779.

[24] H. Sakaue, T. Kinouchi, N. Fujii, T. Takata, N. Fujii, *ACS Omega* **2017**, *2*, 260–267.

[25] V. Y. Torbeev, S. B. H. Kent, *Angew. Chem. Int. Ed.* **2007**, *46*, 1667–1670.

[26] M. Haj-Yahya, H. A. Lashuel, *J. Am. Chem. Soc.* **2018**, *140*, 6611–6621.

[27] T. M. Simon Tang, L. Y. P. Luk, *Org. Biomol. Chem.* **2021**, *19*, 5048–5062.

[28] N. Hirano, F. Shibasaki, R. Sakai, T. Tanaka, J. Nishida, Y. Yazaki, T. Takenawa, H. Hirai, *Eur. J. Biochem.* **1995**, *234*, 336–342.

[29] R. Urade, H. Okudo, H. Kato, T. Moriyama, Y. Arakaki, *Biochemistry* **2004**, *43*, 8858–8868.

[30] G. Dong, P. A. Wearsch, D. R. Peaper, P. Cresswell, K. M. Reinisch, *Immunity* **2009**, *30*, 21–32.

[31] A. Okuda, M. Shimizu, K. Morishima, R. Inoue, N. Sato, R. Urade, M. Sugiyama, *Sci. Rep.* **2021**, *11*, 5655.

[32] K. S. Harris, T. Durek, Q. Kaas, A. G. Poth, E. K. Gilding, B. F. Conlan, I. Saska, N. L. Daly, N. L. van der Weerden, D. J. Craik, M. A. Anderson, *Nat. Commun.* **2015**, *6*, 10199.

[33] R. Yang, Y. H. Wong, G. K. T. Nguyen, J. P. Tam, J. Lescar, B. Wu, *J. Am. Chem. Soc.* **2017**, *139*, 5351–5358.

[34] K. S. Harris, R. F. Guarino, R. S. Dissanayake, P. Quimbar, O. C. McCorkelle, S. Poon, Q. Kaas, T. Durek, E. K. Gilding, M. A. Jackson, D. J. Craik, N. L. van der Weerden, R. F. Anders, M. A. Anderson, *Sci. Rep.* **2019**, *9*, 10820.

[35] F. B. H. Rehm, T. J. Harmand, K. Yap, T. Durek, D. J. Craik, H. L. Ploegh, *J. Am. Chem. Soc.* **2019**, *141*, 17388–17393.

[36] T. M. S. Tang, D. Cardella, A. J. Lander, X. Li, J. S. Escudero, Y.-H. Tsai, L. Y. P. Luk, *Chem. Sci.* **2020**, *11*, 5881–5888.

[37] W. Li, W. Wang, S. Takada, *Proc. Natl. Acad. Sci.* **2014**, *111*, 10550–10555.

[38] H. Kenzaki, N. Koga, N. Hori, R. Kanada, W. Li, K. Okazaki, X.-Q. Yao, S. Takada, *J. Chem. Theory Comput.* **2011**, *7*, 1979–1989.

[39] A. M. Weeks, J. A. Wells, *Chem. Rev.* **2020**, *120*, 3127–3160.

[40] M. W. Popp, J. M. Antos, G. M. Grotenbreg, E. Spooner, H. L. Ploegh, *Nat. Chem. Biol.* **2007**, *3*, 707–708.

[41] M. W. Popp, K. Artavanis-Tsakonas, H. L. Ploegh, *J. Biol. Chem.* **2009**, *284*, 3593–3602.

Entry for the Table of Contents
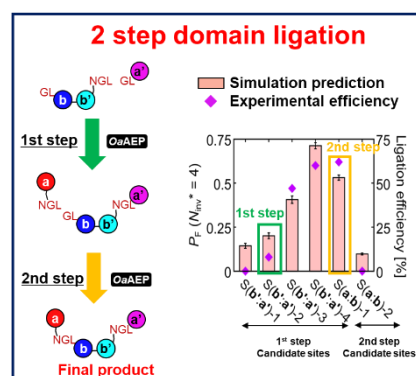
It was experimentally verified in an enzymatic domain-ligation that ligation efficiency depends on ligation sites. Subsequently, a site-dependent ligation-potential index was computationally developed, considering the steric hindrance of ligation site. Referring to this index, the three-domains ligation of ER-60 was succeeded. This index is applicable to identify feasible ligation sites