nature chemistry

Article

https://doi.org/10.1038/s41557-023-01411-x

Mirror-image trypsin digestion and sequencing of D-proteins

Received: 21 February 2023

Accepted: 28 November 2023

Published online: 18 January 2024

Check for updates

Guanwei Zhang [1,2,3] & Ting F. Zhu [2,3]

The development of mirror-image biology systems and related applications is hindered by the lack of effective methods to sequence mirror-image (D-) proteins. Although natural-chirality (L-) proteins can be sequenced by bottom–up liquid chromatography–tandem mass spectrometry (LC–MS/MS), the sequencing of long D-peptides and D-proteins with the same strategy requires digestion by a site-specific D-protease before mass analysis. Here we apply solid-phase peptide synthesis and native chemical ligation to chemically synthesize a mirror-image version of trypsin, a widely used protease for site-specific protein digestion. Using mirror-image trypsin digestion and LC–MS/MS, we sequence a mirror-image large subunit ribosomal protein (L25) and a mirror-image *Sulfolobus solfataricus* P2 DNA polymerase IV (Dpo4), and distinguish between different mutants of D-Dpo4. We also perform writing and reading of digital information in a long D-peptide of 50 amino acids. Thus, mirror-image trypsin digestion in conjunction with LC–MS/MS may facilitate practical applications of D-peptides and D-proteins as potential therapeutic and informational tools.

Mirror-image peptides and proteins composed of D-amino acids and the achiral glycine have been investigated widely as potential therapeutic and enzymatic tools because of their resistance to natural-chirality enzyme digestion and their exceptional biostability[1,2]. However, their resistance to natural-chirality proteases also prevents the sequencing of long D-peptides and D-proteins by traditional bottom–up liquid chromatography–tandem mass spectrometry (LC–MS/MS), because it is necessary to carry out fragmentation through the use of site-specific proteases such as trypsin and endopeptidase Lys-C to generate short peptides, typically 300–2,000 m/z, for tandem mass analysis[3,4].

The total chemical syntheses of mirror-image versions of various enzymes have been realized by combining solid-phase peptide synthesis (SPPS)[5] and native chemical ligation (NCL)[6]. These enzymes include the human immunodeficiency virus type 1 (HIV-1) protease[7], 4-hydroxy-tetrahydrodipicolinate synthase (DapA)[8], *Bacillus amyloliquefaciens* ribonuclease (barnase)[9], African swine fever virus polymerase X (ASFV pol X)[10], *Sulfolobus solfataricus* P2 DNA polymerase IV (Dpo4)[11,12], *Pyrococcus furiosus* (*Pfu*) DNA polymerase[13] and bacteriophage T7 RNA polymerase[14]. The chemically synthesized mirror-image enzymes are up to 883 amino acids (aa), or ~100 kDa[14], and exhibit similar activities on their respective mirror-image substrates as their natural-chirality counterparts[7–14]. However, the mirror-image versions of site-specific proteases capable of digesting long D-peptides and D-proteins have so far not been synthesized.

In this work, we set out to chemically synthesize a mirror-image version of trypsin (Fig. 1a), which cleaves the peptide bond at the C terminus of lysine or arginine. Although trypsin is a relatively small 223-aa enzyme, its in vitro folding might be problematic because the protease digests itself[15,16]. We reasoned that trypsin autolysis could be avoided by chemically synthesizing and in vitro-folding its zymogen form, trypsinogen[17], which does not exhibit protease activity because of an inhibitory propeptide at the N terminus until its cleavage during activation[18–20]. We further tested whether the chemically synthesized and in vitro-folded mirror-image trypsin could enable the digestion and sequencing of long D-peptides and D-proteins.

[1]School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China. [2]School of Life Sciences, New Cornerstone Science Laboratory, Research Center for Industries of the Future, Westlake University, Hangzhou, China. [3]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. e-mail: tzhu@westlake.edu.cn

https://doi.org/10.1038/s41557-023-01411-x

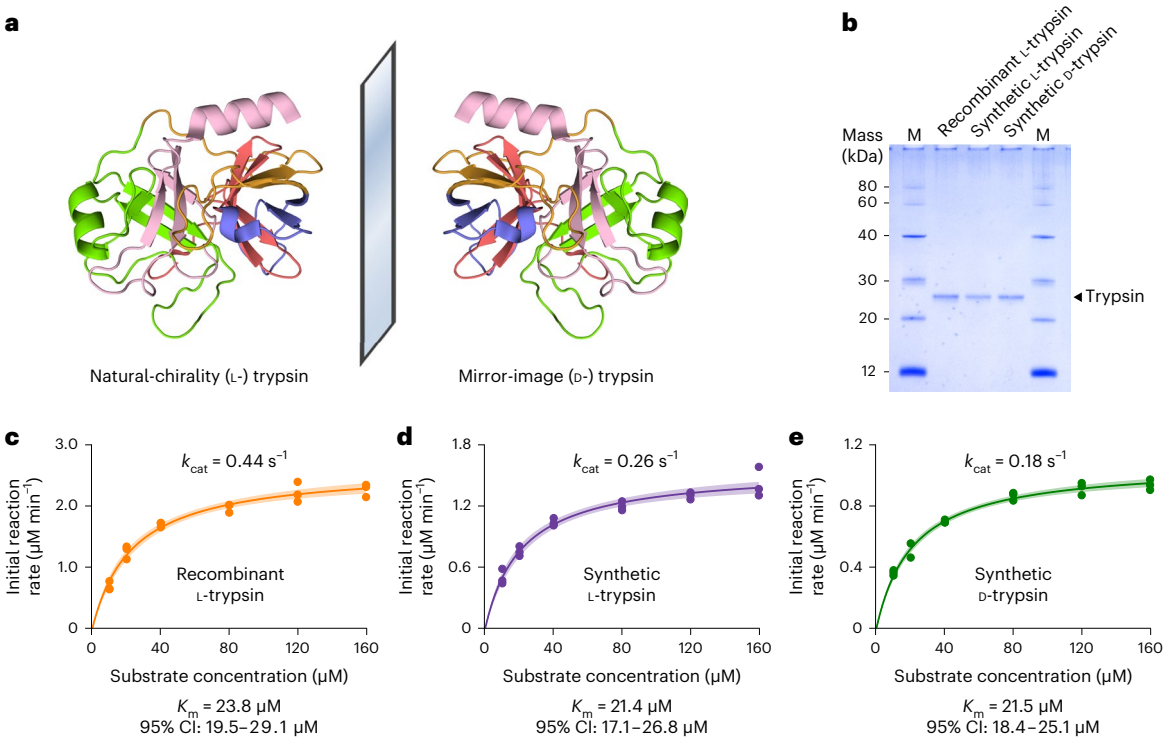


Fig. 1 | Synthetic natural-chirality (L-) and synthetic mirror-image (D-) trypsins. a, Structures of natural-chirality (L-) trypsin (PDB 5XWL) and mirror-image (D-) trypsin (model generated by digital reflection). The amino acid colours correspond to the peptide segment colours used in Extended Data Fig. 1. **b**, Recombinant L-, synthetic L- and synthetic D-trypsin analysed by SDS–PAGE and stained by Coomassie brilliant blue. M, protein marker. The experiment was performed three times with similar results. **c**–**e**, Kinetic assays for digesting substrate peptide L-LYAARLYAVR by the recombinant L- (**c**) and synthetic L-trypsin (**d**), and for digesting substrate peptide D-LYAARLYAVR by the synthetic D-trypsin (**e**). Data are presented as best-fit curves, with individual data points (n = 3) and 95% confidence intervals (CIs).

Results

Synthesis and characterization of mirror-image trypsin

To perform their total chemical synthesis, the natural-chirality and mirror-image porcine trypsinogens were each divided into five peptide segments ranging from 29 to 64 aa in length (Extended Data Fig. 1). All of the peptide segments were prepared by 9-fluorenylmethoxycarbonyl (Fmoc)-SPPS, purified by reversed-phase high-performance liquid chromatography (RP–HPLC), and assembled by hydrazide-based NCL[21], followed by metal-free radical-based desulfurization[22] to convert unprotected cysteine to alanine[23]. After the synthesis, ligation, purification and lyophilization (Supplementary Figs. 1–22), the L- and D- versions of trypsinogen were obtained at milligram scales with the expected molecular mass of 24.4 kDa (Supplementary Table 1). Meanwhile, we expressed and purified the recombinant porcine trypsinogen from *Escherichia coli* (*E. coli*). The recombinant L-, synthetic L- and synthetic D- versions of trypsinogen were folded and autoactivated in vitro (Methods), then analysed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE; Fig. 1b and Supplementary Table 2).

To biochemically characterize the autoactivated recombinant L-, synthetic L- and synthetic D- versions of trypsin, the L- and D- versions of a short substrate peptide (LYAARLYAVR) were chemically synthesized by SPPS and purified by RP–HPLC (Extended Data Fig. 2a,b), and the trypsin-digested products were analysed by RP–HPLC[24]. We found that the synthetic L-peptide was digestible by the recombinant L- and synthetic L-trypsin, but not by the synthetic D-trypsin. The synthetic D-peptide, on the other hand, was digestible by the synthetic D-trypsin, but not by the recombinant L- and synthetic L-trypsin (Extended Data Fig. 2c–h). The reciprocal chiral specificity was consistent with previous studies on the synthetic mirror-image HIV-1 protease[7]. The initial reaction rates were determined at different substrate concentrations, and fitted to the Michaelis–Menten equation, suggesting that the Michaelis constants (K_m) of the recombinant L-, synthetic L- and synthetic D-trypsin were similar, whereas the catalytic constants (k_{cat}) of the synthetic L- and synthetic D-trypsin were about half that of the recombinant L-trypsin (Fig. 1c–e), probably due to the lower purity of the synthetic trypsins (Supplementary Figs. 11 and 22).

Digestion and sequencing of ribosomal protein D-L25

Next, we applied the synthetic D-trypsin to the bottom–up sequencing of small D-proteins. Ribosomal proteins are generally lysine- and arginine-rich and thus are suitable substrates for trypsin digestion. We tested the enzymatic activities of the recombinant L-, synthetic L- and synthetic D-trypsin on the synthetic L- and synthetic D-L25, an *E. coli* ribosomal protein[25]. We found that the synthetic L-L25 was digestible by the recombinant L- and synthetic L-trypsin, but not by the synthetic D-trypsin, whereas the synthetic D-L25 was digestible by the synthetic D-trypsin, but not by the recombinant L- and synthetic L-trypsin (Fig. 2a–f and Extended Data Fig. 3a–e).

The trypsin-digested and desalted peptide fragments of L- and D-L25 were analysed by LC–MS/MS. The peptide-spectrum matches (PSMs; Supplementary Data) resulted in a protein sequence coverage of 98.9% for L-L25 after digestion by the recombinant L- (Fig. 2g) and synthetic L-trypsin (Fig. 2h), and 98.9% for D-L25 after digestion by the synthetic D-trypsin (Fig. 2i), hence effectively validating the D-protein sequence of D-L25. Meanwhile, the cleavage site specificity was confirmed by the precursor intensity of tryptic and nontryptic peptides[16], suggesting that the recombinant L-, synthetic L- and synthetic D-trypsin displayed similar cleavage preferences at the C terminus of lysine or arginine (Extended Data Fig. 4a), with proline at the P1′ site prohibiting trypsin cleavage (Extended Data Fig. 4c)[26,27]. These results suggested

https://doi.org/10.1038/s41557-023-01411-x

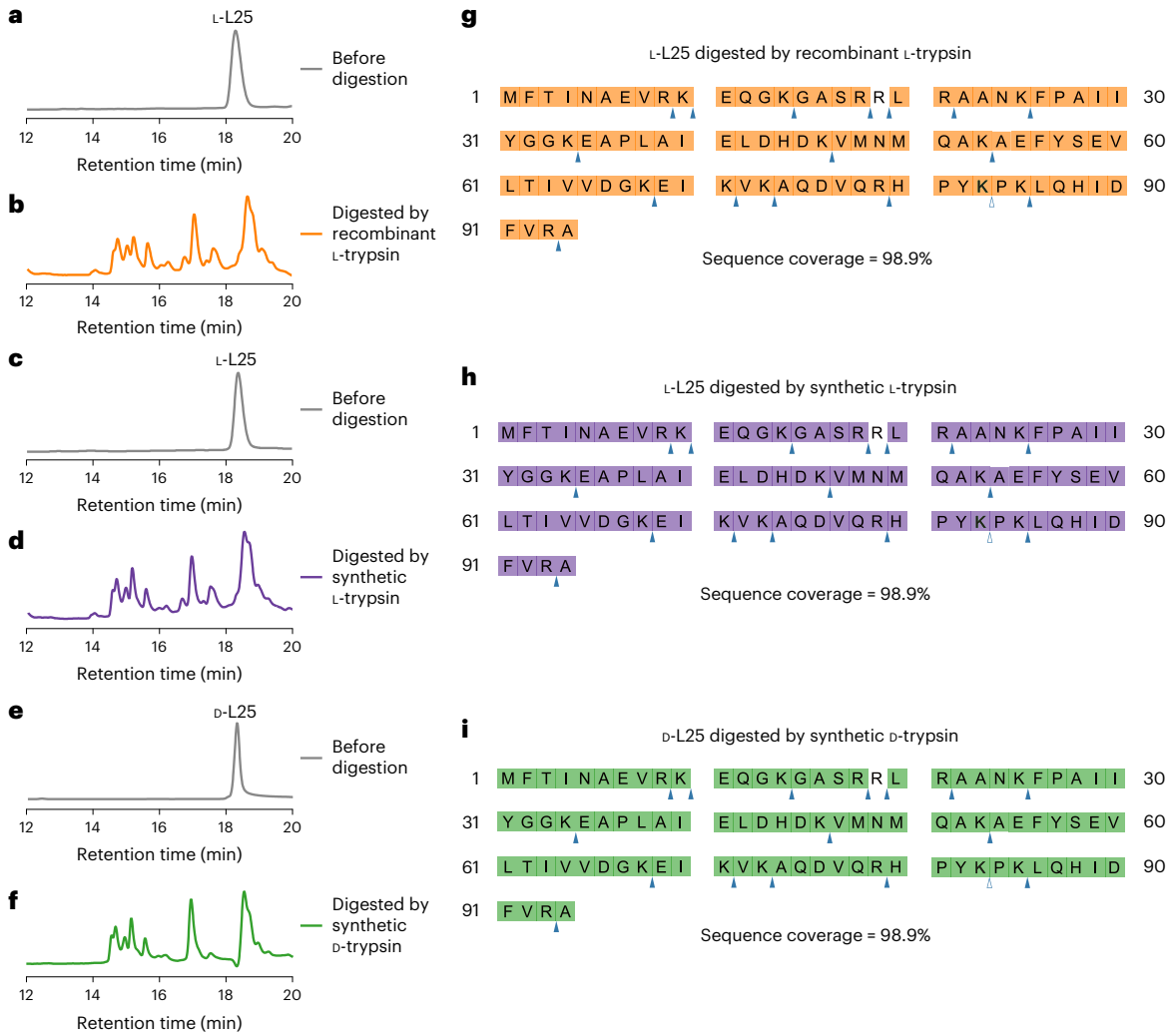


Fig. 2 | Digestion and sequencing of ribosomal protein D-L25. a–f, Analytical RP–HPLC chromatograms of the synthetic L-L25 before (**a**) and after (**b**) digestion by the recombinant L-trypsin, and before (**c**) and after (**d**) digestion by the synthetic L-trypsin, and of the synthetic D-L25 before (**e**) and after (**f**) digestion by the synthetic D-trypsin. **g–i**, Protein sequence coverage of the synthetic L-L25 digested by the recombinant L- (**g**) and synthetic L-trypsin (**h**), and of the synthetic D-L25 by the synthetic D-trypsin (**i**). Amino acids covered by PSMs of ≥1 are highlighted. Blue filled triangles, trypsin cleavage sites. Blue open triangles, prohibited trypsin cleavage sites with proline at the P1′ site. The experiments were performed twice with similar results.

that, for protein sequencing, the synthetic D-trypsin was equally effective and site-specific in digesting small D-proteins including the 94-aa D-L25 as the recombinant L- and the synthetic L-trypsin.

Digestion and sequencing of D-Dpo4

Next, we tested the synthetic D-trypsin on the bottom–up sequencing of a larger D-protein, Dpo4-5m, a mutant version of Dpo4 that facilitated its total chemical synthesis[11,13,28]. Because the 358-aa Dpo4-5m (with an N-terminal His_6 tag) is much larger than the 94-aa ribosomal protein L25, denaturation of Dpo4-5m was performed before trypsin digestion (Methods). We found that the recombinant L-Dpo4-5m was digestible by the recombinant L- and synthetic L-trypsin, but not by the synthetic D-trypsin, whereas the synthetic D-Dpo4-5m was digestible by the synthetic D-trypsin, but not by the recombinant L- and synthetic L-trypsin (Fig. 3a–f and Extended Data Fig. 3f–j).

The trypsin-digested and desalted peptide fragments of L- and D-Dpo4-5m were analysed by LC–MS/MS. The PSMs (Supplementary Data) resulted in protein sequence coverages of 99.2% and 97.8% for L-Dpo4-5m after digestion by the recombinant L- (Fig. 3g) and synthetic L-trypsin (Fig. 3h), respectively, and 99.4% for D-Dpo4-5m after digestion by the synthetic D-trypsin (Fig. 3i), hence effectively validating the D-protein sequence of D-Dpo4-5m. Similarly, the cleavage site specificity was confirmed by the precursor intensity of tryptic and nontryptic peptides[16], suggesting that the recombinant L-, synthetic L- and synthetic D-trypsin displayed similar cleavage preferences at the C terminus of lysine or arginine (Extended Data Fig. 4b), with proline at the P1′ site prohibiting trypsin cleavage (Extended Data Fig. 4d)[26,27]. These results suggested that, for protein sequencing, the synthetic D-trypsin was equally effective and site-specific in digesting large D-proteins including the 358-aa D-Dpo4-5m as the recombinant L- and synthetic L-trypsin.

Distinguishing between different mutants of D-Dpo4

Encouraged by the high protein sequence coverage of D-Dpo4-5m, we sought to apply the mirror-image trypsin digestion and D-protein sequencing method to distinguish between two different mutants of D-Dpo4: D-Dpo4-5m and D-Dpo4-5m-Y12S (a mutant version capable of polymerizing L-RNA, with a tyrosine replaced by serine and all six methionines by norleucines when compared with D-Dpo4-5m)[29].

The trypsin-digested and desalted peptide fragments of the synthetic D-Dpo4-5m and D-Dpo4-5m-Y12S were analysed by LC–MS/MS. To distinguish between the two different mutants of D-Dpo4, we focused

https://doi.org/10.1038/s41557-023-01411-x

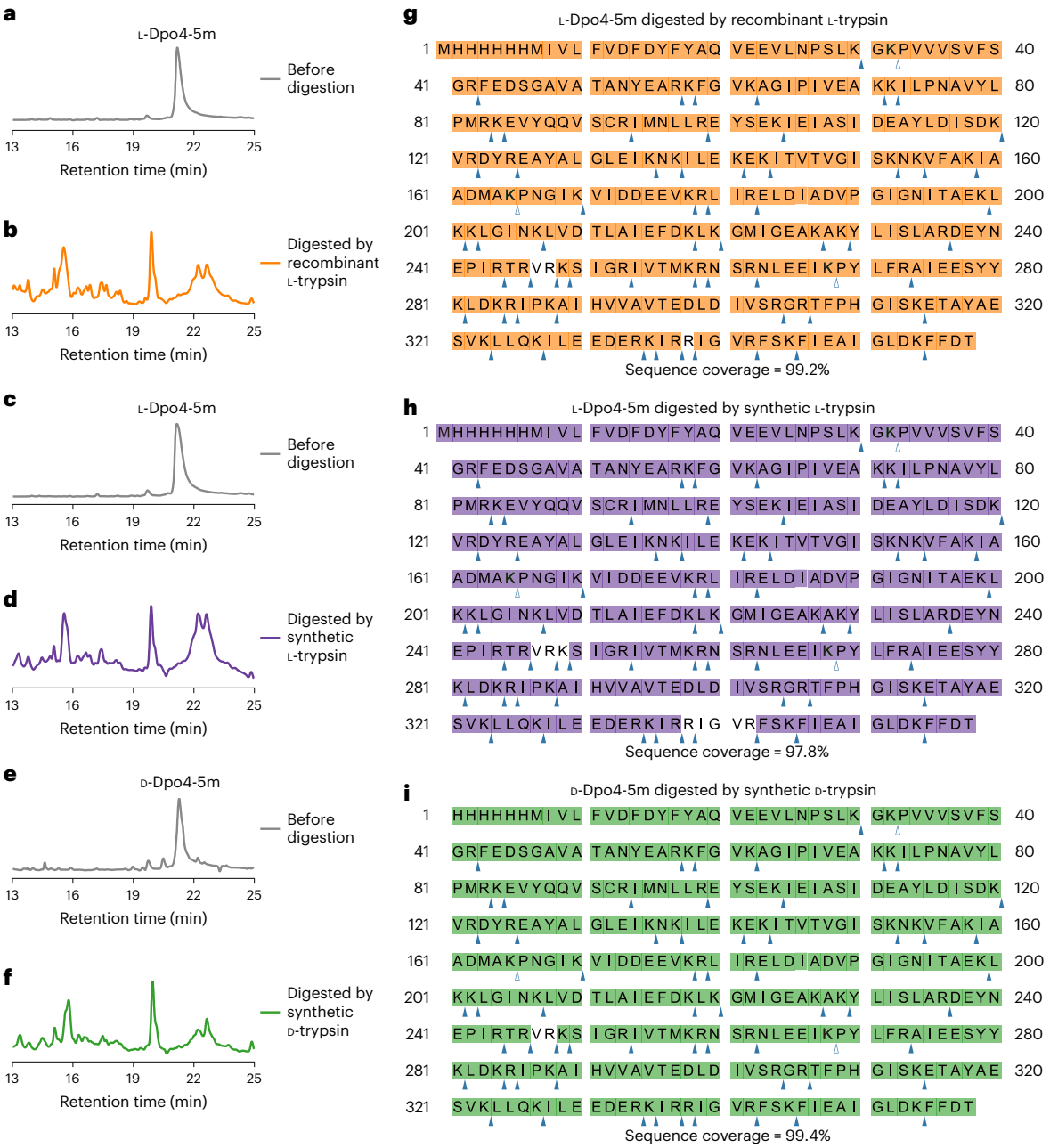


Fig. 3 | Digestion and sequencing of D-Dpo4. a–**f**, Analytical RP–HPLC chromatograms of the recombinant L-Dpo4-5m before (**a**) and after (**b**) digestion by the recombinant L-trypsin, and before (**c**) and after (**d**) digestion by the synthetic L-trypsin, and of the synthetic D-Dpo4-5m before (**e**) and after (**f**) digestion by the synthetic D-trypsin. **g**–**i**, Protein sequence coverage of the recombinant L-Dpo4-5m digested by the recombinant L- (**g**) and synthetic L-trypsin (**h**), and of the synthetic D-Dpo4-5m by the synthetic D-trypsin (**i**). Amino acids covered by PSMs of ≥1 are highlighted. Blue filled triangles, trypsin cleavage sites. Blue open triangles, prohibited trypsin cleavage sites with proline at the P1′ site. The experiments were performed twice with similar results.

on the PSMs (Supplementary Data) containing the aforementioned seven mutated residues. PSMs of D-Dpo4-5m allowed the identification of all the six methionines and the unmutated tyrosine, whereas PSMs of D-Dpo4-5m-Y12S allowed the identification of all the six norleucines and the mutated serine (Table 1). These results suggested that the mirror-image trypsin digestion and D-protein sequencing method was capable of distinguishing between different mutants of a D-protein.

Writing and reading information in a long D-peptide

Although information storage in short L- and D-peptides with lengths of 10–18 aa has been reported[30–32], the retrieval of information from D-peptides longer than ~20 aa remains challenging without a site-specific protease, limiting the amount of information that can be stored in D-peptides. We reasoned that using D-trypsin, information-storing long D-peptides can be digested into short peptides and sequenced by LC–MS/MS, enabling the storage of more information in long D-peptides.

Here we selected a 20-character phrase 'Mirror-image biology' and encoded it into a 50-aa D-peptide, chemically synthesized by SPPS and purified by RP–HPLC, with five arginines to provide trypsin cleavage sites and five index amino acids to provide the order of the trypsin-digested peptide fragments (Fig. 4a). The information-storing D-peptide was digested by the synthetic D-trypsin into five 10-aa D-peptide fragments, and the desalted digestion products were

https://doi.org/10.1038/s41557-023-01411-x

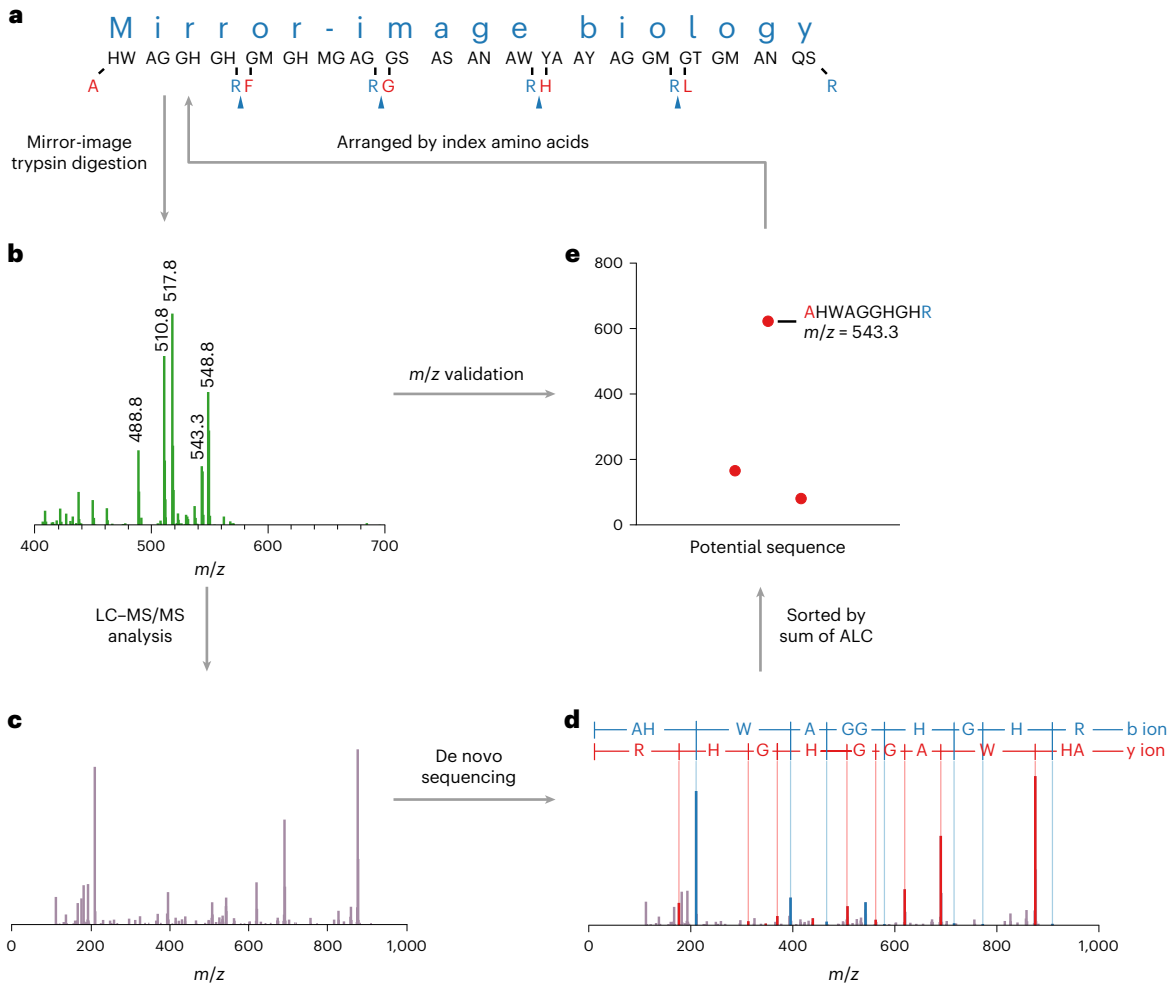


Fig. 4 | Writing and reading information in a long D-peptide. a, Design of an information-storing 50-aa D-peptide, chemically synthesized by SPPS and purified by RP–HPLC. Blue filled triangles, trypsin cleavage sites. Amino acids in red are N-terminal index amino acids of the 10-aa D-peptide fragments, in the order A, F, G, H, L. Amino acids in blue are C-terminal arginines of the 10-aa D-peptide fragments. **b**, ESI–MS spectrum of the mirror-image trypsin-digested information-storing 50-aa D-peptide. **c**,**d**, Tandem mass spectrum (**c**) and de novo sequencing (**d**), with an example of the tandem mass spectra of the 10-aa D-peptide fragment indexed by alanine (AHWAGGHGHR) shown here and the rest shown in Extended Data Fig. 5. **e**, Sorting of de novo sequencing results by the sums of the ALC of the potential 10-aa D-peptide sequences indexed by alanine. The five 10-aa D-peptide sequences (**e** and Extended Data Fig. 5) were arranged by the index amino acids and decoded into the original phrase, 'Mirror-image biology' (**a**). The experiment was performed twice with similar results.

analysed by electrospray ionization–mass spectrometry (ESI–MS; Fig. 4b), followed by LC–MS/MS (Fig. 4c). The de novo sequencing results (Fig. 4d) were filtered (Methods), and the sums of the average local confidence (ALC) of the potential 10-aa D-peptide sequences were sorted to determine their sequences (Fig. 4e and Extended Data Fig. 5), which were arranged by the index amino acids and decoded into the original phrase, 'Mirror-image biology' (Fig. 4a). In comparison, LC–MS/MS analysis of the undigested information-storing 50-aa D-peptide failed to determine its sequence (Extended Data Fig. 6). Therefore, using D-trypsin, information can be retrieved from long D-peptides, expanding the amount of information that can be stored in long D-peptides and securing the information with the synthetic D-trypsin as a 'key', despite the lack of ability to be amplified by mirror-image polymerase chain reaction (PCR) like the information-storing L-DNAs[13].

Discussion

In this work, we demonstrated mirror-image protein sequencing by chemically synthesizing the mirror-image version of a site-specific protease, trypsin, enabling the sequencing of D-proteins and information storage in long D-peptides, echoing the first chemically synthesized D-enzyme, the mirror-image HIV-1 protease[7]. The mirror-image trypsin digestion and D-protein sequencing method may become a useful tool for the quality control of chemically synthesized D-peptide and D-protein drugs[33–37]. Being able to distinguish between different mutants of D-proteins, the mirror-image trypsin digestion and D-protein sequencing method may also be applied to the identification of D-proteins with modified amino acids. Combined with de novo sequencing, mirror-image trypsin digestion may also become a helping hand in discovering D-peptide drugs when applied in conjunction with screening methods such as affinity selection–MS (AS–MS)[38–40], as well as in selecting small-molecule drugs with information-storing D-peptide barcodes[32]. Moreover, the mirror-image versions of trypsin and other proteases can eliminate D-peptides and D-proteins after use as a containment strategy. In addition, the chemically synthesized L-trypsin, free from protease and other protein contamination from traditional protein expression systems[16,41], may provide a contamination-free tool for sample preparation prior to LC–MS/MS analysis in proteomic studies.

The next key step in establishing the mirror-image central dogma of molecular biology is to realize mirror-image translation by synthesizing a mirror-image ribosome[10,14,25,42–44]. Since LC–MS/MS is widely used in the quantitative and semi-quantitative analyses of different protein components in protein and RNA–protein complexes[45–48] including the

Table 1 | Distinguishing between different mutants of D-Dpo4

Mutant D-Dpo4	Residue	PSMs of trypsin-digested D-Dpo4-5m[a]	PSMs of trypsin-digested D-Dpo4-5m-Y12S[a]
D-Dpo4-5m	Met7	1	0
	Tyr18	1	0
	Met82	111	0
	Met95	16	0
	Met163	94	0
	Met222	27	0
	Met257	41	0
D-Dpo4-5m-Y12S	Nle7	0	2
	Ser18	0	2
	Nle82	0	209
	Nle95	0	49
	Nle163	0	220
	Nle222	0	16
	Nle257	0	13

[a]PSMs for distinguishing between the two mutants of D-Dpo4: D-Dpo4-5m and D-Dpo4-5m-Y12S. Y12S corresponds to the Tyr18 to Ser18 mutation in D-Dpo4-5m-Y12S with an N-terminal D-His_6 tag.

ribosome, the mirror-image trypsin digestion and D-protein sequencing method may become a useful tool for validating mirror-image ribosome assembly, as well as the D-protein products of mirror-image translation.

One of the limitations of the current mirror-image trypsin digestion and D-protein sequencing method is that it only digests D-proteins at the C terminus of D-lysine or D-arginine. This can be potentially complemented by chemically synthesizing the mirror-image versions of other widely used site-specific proteases, such as chymotrypsin, endopeptidase Lys-C and endopeptidase Glu-C[4]. Another limitation is that the current synthetic route suffers from low yield (estimated overall synthesis yield of ~0.5%), in part due to the suboptimal synthesis and ligation efficiencies (Supplementary Figs. 1–22), which may affect the yields (Supplementary Table 1), purities (Supplementary Figs. 11 and 22) and activities (Fig. 1d,e) of the synthetic L- and synthetic D-trypsin. This can be potentially improved by more advanced synthesis methods and by split-protein designs with point mutations[13,14].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-023-01411-x.

References

1. Kent, S. B. Novel protein science enabled by total chemical synthesis. *Protein Sci.* **28**, 313–328 (2019).
2. Harrison, K., Mackay, A. S., Kambanis, L., Maxwell, J. W. C. & Payne, R. J. Synthesis and applications of mirror-image proteins. *Nat. Rev. Chem.* **7**, 383–404 (2023).
3. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
4. Zhang, Y. Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).
5. Merrifield, R. B. Solid phase peptide synthesis. 1. Synthesis of a tetrapeptide. *J. Am. Chem. Soc.* **85**, 2149–2154 (1963).
6. Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779 (1994).
7. Milton, R., Milton, S. & Kent, S. B. Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity. *Science* **256**, 1445–1448 (1992).
8. Weinstock, M. T., Jacobsen, M. T. & Kay, M. S. Synthesis and folding of a mirror-image enzyme reveals ambidextrous chaperone activity. *Proc. Natl Acad. Sci. USA* **111**, 11679–11684 (2014).
9. Vinogradov, A. A., Evans, E. D. & Pentelute, B. L. Total synthesis and biochemical characterization of mirror image barnase. *Chem. Sci.* **6**, 2997–3002 (2015).
10. Wang, Z., Xu, W., Liu, L. & Zhu, T. F. A synthetic molecular system capable of mirror-image genetic replication and transcription. *Nat. Chem.* **8**, 698–704 (2016).
11. Jiang, W. et al. Mirror-image polymerase chain reaction. *Cell Discov.* **3**, 17037 (2017).
12. Pech, A. et al. A thermostable D-polymerase for mirror-image PCR. *Nucleic Acids Res.* **45**, 3997–4005 (2017).
13. Fan, C., Deng, Q. & Zhu, T. F. Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image *Pfu* DNA polymerase. *Nat. Biotechnol.* **39**, 1548–1555 (2021).
14. Xu, Y. & Zhu, T. F. Mirror-image T7 transcription of chirally inverted ribosomal and functional RNAs. *Science* **378**, 405–412 (2022).
15. Vestling, M. M., Murphy, C. M. & Fenselau, C. Recognition of trypsin autolysis products by high-performance liquid chromatography and mass spectrometry. *Anal. Chem.* **62**, 2391–2394 (1990).
16. Bunkenborg, J., Espadas, G. & Molina, H. Cutting edge proteomics: benchmarking of six commercial trypsins. *J. Proteome Res.* **12**, 3631–3641 (2013).
17. Kassell, B. & Kay, J. Zymogens of proteolytic enzymes. *Science* **180**, 1022–1027 (1973).
18. Sahin-Toth, M. Human cationic trypsinogen. Role of Asn-21 in zymogen activation and implications in hereditary pancreatitis. *J. Biol. Chem.* **275**, 22750–22755 (2000).
19. Kukor, Z., Toth, M. & Sahin-Toth, M. Human anionic trypsinogen: properties of autocatalytic activation and degradation and implications in pancreatic diseases. *Eur. J. Biochem.* **270**, 2047–2058 (2003).
20. Zhao, M., Wu, F. & Xu, P. Development of a rapid high-efficiency scalable process for acetylated *Sus scrofa* cationic trypsin production from *Escherichia coli* inclusion bodies. *Protein Expr. Purif.* **116**, 120–126 (2015).
21. Fang, G. M. et al. Protein chemical synthesis by ligation of peptide hydrazides. *Angew. Chem. Int. Ed.* **50**, 7645–7649 (2011).
22. Wan, Q. & Danishefsky, S. J. Free-radical-based, specific desulfurization of cysteine: a powerful advance in the synthesis of polypeptides and glycopolypeptides. *Angew. Chem. Int. Ed.* **46**, 9248–9252 (2007).
23. Yan, L. Z. & Dawson, P. E. Synthesis of peptides and proteins without cysteine residues by native chemical ligation combined with desulfurization. *J. Am. Chem. Soc.* **123**, 526–533 (2001).
24. Slechtova, T., Gilar, M., Kalikova, K. & Tesarova, E. Insight into trypsin miscleavage: comparison of kinetic constants of problematic peptide sequences. *Anal. Chem.* **87**, 7636–7643 (2015).
25. Ling, J. J. et al. Mirror-image 5S ribonucleoprotein complexes. *Angew. Chem. Int. Ed.* **59**, 3724–3731 (2020).
26. Keil, B. I. *Specificity of Proteolysis* (Springer, 1992).
27. Yang, H. et al. Precision de novo peptide sequencing using mirror proteases of Ac-LysargiNase and trypsin for large-scale proteomics. *Mol. Cell Proteomics* **18**, 773–785 (2019).
28. Xu, W. et al. Total chemical synthesis of a thermostable enzyme capable of polymerase chain reaction. *Cell Discov.* **3**, 17008 (2017).

29. Wang, M. et al. Mirror-image gene transcription and reverse transcription. *Chem* **5**, 848–857 (2019).
30. Ng, C. C. A. et al. Data storage using peptide sequences. *Nat. Commun.* **12**, 4242 (2021).
31. Zheng, J. S. et al. A mirror-image protein-based information barcoding and storage technology. *Sci. Bull.* **66**, 1542–1549 (2021).
32. Rossler, S. L., Grob, N. M., Buchwald, S. L. & Pentelute, B. L. Abiotic peptides as carriers of information for the encoding of small-molecule library synthesis. *Science* **379**, 939–945 (2023).
33. Eckert, D. M., Malashkevich, V. N., Hong, L. H., Carr, P. A. & Kim, P. S. Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. *Cell* **99**, 103–115 (1999).
34. Mandal, K. et al. Chemical synthesis and X-ray structure of a heterochiral {D-protein antagonist plus vascular endothelial growth factor} protein complex by racemic crystallography. *Proc. Natl Acad. Sci. USA* **109**, 14779–14784 (2012).
35. Chang, H. N. et al. Blocking of the PD-1/PD-L1 interaction by a D-peptide antagonist for cancer immunotherapy. *Angew. Chem. Int. Ed.* **54**, 11760–11764 (2015).
36. Uppalapati, M. et al. A potent D-protein antagonist of VEGF-A is nonimmunogenic, metabolically stable, and longer-circulating in vivo. *ACS Chem. Biol.* **11**, 1058–1065 (2016).
37. Marinec, P. S. et al. A non-immunogenic bivalent D-protein potently inhibits retinal vascularization and tumor growth. *ACS Chem. Biol.* **16**, 548–556 (2021).
38. Zuckermann, R. N., Kerr, J. M., Siani, M. A., Banville, S. C. & Santi, D. V. Identification of highest-affinity ligands by affinity selection from equimolar peptide mixtures generated by robotic synthesis. *Proc. Natl Acad. Sci. USA* **89**, 4505–4509 (1992).
39. Maaty, W. S. & Weis, D. D. Label-free, in-solution screening of peptide libraries for binding to protein targets using hydrogen exchange mass spectrometry. *J. Am. Chem. Soc.* **138**, 1335–1343 (2016).
40. Quartararo, A. J. et al. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat. Commun.* **11**, 3183 (2020).
41. Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **75**, 1454–1462 (2012).
42. Peplow, M. A conversation with Ting Zhu. *ACS Cent. Sci.* **4**, 783–784 (2018).
43. Chen, J., Chen, M. & Zhu, T. F. Translating protein enzymes without aminoacyl-tRNA synthetases. *Chem* **7**, 786–798 (2021).
44. Service, R. F. A big step toward mirror-image ribosomes. *Science* **378**, 345–346 (2022).
45. Cravatt, B. F., Simon, G. M. & Yates, J. R. 3rd The biological impact of mass-spectrometry-based proteomics. *Nature* **450**, 991–1000 (2007).
46. Wang, X. et al. Mass spectrometric characterization of the affinity-purified human 26S proteasome complex. *Biochemistry* **46**, 3553–3565 (2007).
47. Ori, A. et al. Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9**, 648 (2013).
48. Chen, S. S. & Williamson, J. R. Characterization of the ribosome biogenesis landscape in *E. coli* using quantitative mass spectrometry. *J. Mol. Biol.* **425**, 767–779 (2013).

Methods

Materials

The D-DNA oligos and the porcine trypsinogen gene for recombinant protein expression were ordered from Genewiz. Tris base and guanidine hydrochloride (Gn·HCl) were purchased from Amresco. Phosphate buffered saline (PBS), L-cystine and L-cysteine were purchased from Solarbio Life Sciences. Urea, $CaCl_2$ and Amicon Ultra centrifugal filter (0.5 ml, 10,000 MWCO) were purchased from Sigma. TransStart FastPfu Fly DNA polymerase, the pEASY-Uni seamless cloning and assembly kit and ProteinRuler I were purchased from TransGen Biotech. The Sep-Pak tC18 cartridge (100 mg sorbent per cartridge) was purchased from Waters Corp. The 2-chlorotrityl chloride resin was purchased from Tianjin Nankai Hecheng Science and Technology. Wang ChemMatrix resin, Fmoc-L-amino acids, Fmoc-D-amino acids, Boc-L-Ser(Fmoc-L-Val)-OH, Boc-D-Ser(Fmoc-D-Val)-OH and *O*-(6-chlorobenzotriazol-1-yl)-*N*,*N*,*N*′,*N*′-tetramethyluronium hexafluorophosphate (HCTU) were purchased from GL Biochem. *N*,*N*-dimethylformamide (DMF), *N*,*N*-diisopropylethylamine (DIEA), trifluoroacetic acid (TFA), thioanisole, triisopropylsilane (TIPS), 1,2-ethanedithiol (EDT), $PdCl_2$ and 2,2′-azobis[2-(2-imidazolin-2-yl) propane]dihydrochloride (VA-044) were purchased from J&K Scientific. 4-Mercaptophenylacetic acid (MPAA) was purchased from Alfa Aesar Chemicals. Piperidine, $Na_2HPO_4{\cdot}12H_2O$, $NaH_2PO_4{\cdot}2H_2O$ and $NaNO_2$ were purchased from Sinopharm Chemical Reagent. NaCl, NaOH and HCl were purchased from Sinopharm Chemical Reagent. Tris(2-carboxyethyl)phosphine hydrochloride (TCEP·HCl), 9-fluorenylmethyl carbazate (Fmoc-hydrazide), ethyl cyanoglyoxylate-2-oxime (Oxyma), *N*,*N*′-diisopropylcarbodiimide (DIC) and DL-1,4-dithiothreitol (DTT) were purchased from Adamas Reagent. Glutathione reduced (GSH) was purchased from Acros Organics. Anhydrous ether was purchased from Beijing Tongguang Fine Chemicals. Acetonitrile (CH_3CN, HPLC grade) was purchase from J. T. Baker.

Trypsinogen expression and purification

The porcine trypsinogen gene was amplified by PCR using TransStart FastPfu Fly DNA polymerase and cloned into the pET-28c vector with the pEASY-Uni seamless cloning and assembly kit. The recombinant trypsinogen was expressed in *E. coli* strain BL21(DE3) in lysogeny broth (LB) medium. The induced cells were collected and resuspended in PBS. The cell lysate was disrupted by sonication at 4 °C for 10 min, and the proteins were subsequently precipitated by centrifugation at 20,000*g* at 4 °C for 40 min. The precipitate was further washed three times with PBS. The trypsinogen in inclusion bodies was solubilized in 8 M Gn·HCl, purified by RP–HPLC, and lyophilized.

Fmoc-SPPS

All the peptides were synthesized by Fmoc-SPPS on Liberty Blue automated microwave peptide synthesizers (CEM Corp.). Isoacyl dipeptide[49] was incorporated at positions Val199–Ser200 in L-/D-trypsinogen-5. The peptides with C-terminal carboxylate, such as L-/D-trypsinogen-5, were synthesized on Wang ChemMatrix resin (0.6 mmol g^{-1}) preloaded with the first C-terminal residue, and the other peptides were synthesized on Fmoc-hydrazine 2-chlorotrityl resin (0.53 mmol g^{-1}) to prepare peptide hydrazides[50], the scale of which was typically 0.25 mmol. The first residue on the Wang ChemMatrix resin was manually attached by a double coupling method: in the first coupling reaction, the amino acid was coupled at 30 °C for 40 min with 1 mmol amino acid, 0.98 mmol HCTU and 2 mmol DIEA dissolved in 4 ml DMF, after which the resin was washed with DMF. Without deprotection, the second coupling reaction was performed at 25 °C overnight with 1 mmol amino acid, 1 mmol Oxyma and 1 mmol DIC dissolved in 4 ml DMF. All the resins were swelled in DMF for 5–10 min before use. The Fmoc groups of the resins and coupled amino acids were removed by treatment with 20% piperidine and 0.1 M Oxyma in DMF. The coupling of amino acids except Fmoc-Cys(Trt)-OH (Trt, triphenylmethyl), Fmoc-Cys(Acm)-OH (Acm, acetamidomethyl) and Fmoc-His(Trt)-OH was performed in 10 ml DMF with 1 mmol amino acid, 1 mmol Oxyma and 2 mmol DIC at 85 °C for 3 min. The coupling of Fmoc-Cys(Trt)-OH, Fmoc-Cys(Acm)-OH and Fmoc-His(Trt)-OH was performed at 50 °C for 10 min to reduce side reactions at higher temperatures. The coupling of trifluoroacetyl thiazolidine-4-caboxylic acid-OH (Tfa-Thz-OH) was performed with Oxyma/DIC activation at room temperature overnight[51]. The synthesized peptides were cleaved from resin using H_2O/thioanisole/TIPS/EDT/TFA (0.5/0.5/0.5/0.25/8.25, vol/vol) under agitation at 27 °C for 2.5 h. Most of the TFA in the mixture was removed by N_2 blowing, and cold ether was added to precipitate the crude peptides. After centrifugation, the supernatant was discarded and the precipitate was washed twice with ether. The crude peptides were dissolved in CH_3CN/H_2O or 6 M Gn·HCl, analysed by RP–HPLC and ESI–MS, and purified by semi-preparative RP–HPLC.

Native chemical ligation

The peptide segments were assembled by hydrazide-based NCL with a convergent assembly strategy[52]. The C-terminal peptide hydrazide segment (4–6 mM) was dissolved in acidified ligation buffer (6 M Gn·HCl, 0.1 M NaH_2PO_4, pH 3.0), with pH monitored by a FiveEasy Plus pH meter and InLab Micro pH electrode (Mettler Toledo). The mixture was cooled in an ice–salt bath at −10 °C, after which 25 mM $NaNO_2$ in acidified ligation buffer was added. The reaction mixture was kept in the ice–salt bath under stirring for 25 min, after which 100 mM MPAA in 6 M Gn·HCl, 0.1 M Na_2HPO_4, pH 5–6 was added. After the addition of the N-terminal cysteine peptide (to a 2–3 mM final concentration for both the C-terminal and N-terminal peptide segments), the pH of the reaction mixture was adjusted to 6.5 at room temperature. After overnight reaction, 150 mM TCEP in ligation buffer (pH 7.0) was added to dilute the reaction mixture twice, under stirring at room temperature for 30 min. Next, the ligation product was analysed by RP–HPLC and ESI–MS, and purified by semi-preparative RP–HPLC. The preparations of L-/D-trypsinogen-10 and L-/D-trypsinogen-11 suffered from low ligation yields (Supplementary Figs. 10, 11, 21 and 22), probably because the solubility of L-/D-trypsinogen-5 and L-/D-trypsinogen-10 decreased after the *O*-acyl isopeptide bond being converted to the native *N*-acyl peptide bond[53].

Desulfurization

Metal-free radical-based desulfurization[22] was performed by dissolving the cysteine-containing peptides (3 mg ml^{-1}) in desulfurization buffer (6 M Gn·HCl, 0.1 M Na_2HPO_4, 200 mM TCEP, 40 mM GSH, 20 mM VA-044, pH 6.8) under stirring at 37 °C overnight, and the desulfurization product was analysed by RP–HPLC and ESI–MS, and purified by semi-preparative RP–HPLC.

Acetamidomethyl deprotection

Pd-assisted Acm deprotection[54] was performed by dissolving the Acm-protected peptides in Acm deprotection buffer (6 M Gn·HCl, 0.1 M Na_2HPO_4, 40 mM TCEP, pH 7.0), after which 20 mM (final concentration) $PdCl_2$ was added, under stirring at 25 °C overnight, with 50 mM (final concentration) DTT added to quench the reaction. The reaction mixture was stirred for 1 h, analysed by RP–HPLC and ESI–MS, and purified by semi-preparative RP–HPLC.

RP–HPLC and ESI–MS

All the RP–HPLC analysis and purification experiments were performed on Shimadzu Prominence HPLC systems (Shimadzu Corp.) with SPD-20A UV–vis detectors and LC-20AT solvent delivery units. Ultimate XB-C4 120 Å (5 μm, 21.2 × 250 mm, Welch Materials) and C18 120 Å (5 μm, 21.2 × 250 mm) columns were used to purify the crude peptides at a flow rate of 8 ml min^{-1}. Ultimate XB-C4 120 Å (5 μm, 10 × 250 mm) columns were used to separate the ligation products at a flow rate of

4 ml min^{-1}. Ultimate XB-C4 300 Å (5 µm, 4.6 × 250 mm) and Inertsil C4 150 Å (5 µm, 4.6 × 250 mm, GL Sciences) columns were used to monitor the ligation reactions and analyse the purity of the ligation products and the trypsin-digested peptides or proteins at a flow rate of 1 ml min^{-1}. The molecular mass with standard deviation (s.d.) of each purified peptide segment and ligation product was characterized by ESI–MS on a Shimadzu LC/MS-2020 mass spectrometer (Shimadzu Corp.). The final products (L-trypsinogen-11 and D-trypsinogen-11) were further characterized by high-resolution ESI–MS on a Waters SYNAPT G2-Si HDMS mass spectrometer (Waters Corp.), with the purities of the final products estimated without deconvolution.

Folding and activation of trypsinogen in vitro

Lyophilized recombinant L-, synthetic L- and synthetic D- versions of trypsinogen were dissolved in denaturation buffer (8 M urea, 20 mM DTT, 20 mM Tris-HCl, pH 8.5) at room temperature for 4 h. Trypsinogen folding was performed by diluting the protein with 39× volumes of renaturation buffer (20 mM Tris-HCl, 2 M urea, 1 mM L-cystine, 3 mM L-cysteine, pH 7.5) under stirring at 4 °C for 12 h. After folding, the precipitates were removed by centrifugation at 12,000*g* at 4 °C for 30 min, followed by ultrafiltration using an Amicon Ultra centrifugal filter (0.5 ml, 10,000 MWCO). The trypsinogen was autoactivated in activation buffer (40 mM Tris-HCl, 0.1 M NaCl, 10 mM $CaCl_2$, pH 8.0) at 37 °C for 12 h, analysed by 15% SDS–PAGE, and scanned by a ChemiDoc XRS+ system (Bio-Rad Laboratories). The autoactivated trypsin was aliquoted and stored at −20 °C.

Trypsin kinetic assay

Substrate peptides (L-/D-LYAARLYAVR), chemically synthesized by SPPS and purified by RP–HPLC, were used for the trypsin kinetic assays. The substrate peptides were dissolved and diluted into activation buffer (40 mM Tris-HCl, 0.1 M NaCl, 10 mM $CaCl_2$, pH 8.0), digested at 37 °C for 10 min with a final trypsin concentration of 100 nM, and quenched by adding 1× volume of 1% TFA in H_2O. The trypsin-digested substrate peptides were analysed by RP–HPLC using an Inertsil C4 150 Å (5 µm, 4.6 × 250 mm) column, with the peak areas of UV absorption at 214 nm used to calculate the concentrations of the substrate peptides and trypsin-digested products with a standard curve. The initial reaction rates were calculated using the concentrations of the trypsin-digested products divided by the reaction time, and least-square curve fitting was used to calculate K_m and k_{cat} in the Michaelis–Menten equation using GraphPad Prism software (version 10.0.2, Dotmatics Ltd). According to the different k_{cat} values, we adjusted the concentrations of recombinant L-, synthetic L- and synthetic D-trypsin to compensate for their different activities (enzyme-to-substrate ratios of 1/133, 1/80, 1/53 (wt/wt), respectively).

Trypsin digestion of peptides and proteins

The synthetic L- and synthetic D- versions of *E. coli* ribosomal protein L25, synthetic D-Dpo4-5m and synthetic D-Dpo4-5m-Y12S were synthesized in our previous work[11,13,25,28,29]. The synthetic L- and synthetic D- versions of L25, and synthetic long D-peptide for information storage were directly dissolved and diluted into activation buffer (40 mM Tris-HCl, 0.1 M NaCl, 10 mM $CaCl_2$, pH 8.0) before trypsin digestion. The recombinant L-Dpo4-5m, synthetic D-Dpo4-5m and synthetic D-Dpo4-5m-Y12S were denatured in a buffer containing 8 M urea and 5 mM DTT for 1 h, after which 12.5 mM iodoacetamide was added to alkylate the cysteines in the dark at room temperature for 30 min. Iodoacetamide was inactivated by exposure to room light for 15 min, and the protein solution was diluted with 5× volumes of activation buffer before trypsin digestion. The trypsin digestion was performed at 37 °C for 12 h, quenched by adding 1× volume of 1% TFA in H_2O, and analysed by analytical RP–HPLC using an Inertsil C4 150 Å (5 µm, 4.6 × 250 mm) column with a gradient of 5–95% CH_3CN (with 0.1% TFA) in H_2O (with 0.1% TFA) over 30 min.

LC–MS/MS analysis of trypsin-digested peptides and proteins

The trypsin-digested peptides and proteins were desalted by a Sep-Pak tC18 cartridge, dried using a CV200 vacuum centrifugal concentrator (Beijing JM Technology), and dissolved in 20 µl 0.1% TFA in H_2O, of which 6 µl was used for LC–MS/MS analysis. The samples were separated using a 100-µm × 150-mm fused silica capillary column, packed in-house with ReproSil-Pur C18-AQ 1.9-µm resin (Dr Maisch), with a gradient of 5–95% CH_3CN (with 0.1% formic acid) in H_2O (with 0.1% formic acid) over 60 min at a flow rate of 0.30 µl min^{-1}, directly interfaced with an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific). For sequencing the ribosomal protein L25 and different mutants of Dpo4, the tandem mass spectra were searched by Proteome Discoverer software (version 2.5, Thermo Fisher Scientific) against a protein database containing the proteome of *E. coli* (UniProt Taxonomy 83333, 4,530 proteins), porcine trypsin (UniProt P00761) and different mutants of Dpo4 (refs. 11,13,28,29) (with norleucines in Dpo4-5m-Y12S replaced by leucines of the same mass), with the following settings: full trypsin specificity (for calculating protein sequence coverage) or no enzyme specificity (for calculating the frequency of non-specific cleavage), two missed cleavages allowed, oxidation of methionine (+16.00 Da) and deamination of asparagine and glutamine (+0.98 Da) set to variable modifications, carbamidomethyl of cysteine (+57.02 Da) set to static modification for sequencing the mutants of Dpo4, precursor ion mass tolerance set to 3 ppm with nonlinear mass recalibration applied, and fragment ion mass tolerance set to 0.02 Da. The false discovery rate was set to 0.01 based on the *q*-value calculated by Percolator in Proteome Discoverer software. For the analysis of cleavage site specificity, the intensity of the precursor ion of each PSM was used to calculate the amino-acid frequency at the P1 site. For information storage in a long D-peptide, PEAKS Studio software (version 8.5, BioInformatics Solutions) was used for de novo sequencing, with the following settings: no enzyme specificity, oxidation of methionine (+16.00 Da) and deamination of asparagine and glutamine (+0.98 Da) set to variable modifications, precursor ion mass tolerance set to 10 ppm, and fragment ion mass tolerance set to 0.02 Da, with the ALC calculated by PEAKS Studio software.

Writing and reading information in a long D-peptide

To encode the 128 American Standard Code for Information Interchange (ASCII) codes, 12 different amino acids were used as two-letter codes ($12^2 > 128$). Among the 20 proteinogenic amino acids, lysine and arginine were used as cleavage sites, proline at the P1′ site prohibits trypsin cleavage, aspartate and glutamate at the P2, P3, P1′ or P2′ sites inhibit trypsin cleavage[24], cysteine forms disulfide bond, and isoleucine cannot be distinguished from leucine by MS. Among the remaining 13 amino acids, the hydrophobic amino acid valine was not chosen in order to improve the solubility of the synthetic long D-peptides. The remaining 12 amino acids (alanine, phenylalanine, glycine, histidine, leucine, methionine, asparagine, glutamine, serine, threonine, tryptophan and tyrosine) were used for long D-peptide information storage. The printable ASCII codes were converted to duodecimal numbers, and to the corresponding amino acids (Supplementary Table 3). Arginines were inserted into the C terminus of each fragment to provide trypsin cleavage sites, and index amino acids were inserted into the N terminus of each fragment to provide the order of the trypsin-digested peptide fragments. A 20-character phrase, 'Mirror-image biology', was encoded into a 50-aa D-peptide, chemically synthesized by SPPS and purified by RP–HPLC. The information-storing 50-aa D-peptide was digested by the synthetic D-trypsin into five 10-aa D-peptide fragments, and the desalted digestion products were analysed by ESI–MS, followed by LC–MS/MS. The de novo sequencing results were filtered using Microsoft Excel 2019 software (version 16.43, Microsoft Corp.) by a length of 10 aa, an ALC above 80%, with an ESI–MS *m/z* error of ±0.5. The sums of the ALC of the potential 10-aa D-peptide sequences were sorted to determine their sequences, which were arranged by the index amino acids and decoded into the original phrase, 'Mirror-image biology'.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability
All data are available in the main text or the Supplementary Information. The *E. coli* proteome database (Taxonomy 83333) was downloaded from UniProt (https://www.uniprot.org). The LC–MS/MS data were deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD046228. Source data are provided with this paper.

References

49. Coin, I. The depsipeptide method for solid-phase synthesis of difficult peptides. *J. Pept. Sci.* **16**, 223–230 (2010).
50. Huang, Y.-C. et al. Facile synthesis of C-terminal peptide hydrazide and thioester of NY-ESO-1 (A39-A68) from an Fmoc-hydrazine 2-chlorotrityl chloride resin. *Tetrahedron* **70**, 2951–2955 (2014).
51. Huang, Y. C. et al. Synthesis of L- and D-ubiquitin by one-pot ligation and metal-free desulfurization. *Chemistry* **22**, 7623–7628 (2016).
52. Fang, G. M., Wang, J. X. & Liu, L. Convergent chemical synthesis of proteins by ligation of peptide hydrazides. *Angew. Chem. Int. Ed.* **51**, 10347–10350 (2012).
53. Sohma, Y. et al. 'O-Acyl isopeptide method' for the efficient synthesis of difficult sequence-containing peptides: use of 'O-acyl isodipeptide unit'. *Tetrahedron Lett.* **47**, 3013–3017 (2006).
54. Maity, S. K., Jbara, M., Laps, S. & Brik, A. Efficient palladium-assisted one-pot deprotection of (acetamidomethyl) cysteine following native chemical ligation and/or desulfurization to expedite chemical protein synthesis. *Angew. Chem. Int. Ed.* **55**, 8108–8112 (2016).

Acknowledgements
We thank J. Chen, Q. Deng, C. Fan, H. Liu, G. Wang, Y. Xu, J. Zhang and R. Zhao for assistance with the experiments, and H. Deng and X. Tian at the Tsinghua Technology Center for Protein Sciences, C. C. Wong at the Peking Union Medical College Hospital State Key Laboratory of Complex Severe and Rare Diseases, and T. Guo and Y. Zhu at the Westlake iMarker Lab for assistance with the tandem mass analysis. The work was supported by the National Natural Science Foundation of China (grant nos. 21925702 and 32050178), the Research Center for Industries of the Future (RCIF) at Westlake University, the Westlake Education Foundation, the New Cornerstone Science Foundation, the Tsinghua-Peking Center for Life Sciences (CLS) and the Beijing Frontier Research Center for Biological Structure. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

Author contributions
G.Z. performed the experiments. Both authors analysed the data and wrote the paper. T.F.Z. designed and supervised the study.

Competing interests
A provisional US patent application (no. 63/546,881) has been filed by Westlake University with T.F.Z. and G.Z. listed as inventors. The authors declare no other competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41557-023-01411-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-023-01411-x.

Correspondence and requests for materials should be addressed to Ting F. Zhu.

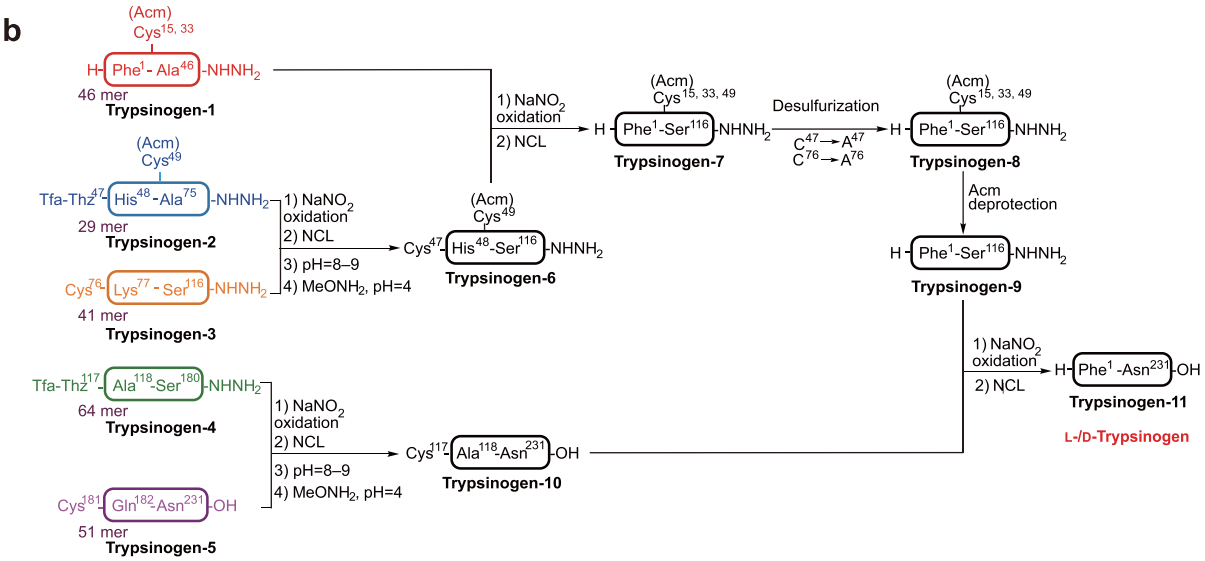
Peer review information *Nature Chemistry* thanks Michael Kay, Stephen Kent and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41557-023-01411-x

a

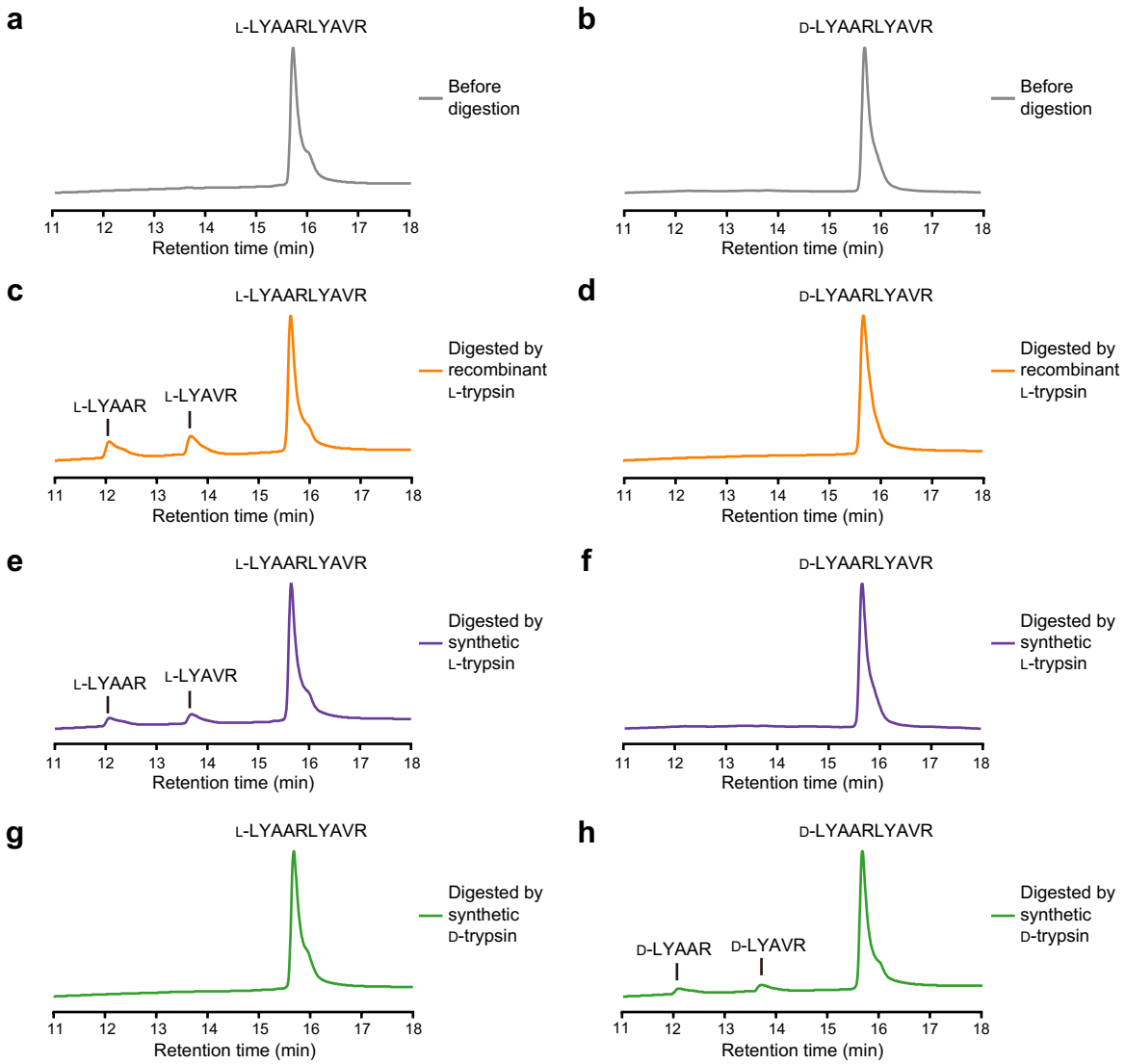
FPTDDDDKIV[10] GGYTCAANSI[20] PYQVSLNSGS[30] HFCGGSLINS[40] QWVVSAAHCY[50] KSRIQVRLGE[60]
HNIDVLEGNE[70] QFINAAKIIT[80] HPNFNGNTLD[90] NDIMLIKLSS[100] PATLNSRVAT[110] VSLPRSCAAA[120]
GTECLISGWG[130] NTKSSGSSYP[140] SLLQCLKAPV[150] LSDSSCKSSY[160] PGQITGNMIC[170] VGFLEGGKDS[180]
CQGDSGGPVV[190] CNGQLQGIVS[200] WGYGCAQKNK[210] PGVYTKVCNY[220] VNWIQQTIAA[230] N

b



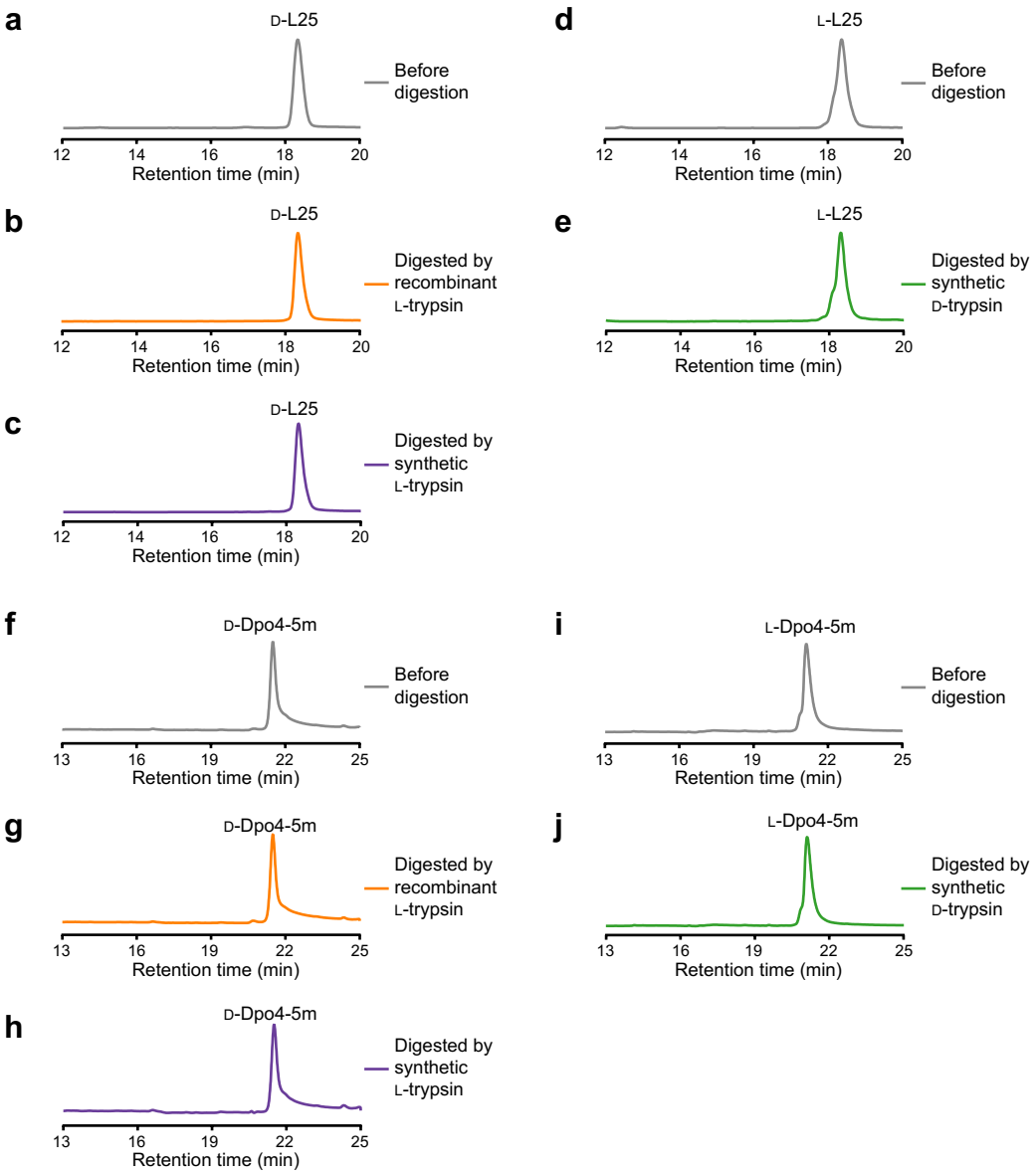
Extended Data Fig. 1 | Design of the synthetic L-/D-trypsinogen. a, Amino acid sequence of the synthetic L-/D-trypsinogen (UniProt P00761). The amino acid colours correspond to the peptide segment colours used in **b** and Fig. 1a. **b**, Synthetic route for the total chemical synthesis of L-/D-trypsinogen.

https://doi.org/10.1038/s41557-023-01411-x



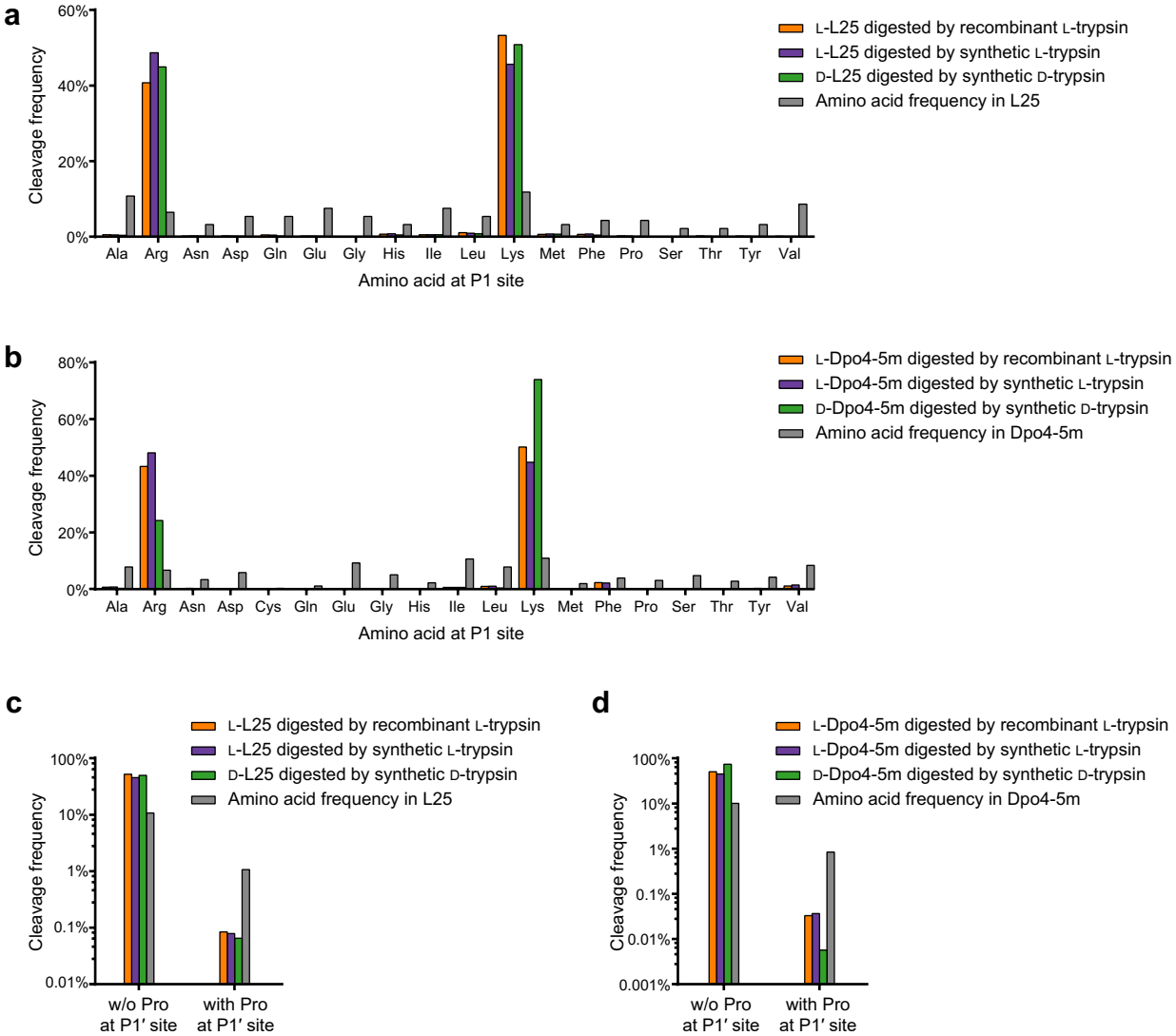
Extended Data Fig. 2 | Chiral specificity of trypsin digestion of substrate peptides. a,**b**, Analytical RP–HPLC chromatograms of the chemically synthesized L-LYAARLYAVR (**a**) and D-LYAARLYAVR (**b**) before digestion. **c**,**d**, Analytical RP–HPLC chromatograms of L-LYAARLYAVR (**c**) and D-LYAARLYAVR (**d**) digested by the recombinant L-trypsin. **e**,**f**, Analytical RP–HPLC chromatograms of L-LYAARLYAVR (**e**) and D-LYAARLYAVR (**f**) digested by the synthetic L-trypsin. **g**,**h**, Analytical RP–HPLC chromatograms of L-LYAARLYAVR (**g**) and D-LYAARLYAVR (**h**) digested by the synthetic D-trypsin. The experiments were performed three times with similar results.

https://doi.org/10.1038/s41557-023-01411-x



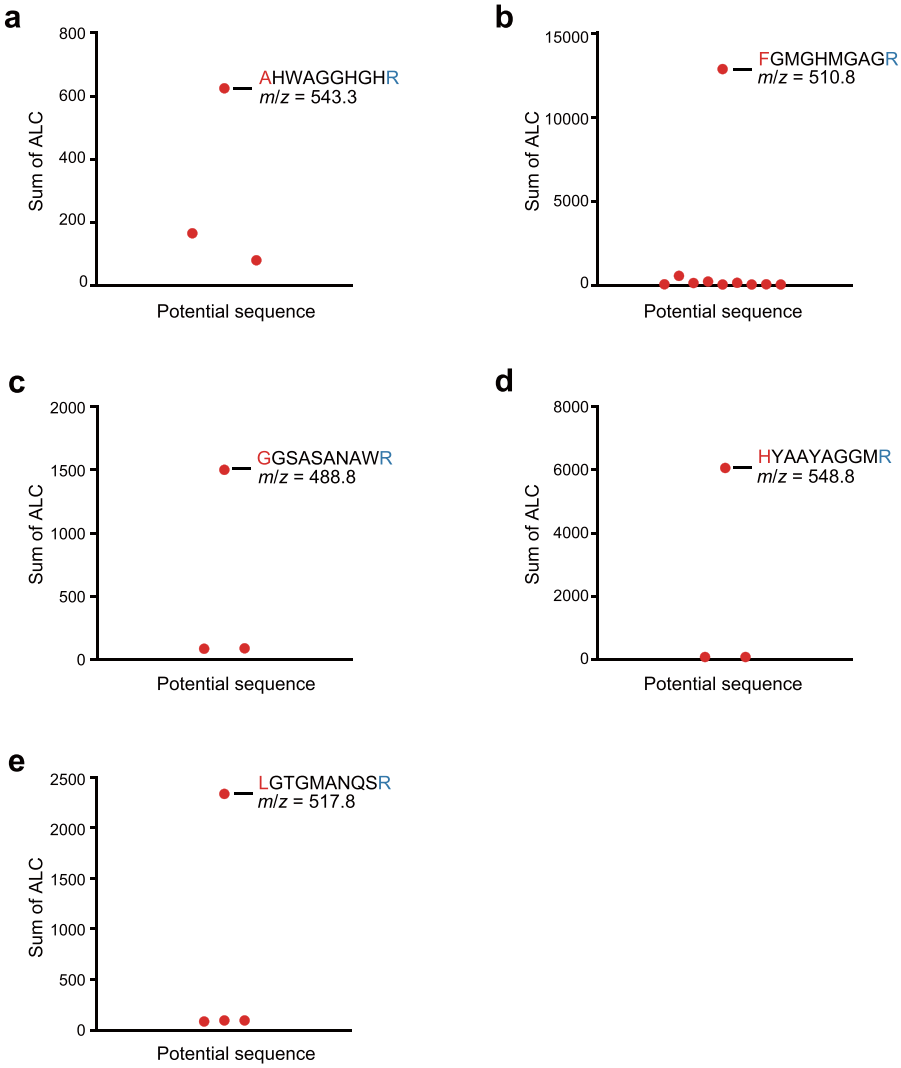
Extended Data Fig. 3 | Chiral specificity of trypsin digestion of ribosomal protein L25 and Dpo4. a–**c**, Analytical RP–HPLC chromatograms of the synthetic D-L25 before (**a**) and after digestion by the recombinant L- (**b**) and synthetic L-trypsin (**c**). **d**,**e**, Analytical RP–HPLC chromatograms of the synthetic L-L25 before (**d**) and after digestion by the synthetic D-trypsin (**e**). **f**–**h**, Analytical RP–HPLC chromatograms of the synthetic D-Dpo4-5m before (**f**) and after digestion by the recombinant L- (**g**) and synthetic L-trypsin (**h**). **i**,**j**, Analytical RP–HPLC chromatograms of the recombinant L-Dpo4-5m before (**i**) and after digestion by the synthetic D-trypsin (**j**). The experiments were performed twice with similar results.

https://doi.org/10.1038/s41557-023-01411-x



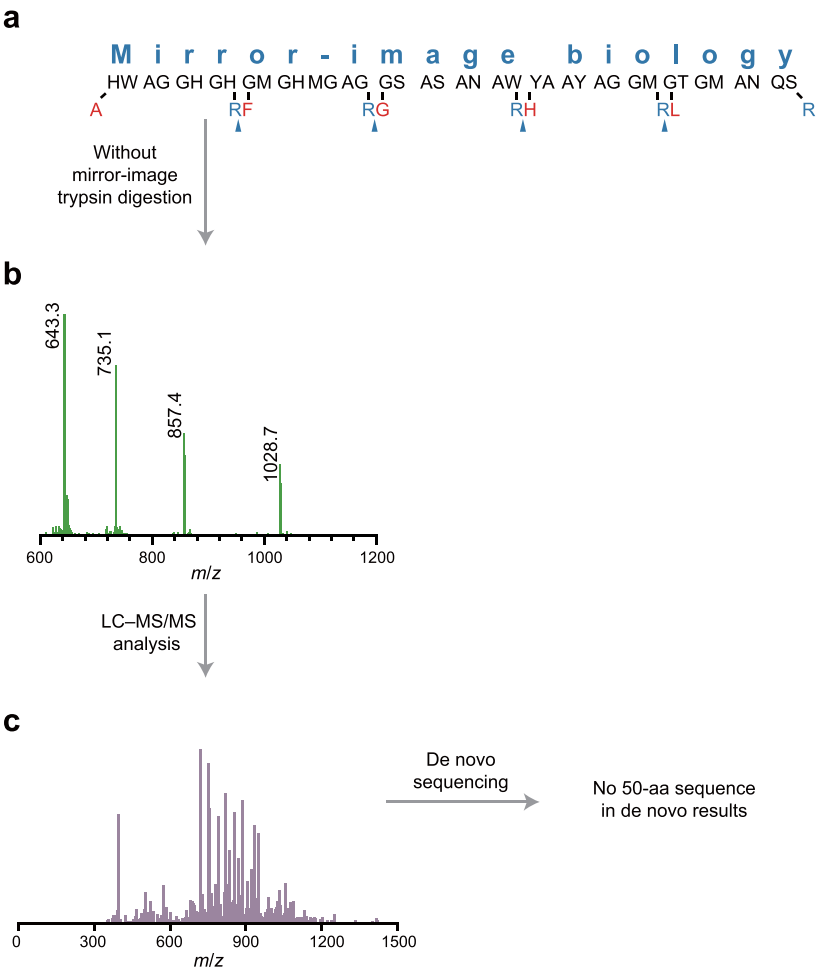
Extended Data Fig. 4 | Cleavage site specificity of trypsin digestion of ribosomal protein L25 and Dpo4. a, Observed cleavage frequency at the P1 site of L-L25 digested by the recombinant L- and synthetic L-trypsin, and of D-L25 by the synthetic D-trypsin. **b**, Observed cleavage frequency at the P1 site of L-Dpo4-5m digested by the recombinant L- and synthetic L-trypsin, and of D-Dpo4-5m by the synthetic D-trypsin. **c**, Observed cleavage frequency of L-L25 digested by the recombinant L- and synthetic L-trypsin, and of D-L25 by the synthetic D-trypsin, at trypsin cleavage sites with lysine at the P1 site and with or without proline at the P1′ site, displayed on a log scale. **d**, Observed cleavage frequency of L-Dpo4-5m digested by the recombinant L- and synthetic L-trypsin, and of D-Dpo4-5m by the synthetic D-trypsin, at trypsin cleavage sites with lysine at the P1 site and with or without proline at the P1′ site, displayed on a log scale. The experiments were performed twice with similar results.

https://doi.org/10.1038/s41557-023-01411-x



Extended Data Fig. 5 | Sorting of de novo sequencing results. a–**e**, Sorting of de novo sequencing results by the sums of the ALC of the potential 10-aa D-peptide sequences indexed by alanine (**a**, also shown in Fig. 4e), phenylalanine (**b**), glycine (**c**), histidine (**d**), and leucine (**e**). The experiment was performed twice with similar results.

https://doi.org/10.1038/s41557-023-01411-x



Extended Data Fig. 6 | LC–MS/MS analysis of the undigested information-storing 50-aa D-peptide. a, Design of an information-storing 50-aa D-peptide, also shown in Fig. 5a. **b**,**c**, ESI–MS spectrum of the undigested information-storing 50-aa D-peptide (**b**), with an example of the tandem mass spectra of the undigested D-peptide shown (**c**). No 50-aa sequence was present in the de novo sequencing results. The experiment was performed twice with similar results.

nature portfolio

Corresponding author(s): Ting Zhu

Last updated by author(s): Oct 20, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All RP–HPLC, ESI–MS, HR–ESI–MS, and LC–MS/MS analyses were performed on the Shimadzu Prominence HPLC system, the Shimadzu LC/MS-2020 mass spectrometer, the Waters SYNAPT G2-Si HDMS mass spectrometer, and the Orbitrap Exploris 480 mass spectrometer, respectively. Polyacrylamide gels were scanned by the ChemiDoc XRS+ system.
Data analysis	Proteome Discoverer (version 2.5), PEAKS Studio (version 8.5), Microsoft Excel 2019 (version 16.43), GraphPad Prism (10.0.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data are available in the main text or the Supplementary Information. The E. coli proteome database (Taxonomy 83333) was downloaded from UniProt (https://

www.uniprot.org). The LC–MS/MS data were deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD046228. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample sizes of experiments are described in the paper and the Supplementary Information. No sample size calculation was performed for the biochemical experiments.
Data exclusions	The raw tandem mass spectra were processed by Proteome Discoverer or PEAKS Studio to remove low-quality peptide sequences.
Replication	The L-/D-trypsin was synthesized consistently. All experiments were performed two or three times independently as described in the figure legends. All replication attempts were successful, with a representative result shown in certain cases.
Randomization	This study did not involve samples being allocated into experimental groups, and thus statistical hypothesis issues related to randomization did not apply.
Blinding	This study did not involve experiments where the outcome would be influenced by blinding, and thus statistical hypothesis issues related to blinding did not apply.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging