ACS central science

http://pubs.acs.org/journal/acscii

Research Article

Accurate Models of Substrate Preferences of Post-Translational Modification Enzymes from a Combination of mRNA Display and Deep Learning

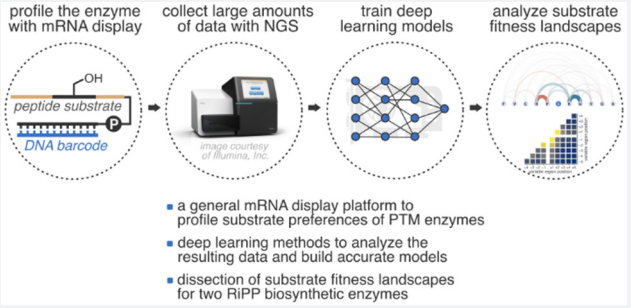
Alexander A. Vinogradov,* Jun Shi Chang, Hiroyasu Onaka, Yuki Goto, and Hiroaki Suga*

Cite This: *ACS Cent. Sci.* 2022, 8, 814–824

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Promiscuous post-translational modification (PTM) enzymes often display nonobvious substrate preferences by acting on diverse yet well-defined sets of peptides and/or proteins. Understanding of substrate fitness landscapes for PTM enzymes is important in many areas of contemporary science, including natural product biosynthesis, molecular biology, and biotechnology. Here, we report an integrated platform for accurate profiling of substrate preferences for PTM enzymes. The platform features (i) a combination of mRNA display with next-generation sequencing as an ultrahigh throughput technique for data acquisition and (ii) deep learning for data analysis. The high accuracy (>0.99 in each of two studies) of the resulting deep learning models enables comprehensive analysis of enzymatic substrate preferences. The models can quantify fitness across sequence space, map modification sites, and identify important amino acids in the substrate. To benchmark the platform, we performed profiling of a Ser dehydratase (LazBF) and a Cys/Ser cyclodehydratase (LazDEF), two enzymes from the lactazole biosynthesis pathway. In both studies, our results point to complex enzymatic preferences, which, particularly for LazBF, cannot be reduced to a set of simple rules. The ability of the constructed models to dissect such complexity suggests that the developed platform can facilitate a wider study of PTM enzymes.



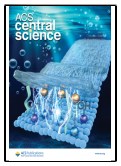
INTRODUCTION

Enzymes which perform post-translational modification (PTM) of peptides and proteins often display nontrivial preferences by acting on a wide range of substrates. The nuanced and, in many cases, poorly understood nature of substrate recognition and engagement by PTM enzymes has come under scrutiny in several contexts. For example, during the biosynthesis of ribosomally synthesized and post-translationally modified peptides (RiPPs),[1,2] notably, lanthipeptides[3,4] and cyanobactins,[5] a single set of PTM enzymes can modify disparate substrates to assemble multiple natural products.[6,7] Catalytic promiscuity of RiPP biosynthetic enzymes suggests numerous bioengineering applications,[8−11] and accordingly, much effort has been dedicated to understanding the molecular basis for such behavior.[12−17] Likewise, in human biology, dense PTM networks controlled by hundreds of promiscuous enzymes orchestrate virtually every aspect of cell behavior, and thus, investigating how PTM enzymes discriminate their substrates is integral to form a holistic appreciation of cellular processes.[18−20]

Substrate specificity profiling studies are a natural first step when studying catalysis by promiscuous PTM enzymes. Numerous approaches developed over the years[21−25] enable streamlined analysis of substrate fitness landscapes, but every method comes with its own limitations. Platforms based on the screening of synthetic peptide microarrays[26−30] and saturation mutagenesis approaches[31−33] have relatively low throughput and can suffer from limited generalizability and accuracy. For example, microarray-derived substrate preferences of sirtuin lysine deacetylases mismatch known cellular substrates.[34] In vivo library construction methods, particularly yeast and phage display, offer a much higher throughput (up to ~10^9 peptides for testing compared to 10^3−10^4 for microarrays), but developing experimental schemes for phage/yeast display can be technically difficult, and these approaches to date have mainly focused on studying proteases.[35−37]

Inference from large amounts of data is another challenge common for high-throughput methods. The de facto standard approach is the computation of position-wise amino acid enrichment scores (usually displayed as WebLogo sequence alignment plots),[38] which overcompresses available information and inevitably loses the nuance. Machine learning/deep learning

Received: February 27, 2022
Published: May 26, 2022



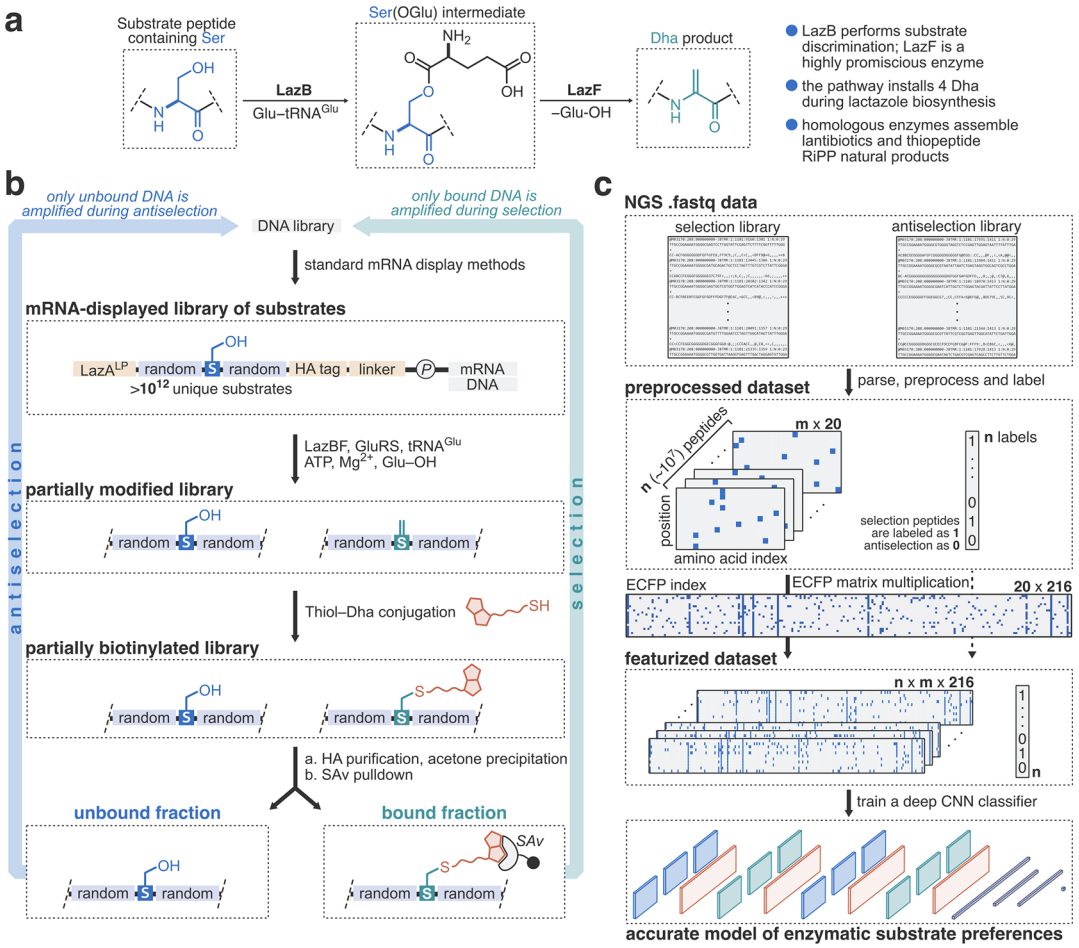


Figure 1. An overview of the workflow for the profiling of LazBF substrate preferences. (a) Chemical reaction catalyzed by LazBF. (b) Schematic overview of mRNA display-based selection/antiselection setups. For the full protocol, see Supporting Information 2.3. Ⓟ refers to the puromycin linker used to display the peptides onto cognate mRNAs. Both selection and antiselection assays can be repeated several times to produce libraries of progressively increasing (or decreasing) substrate fitness. (c) Schematic overview of the data analysis pipeline. NGS selection and antiselection data sets are parsed, preprocessed, and labeled. Peptides are represented as positionally encoded matrices of ECFPs, and a supervised CNN classifier is trained on the resulting data to produce models of LazBF substrate preferences. For a complete description of the data analysis pipeline, see Supporting Information 2.5.

methods represent a promising alternative to the purely statistical treatment of data. Deep learning has in recent years proven its ability to make meaningful generalizations in a variety of complex tasks, but it requires large amounts of clean, curated data to train and evaluate the models.[39,40] To date, the substrate profiling studies which utilized deep learning were either data-limited, due to their reliance on peptide microarrays for data acquisition,[41] or used heterogeneous data sets,[42−45] which have led to models with modest predictive power.

Messenger RNA (mRNA) display-based enzyme-profiling strategies[46,47] have recently gained traction as a viable alternative to the established methods. As a fully in vitro platform, mRNA display can access combinatorial libraries of vast diversity (>10^{12} unique peptides).[48,49] The technique also allows for elaborate manipulation of the libraries (extensive genetic code reprogramming, affinity purification, chemical labeling, etc.) and therefore supports the development of workflows inaccessible to in vivo methods. Still, mRNA display approaches have thus far revolved around single-point saturation mutagenesis[46,47] and, as such, have typically profiled only hundreds of enzyme substrates at once, not taking full advantage of the available library diversity.

Here, we report the development of a general platform for assaying substrate fitness landscapes of PTM enzymes (Figure 1). Our approach integrates mRNA display as the data-generating engine with deep learning workflows to analyze and learn from the resulting data. Using two RiPP biosynthetic enzymes catalyzing distinct reactions, we demonstrate that mRNA display-based substrate selections can provide large amounts of clean, labeled data to train supervised deep learning classifiers of enzymatic substrate preferences. The resulting models accurately predict substrate fitness from a primary sequence and generalize well across the peptide sequence space. The models have a degree of interpretability that allows for mapping of modification sites and identification of important residues in the substrate. Altogether, we believe that the described pipeline is a powerful tool for studying the dynamics of PTM enzyme/substrate interactions.

RESULTS AND DISCUSSION

Development of the mRNA Display Scheme for LazBF Profiling. For this study, we focused on PTM enzymes participating in the biosynthesis of lactazole A,[50] a natural product belonging to the thiopeptide family of RiPPs.[51] The compound is a promising bioengineering target because its five biosynthetic enzymes (LazBCDEF) can convert a wide variety

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

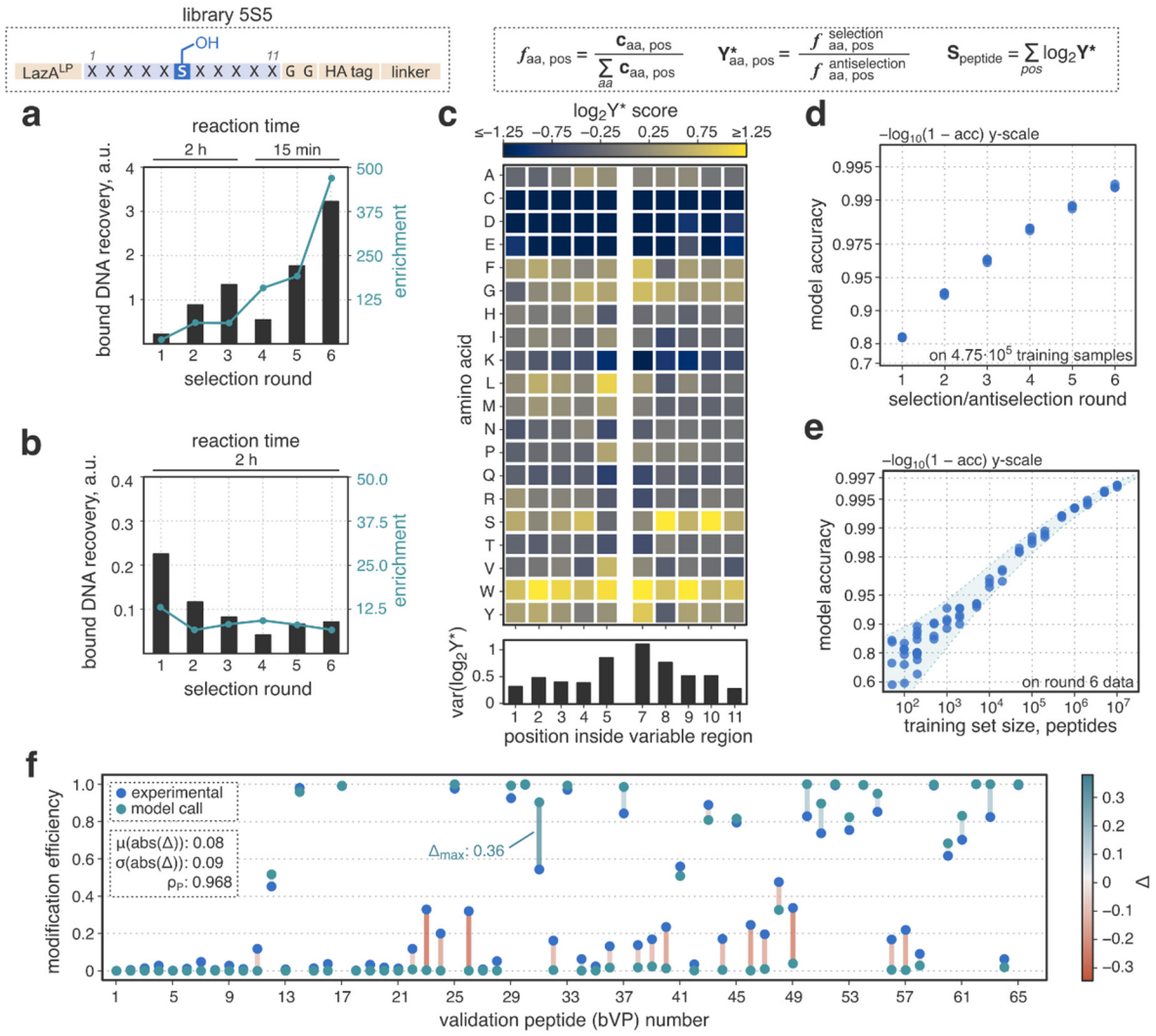


Figure 2. mRNA display profiling of LazBF leads to enriched peptide populations suitable for deep learning applications. (a, b) Summary of the selection (a) and antiselection (b) experiments. Plotted are respective DNA recovery and enrichment values measured by qPCR after every round of mRNA display. (c) Data set convergence at the amino acid level as measured by $\log_2 Y^*$ scores. Amino acid *aa* in position *pos* is enriched in the selection data set compared to the antiselection one if $\log_2 Y^*_{aa,\,pos}$ is greater than 0. See also the definitions in the figure header and Supporting Information 2.1; $c_{aa,\,pos}$ is the number of NGS reads with amino acid *aa* in position *pos* in a data set. (d) CNN classifier accuracy as a function of the number of mRNA display rounds. The models were trained on 4.75×10^5 samples from the respective data sets, using 0.25×10^5 unseen samples for validation. Multiple rounds of mRNA display lead to cleaner data sets and, hence, more accurate models. (e) CNN classifier accuracy as a function of the training data set size. The models were trained on round 6 data. Model accuracy scales with the size of the training data set. (f) Validation of model predictions against experimental data. 65 validation peptides (bVP1−65; all encoded in library 5S5; see also Table S4) were expressed by the FIT system and treated with LazBF/GluRS/tRNA^{Glu} for 2 h. Reaction outcomes were analyzed by LC-MS as described in Supporting Information 2.8. Model predictions showed good agreement with the experiment.

of sequence-randomized precursor peptides to lactazole-like thiopeptides.[47,52] LazBF, a split Ser dehydratase homologous to class I lanthipeptide synthetases (Figure 1a),[3] plays a central role during lactazole biosynthesis because its operation to install four dehydroalanine (Dha) residues into precursor peptide LazA requires precise timing and selectivity.[53] Mechanistically, LazBF operates via a two-step process akin to class I Ser/Thr dehydratases.[54,55] The glutamylation domain (LazB) binds the N-terminal leader peptide (LP) region of LazA (LazA^{LP}) and promotes Ser glutamylation in the downstream core peptide (CP) section using Glu-tRNA^{Glu} as the acyl donor. In the second step, the elimination domain (LazF) catalyzes a retro-Michael reaction in the Ser(OGlu) intermediate to yield the Dha-containing product.[56] Although preliminary enzyme characterization indicated that LazBF prefers a Trp residue in position +1 relative to the modification site, the enzyme also displayed more elaborate preferences which eluded generalization.[47,53] Here, we sought to develop an mRNA display/deep learning-based platform for comprehensive profiling of LazBF substrate fitness landscapes.

We envisioned training a supervised learning classifier that could predict the fitness of LazBF substrates from their primary sequence. To that end, the acquisition of two mRNA display data sets (one corresponding to substrates of high fitness and another for peptides which are not dehydrated by LazBF) was necessary. We anticipated that the treatment of a diverse library of mRNA-displayed peptides with LazBF would dehydrate some, but not all, library members (Figure 1b). The modified peptides, i.e., those containing a Dha residue, are reactive toward thiols[57] and can be selectively conjugated to a biotinylated probe

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

(biotin-PEG_2-SH; Figure S1d). The labeling reaction enables the separation of modified and unmodified substrates using a streptavidin (SAv) pulldown, which selectively isolates biotinylated products. The subsequent PCR amplification of either the SAv-bound or unbound fraction recovers DNA libraries encoding peptides of increased or decreased fitness, respectively. Iterative repetition of this process should produce increasingly enriched peptide populations. During a "selection", SAv-bound DNA is amplified to enrich for substrates of high fitness, while an "antiselection" recovers the unbound DNA fraction to generate a data set of poor substrates.

To establish the assay, we designed an mRNA library encoding peptides bearing the $LazA^{LP}$ sequence followed by a randomized CP, HA tag for affinity purification and a flexible C-terminal linker (library 5S5; Figure 2). Every CP contained a potentially modifiable Ser residue flanked by five random amino acids on either side (theoretical diversity: 1×10^{13} sequences) to establish substrate recognition requirements around the dehydration site. Our preliminary experiments indicated that library 5S5 contained substrates of differential fitness. First, we selected three such peptides (bAP1−3, in order of decreasing fitness; Figure S1) to establish the experimental conditions. The treatment of the peptides expressed by the flexible in vitro translation (FIT) system[58] with 2 μM LazBF, 20 μM $tRNA^{Glu}$, and 1 μM GluRS for 2 h led to the quantitative dehydration of bAP1, partial modification of bAP2, and virtually no reaction for bAP3 (Figure S1a−c). Further incubation of the reaction products with 5 mM biotin-PEG_2-SH at pH 8.5 on ice for 17 h resulted in selective and nearly quantitative biotinylation of Dha-containing peptides, indicating the feasibility of the envisioned experimental scheme (Figure S1e).

Next, we tested whether these substrates could be differentiated in an mRNA display format. Peptide-mRNA/DNA chimeras prepared following the standard techniques (Supporting Information 2.2 and 2.3)[49] were modified under the aforementioned conditions, and, following SAv pulldown, the amount of DNA in either bound or unbound fraction was quantified by qPCR. A large difference in DNA recovery between bAP1 and bAP3 was observed (r; defined as the ratio of DNA in the bound over the unbound fractions; r_{bAP1} = 4.6 vs r_{bAP3} = 0.008; Figure S1f) but only when intermediate HA-affinity purification and acetone precipitation steps (aimed to eliminate unreacted biotin-PEG_2-SH and mRNAs that failed to display peptides) were included (data not shown). Enrichment, defined as the ratio of DNA recovery in the experiment over the negative control (an analogous assay where LazB is omitted from the enzyme mix), also pointed to the enzyme-dependent DNA recovery in the bound fraction ($enrichment_{AP1}$ = 223 vs $enrichment_{AP3}$ = 1.2; Figure S1f). Combined, these data indicate that the developed mRNA display pipeline can discriminate LazBF substrates based on their fitness.

With these protocols, we performed six rounds of selection and antiselection for library 5S5 following the established conditions, except, starting with round 4 of the selection experiment, the LazBF incubation time was shortened to 15 min to adjust selection pressure. The enrichment values increased between rounds during the selection experiment (Figure 2a), suggesting that the substrate populations of progressively higher fitness were obtained. In contrast, for antiselection, after the initial decrease in round 2, DNA recovery and enrichment remained relatively constant (Figure 2b). Next-generation sequencing (NGS) of the resulting DNA libraries revealed that, even after six rounds of selection/antiselection, the substrate populations remained highly diverse and had no convergence at the peptide level, which stands in contrast to traditional affinity-based mRNA display selection workflows (Figure S2). To analyze convergence at the amino acid level, we computed $Y^*_{aa,\ pos}$ scores as a measure of enrichment for amino acid *aa* in position *pos* in the selection data set compared to the antiselection one (Figure 2c). Thus, amino acid *aa* in position *pos* appears to be favored by the enzyme if its log transformed Y* score, $\log_2 Y^*$, is greater than zero, and disfavored when $\log_2 Y^* < 0$. This analysis recapitulated our previous[47,53] findings: for example, Trp in position 7, i.e., position +1 to the fixed Ser residue, had the highest Y* score ($\log_2 Y^* = 2.53$), whereas Asp and Glu, which are known to be disfavored by LazBF,[47,53] had a negative $\log_2 Y^*$ in every position. Overall, the amino acids around the designed modification site (positions 5 and 7) were subject to a stronger discrimination than those further away from Ser6 (compare position-wise variances of $\log_2 Y^*$ scores; Figure 2c). For any library member, a statistical fitness score, S, can be computed as the sum of $\log_2 Y^*$ for every amino acid in the variable region. We found that representing peptides in the S-space is an effective way to perform data set-wide analysis of substrate populations. For example, consistent with the qPCR results, this analysis revealed (Figure S3) that the selection generated a highly enriched substrate subpopulation (1.7σ higher than the naïve library), whereas the antiselection did not because the antiselection peptides resembled the naïve library members ($\Delta 0.5\sigma$). Altogether, these data suggest that the assay produced enriched yet highly diverse substrate populations suitable for further analysis.

Development and Validation of Deep Learning Models. Next, we turned to the development of a deep learning workflow (Figure 1c). We sought a scalable and generalizable pipeline to build interpretable models which can facilitate downstream enzymatic studies. After considerable experimentation, we opted for a straightforward data preprocessing routine: NGS data were in silico translated, denoised, and demultiplexed, after which the resulting peptide data sets were labeled (Supporting Information 2.5; Table S3). All selection and antiselection peptides received a label of "1" and "0", respectively. A number of more sophisticated workflows, which included data preclustering, outlier detection, or quantification of relative fitness from NGS read counts, were rejected as they consistently led to models of a poorer performance. Peptide sequences were represented as matrices of positionally encoded amino acid-wise extended connectivity fingerprints (ECFPs; Supporting Information 2.5, Figures S4 and S5),[59] a technique that has been recently applied to train models which take peptide sequences as input data.[60,61] ECFP representations are built directly from the chemical structures of constituent amino acids, and thus, they bypass the limitation of many popular approaches based on biophysical descriptors,[62,63] which are typically limited to 20 proteinogenic amino acids. At the same time, ECFP representations are more interpretable than one-hot encoding and related techniques. A deep convolutional neural network (CNN; 2.5 million parameters; Supporting Information 2.5 and Figure S6) was selected as the model architecture, primarily due to its fast training. However, we note that neither the choice of the model architecture (also tested were recurrent networks, transformers, and fully connected networks) nor the nature of peptide representation was particularly critical from the accuracy perspective.

With these methods, we turned to benchmarking the overall workflow. First, we ascertained whether multiple rounds of

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

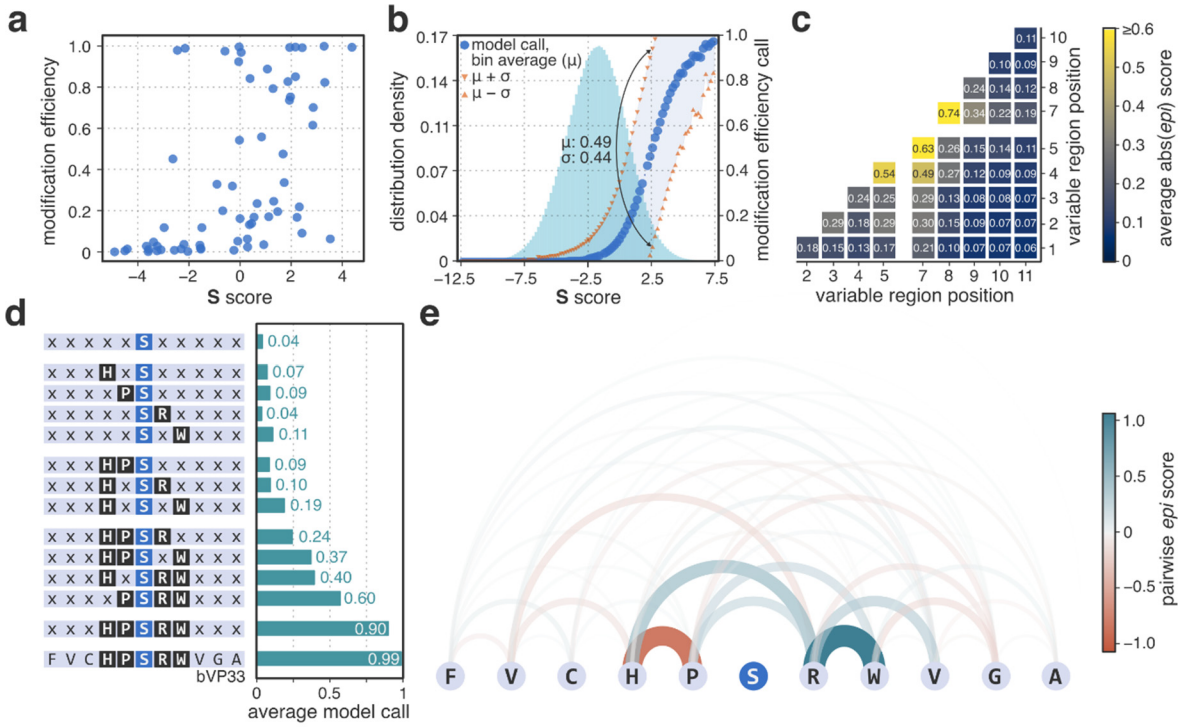


Figure 3. Model enables high-level analysis of LazBF substrate fitness landscapes. (a) Experimentally measured modification efficiencies of validation peptides (bVP1−65; Table S4) as a function of their *S* scores. *S* scores cannot be used to reliably predict fitness of bVP peptides. (b) Distribution of model predictions in the *S*-space. Substrate fitness of 5×10^6 random 5S5 peptides was evaluated with the model. Plotted are binned statistics of model predictions in the S-space. The overall distribution of the peptides in the same space is displayed for reference. The analysis reveals that at best *S* scores can be reliably used as antideterminants of substrate fitness (when $S < -5$). (c) Pairwise epistasis between variable positions in the CP of 5S5 peptides. The model was utilized to compute abs (*epi*) scores using predictions for 5×10^6 sequences from b). The resulting values can be used to estimate how strongly amino acids in the substrate affect each other's fitness. Higher values correspond to stronger second-order effects. See Supporting Information 2.1 for computation details. (d) Analysis of epistatic interactions in bVP33. Average model calls were computed for 2×10^4 partially random in silico generated peptides in each case; "*x*" denotes any amino acid except Ser. (e) Visualization of all pairwise epistatic interactions in bVP33. Strong epistasis inside the His4-Pro5-Ser6-Arg7-Trp8 motif contributes to the high fitness of the peptide.

mRNA display were important by training CNN models on NGS data for each selection/antiselection round using 4.75×10^5 and 0.25×10^5 samples for training and validation, respectively (Figure 2d). Model accuracy increased from 0.823 for round 1 data to 0.992 for round 6, indicating that multiple rounds of mRNA display can furnish progressively cleaner data sets for deep learning. The amount of training data also proved important. Although reasonable models could be trained on as few as 10^2 peptides from the round 6 data set (Figure 2e), the log−log plot of the accuracy versus the number of training samples was nearly linear, with model accuracy reaching up to 0.997 when 10^7 peptides were used for training. Notably, saturation in method performance was not observed in either experiment, which suggests that running more rounds of mRNA display and/or increasing the sequencing depth could further improve the accuracy of the method. The latter approach might be particularly straightforward because the throughput of contemporary NGS instruments reaches 10^{10} reads/day.[64] We also benchmarked our workflow against several traditional machine learning methods (*k*-nearest neighbors, adaptive/gradient boosting, logistic regression, and random forest classifiers) and found that deep neural networks were consistently superior (Figure S7a).

The experiments above evaluated model performance in simple classification tasks where a model is tasked with assigning library 5S5 peptides as belonging to either the selection or antiselection data sets, with NGS data used as the ground truth. In the final experiment, we evaluated whether the models could make more biochemically meaningful predictions, i.e., whether they generalize beyond NGS data and agree with experimentally determined substrate fitness values. To this end, we semi-randomly selected 65 library 5S5 members to ensure a fair test of the model performance ("validation peptide set", bVP1−65; see Supporting Information 2.6 for sequence choices and Table S4) and experimentally investigated their dehydration by LazBF in batch format. The peptides expressed by the FIT system were incubated with LazBF/$tRNA^{Glu}$/GluRS for 2 h under the same conditions as for the mRNA display pipeline. Reaction outcomes were quantified by LC-MS and summarized as modification efficiency values (see Supporting Information 2.8 for details). The model training pipeline was modified to exclude all validation peptide sequences from the training data set using a Hamming distance ≤ 2 as the cutoff value. Overall, we found that the model predictions tracked the experimental values (Pearson correlation coefficient, $\rho_P = 0.968$; Figure 2f), indicating that despite being trained as a classifier, the model could also quantify substrate fitness with the mean prediction error of 0.08 ± 0.09 ($\pm\sigma$). The ability to quantify substrate fitness was in line with the model's performance on NGS data sets; that is, the quantification accuracy depended on the amount of training data and the number of mRNA display rounds (Figure S8a,b). The model excelled at identifying high fitness substrates, whereas underprediction of reaction yields for

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

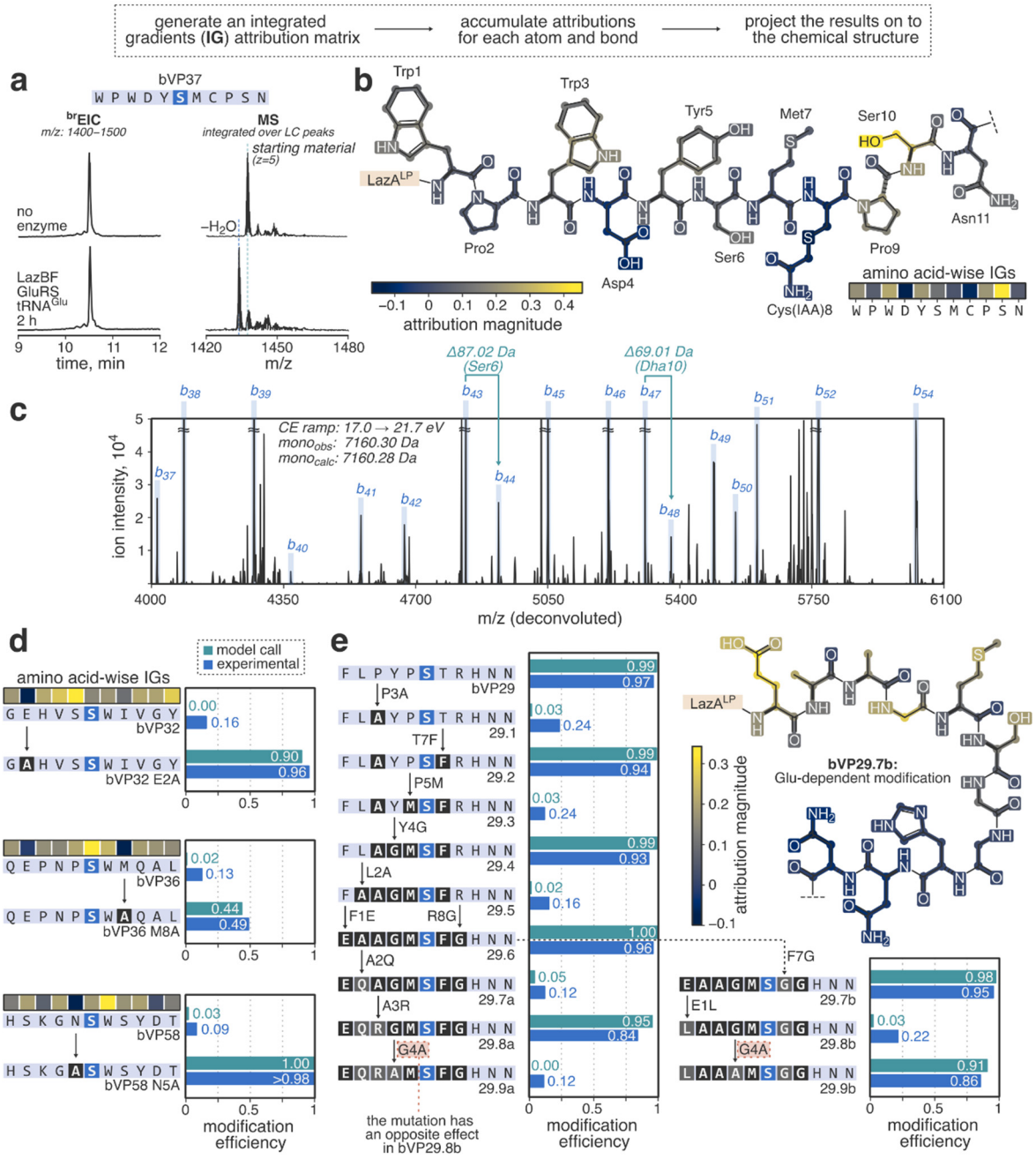


Figure 4. Model-guided dissection of the substrate preferences of LazBF. (a) LC-MS analysis of bVP37 dehydration by LazBF [a broad extracted ion chromatogram (^{br}EIC) and a composite MS spectrum integrated over substrate-derived peaks showing the overall product distribution; see Supporting Information 2.8 for LC-MS details]. (b) Atom- and bond-wise accumulated IG attributions for bVP37. The model suggests that Ser10 is the primary determinant of the high modification efficiency. (c) A zoomed-in section of a charge-deconvoluted CID fragmentation spectrum for singly dehydrated bVP37; *y*-ion assignments and neutral molecule losses are omitted for clarity. The spectrum allows unambiguous assignment of the dehydration site to Ser10, consistent with the model's suggestion. See Figures S10–12 for more examples. (d) Amino acid-wise IGs provide an intuition for relative amino acid contributions to the total substrate fitness. Experimentally measured increase in modification efficiency for three single-point mutants of bVP32, 36, and 58 underscores the model's ability to identify amino acids critical for LazBF-mediated dehydration. See Figure S13 for more examples. (e) Substrate space traversal study for bVP29 (see also the accompanying text). The model was employed to find a sequence of bVP29 mutants which alter the substrate fitness at each step. The route identified by the model was validated experimentally. Collectively, this study points to the complex and unintuitive substrate preferences of LazBF.

moderately poor peptides (those with modification efficiencies of ∼0.05−0.15) was the most common source of error.

Altogether, these data demonstrate that the developed mRNA display/deep learning platform can produce accurate models capable of extrapolating substrate fitness across the peptide sequence space. In the subsequent series of experiments, we deployed the model to understand the high-level features of the LazBF substrate space.

Model-Guided Population-Level Analysis of LazBF Substrates. In striking contrast to the performance of deep learning models, mRNA display-based statistical metrics such as the *S* score bore close to no predictive power for the validation

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

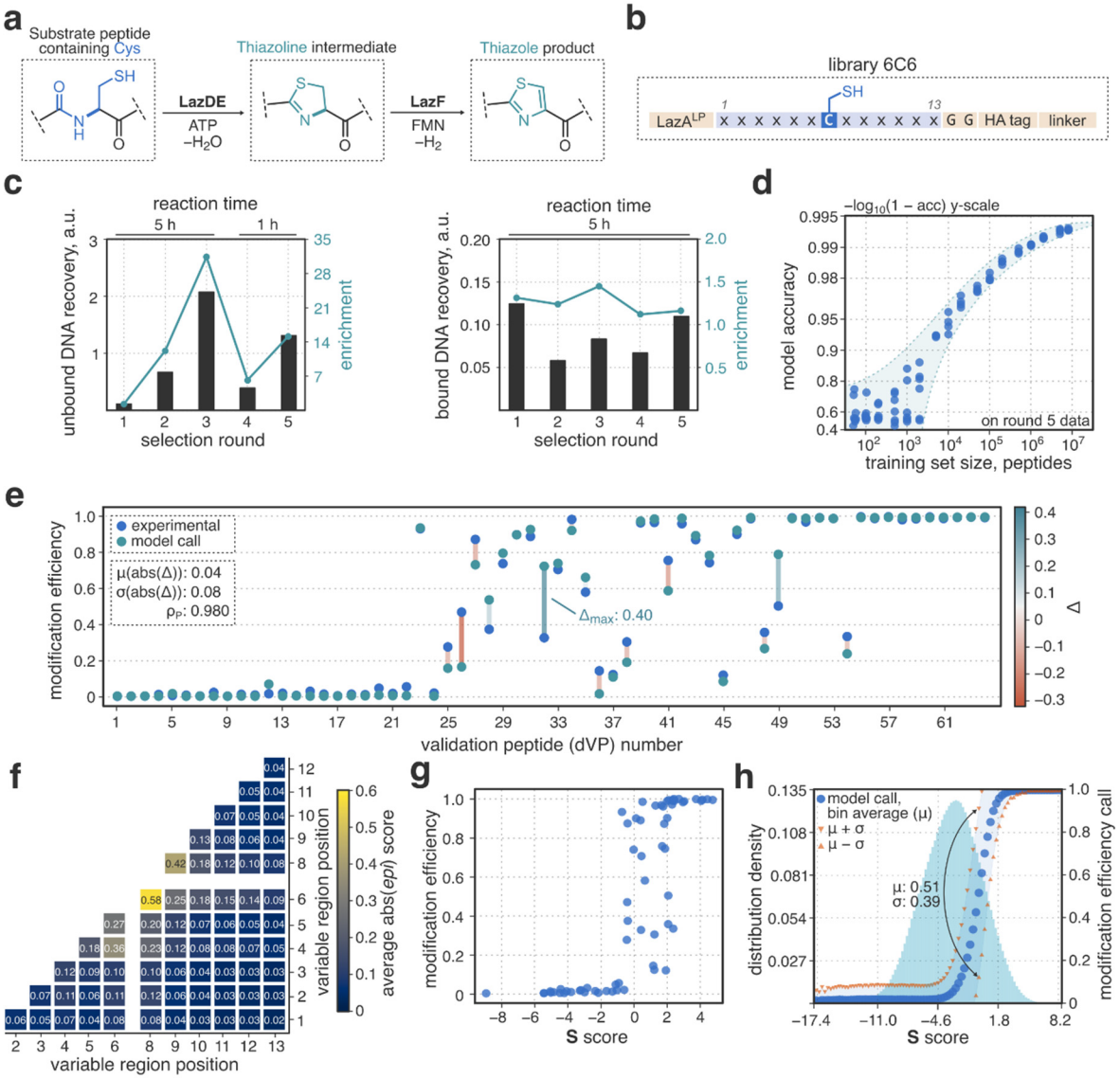


Figure 5. Substrate specificity profiling for LazDEF. (a) Chemical reactions catalyzed by LazDEF. (b) Design of the LazDEF substrate library, library 6C6. (c) Summary of the selection and antiselection experiments. Plotted are respective DNA recovery and enrichment values measured by qPCR after every round of mRNA display. (d) CNN classifier accuracy as a function of training data set size. The models were trained on round 5 data. (e) Validation of model predictions against experimental data. A total of 64 validation peptides (dVP1−64; Table S5) were expressed by the FIT system and treated with LazDEF for 5 h. Reaction outcomes were analyzed by LC-MS as described in Supporting Information 2.8. Model predictions show good agreement with the experiment. (f) Pairwise epistasis between variable positions in the CP of 6C6 peptides. The model was utilized to compute abs(*epi*) scores using predictions for 5×10^6 sequences from panel h). The resulting values can be used to estimate how strongly amino acids in the substrate affect each other's fitness. Higher values correspond to stronger second-order effects. Compared to the results for LazBF, LazDEF substrates are characterized by weaker pairwise epistatic interactions, which aids in explaining the results in panels (g) and (h). See Supporting Information 2.1 for computation details. (g) Experimentally measured modification efficiencies of validation peptides as a function of their *S* scores. Compared to the LazBF results (Figure 3a), the *S* scores for LazDEF substrates prove more informative. (h) Distribution of model predictions in the *S*-space. Substrate fitness of 5×10^6 random 6C6 peptides was evaluated with the model. Plotted are binned statistics of model predictions in the *S*-space. The overall distribution of the peptides in the same space is displayed for reference. In the interval $[-3, 2]$, which accounts for 46% of the total peptide space, *S* scores are an unreliable metric of substrate fitness.

peptide set (Figure 3a). To see whether this is generally true for LazBF substrates, we generated 5×10^6 random peptides from library 5S5 in silico and estimated their fitness using the model. The analysis of the distribution of model predictions in the S-space demonstrated that statistical enrichments could confidently point to a small fraction of poor substrates ($S < -5$), but for high fitness peptides, the uncertainty of the prediction was too high to be practically useful (Figure 3b). For example, an average peptide with $S = 2.5$ had a predicted modification efficiency of 0.49 ± 0.45 ($\pm\sigma$). Representing the outcomes of high-throughput enzyme-profiling experiments as positional amino acid preferences is a common practice. Our results (see also the data for LazDEF below) suggest that at least for some RiPP enzymes such a practice should be exercised with caution, although it remains to be established how general this phenomenon is.

Statistically, poor performance of *S* scores in predicting substrate fitness points to prevalent higher-order effects; i.e., the

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

fitness of an amino acid in a given position strongly depends on the rest of the substrate sequence and should not be treated as an independent variable. To quantify some of these effects, we employed the model to compute pairwise epistasis between substrate amino acids in various positions and summarized the results as *epi* score values (for definition, see Supporting Information 2.1). A positive *epi* score corresponds to a synergistic effect between amino acids *aa1* and *aa2* in positions *pos1* and *pos2*, respectively. Conversely, a negative *epi* score indicates that on average a substrate containing *aa1* and *aa2* in positions *pos1* and *pos2* has a lower-than-expected fitness if statistical independence of amino acids was assumed (see Figure S9 for examples). Averaging of absolute *epi* scores over *aa1* and *aa2* can be utilized to estimate how strongly *pos1* and *pos2* influence each other. This analysis showed (Figure 3c) that amino acids around the modification site (positions 4, 5, 7, and 8) have stronger pairwise epistasis than those distal from Ser6, although a number of long-range interactions was still noticeable (for example, position 1 to position 7; *epi* = 0.21). Overall, such second-order effects dominated the substrate fitness landscape for LazBF, which explains why simple amino acid enrichment metrics had limited predictive power. For instance, validation peptide bVP33 underwent near quantitative dehydration by LazBF (0.97) due to the presence of His4-Pro5-Ser6-Arg7-Trp8 motif. Multiple pairwise epistatic interactions within the motif (Figure 3d,e) facilitated substrate fitness despite the low statistical fitness score (S = 0.04), and no single amino acid was solely responsible for the high modification efficiency.

The diversity and abundance of epistatic interactions in LazBF substrates suggest that the enzyme likely makes extensive but transient contacts with the substrate's CP during the two-step catalysis. Despite the variety of high fitness substrates, LazBF is less promiscuous than it might appear, as only 4% of library 5S5 peptides were predicted to undergo efficient dehydration (Figure 3d).

Model-Guided Peptide-Level Analysis of LazBF Substrates. Integrated gradients (IGs) are a popular method for interpreting predictions of deep learning models.[65] As an attribution technique, IGs seek to understand how individual input features affect a particular prediction by the model. Because in our case peptides are represented as a matrix of ECFPs, IGs can be projected onto the chemical structures of input sequences to visualize model attributions at an atomic resolution. We found this technique insightful in two ways. First, IG attributions facilitated the assignment of PTM sites. For several validation set peptides containing multiple Ser residues in the CP region, the treatment with LazBF yielded chromatographically homogeneous singly dehydrated species (see bVP17, 25, 37, and 51 in Figures S10a, S11a, 4a, and S12a, respectively), pointing to selective modification of one Ser residue. For bVP37, the model attributed its high score prediction (0.985) primarily to Ser10 (Figure 4b), while Ser6 was deemed less important, suggesting that the dehydration occurred at the former residue. Tandem mass-spectrometry (MS/MS) of dehydrated bVP37 unambiguously located the modification site to Ser10 (Figure 4c), and similar analysis performed for bVP17, 25, and 51 confirmed that IG attributions can be utilized to predict modification sites (Figures S10, S11, and S12). Second, the technique could also be leveraged to dissect the contributions of individual amino acids toward the overall substrate fitness. For several validation set peptides, amino acid-wise IG attributions designated a single amino acid, often far from the modification site (Figures 4d and S13), as the major reason for a poor dehydration efficiency. Indeed, single-point mutations at specified amino acids improved the experimentally observed substrate fitness in every case.

The model—together with the aforementioned techniques—enabled a detailed evaluation of LazBF's catalytic promiscuity. Ultimately, we found that the complexity of the substrate landscape, as hinted at by the analysis of epistatic interactions, precludes reasonable simplifications to a set of straightforward rules. To illustrate the intricacy of LazBF substrate preferences, we performed a sequence space traversal study for one validation set peptide, bVP29. Specifically, we utilized the model to find a chain of mutations which drastically alter substrate fitness at each step (Figure 4e). The model pointed to numerous inconspicuous amino acid replacements distal from the modification site which either abrogated (for example, L2A mutation in bVP29.4) or restored (A3R in bVP29.7a) LazBF-mediated dehydration at Ser6. Altogether, we found that (i) the presence of an aromatic amino acid next to the modification site or elsewhere in the CP is not absolutely required for modification (bVP29.7b and bVP29.9b); (ii) even though in general negatively charged Glu/Asp in the CP region strongly decrease substrate fitness, some peptides instead rely on the presence of Glu for dehydration (see E1L mutation in bVP29.7b and the corresponding IG attributions); and (iii) analogous mutations in homologous substrates (G4A for bVP29.8a and bVP29.8b) can lead to opposite outcomes.

Given the uncovered complexity of LazBF preferences, and hence the infeasibility of manual annotations of substrate fitness for the enzyme, we argue that the models constructed with our platform represent a powerful tool to facilitate the study of promiscuous lanthipeptide and thiopeptide dehydratases. Our results demonstrate that the substrate preferences for LazBF, as obfuscated as they might be, are discernible, and with enough training data, deep learning can furnish models which are both accurate and generalizable.

LazDEF Profiling. In the final series of experiments, we explored how well the developed platform can be expanded to other PTMs. We chose LazDEF, another component of the lactazole biosynthesis pathway, as the model for this study. LazDE is a YcaO family enzyme[66] which cyclodehydrates Cys and Ser residues in $LazA^{CP}$ during lactazole biosynthesis (Figure 5a) to yield thiazolines and oxazolines, respectively.[52] The dehydrogenase domain of LazF further aromatizes these structures to azoles via FMN-dependent dehydrogenation.[52] As with LazBF, LazDEF is known to process non-native substrates, but the extent of such promiscuity has not been fully elucidated.[47,53]

To profile the enzyme, we designed mRNA display library 6C6 (Figure 5b), where the CP region contained a fixed Cys residue flanked by six random amino acids on either side. To discriminate LazDEF-modified products (i.e., peptides containing a thiazoline/thiazole residue) from unmodified substrates (peptides bearing Cys6), we opted to use iodoacetamide-based chemistry to selectively biotinylate the latter (Figure S14). Thus, the selection protocol was modified to collect and amplify the unbound DNA fraction, while the antiselection amplified SAv pulldown products. In total, we performed five rounds of selection and antiselection (Figure 5c). Consistent with the LazBF study, the selection recovery and enrichment values increased from round to round except for round 4, when the selection stringency was adjusted, whereas antiselection statistics hovered around the same values. Likewise, the resulting sequence populations had strong enrichments at the amino acid

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

level (Figure S15) but did not converge at the peptide level (normalized Shannon entropy, $H_{selection}$ = 0.992), which provided an ample amount of training data for deep learning. Training a CNN classifier on round 5 data led to accurate models, where—similar to the LazBF experiments—the model accuracy was proportional to the number of training samples, reaching up to 0.993 for 8×10^6 input peptides (Figures 5d and S8c), and deep learning-based classifiers also outperformed traditional machine learning methods (Figure S7b). Validation of model predictions against experimentally measured modification efficiency values for 64 peptides confirmed the ability of the model to generalize beyond NGS data sets (Figure 5e, Table S5). The LazDEF model predictions were in good agreement with the experimental values [ρ_P = 0.980; $\mu(\text{abs}(\Delta))$ = 0.04 ± 0.08 ($\pm\sigma$)], indicating that the model might be leveraged for quantitative estimation of LazDEF substrate fitness. Taken together, these results show the flexibility of the developed mRNA display platform and its ability to profile PTM enzymes catalyzing diverse chemical reactions.

In contrast to the similar mRNA display outcomes, LazDEF and LazBF had divergent substrate fitness landscapes. The difference mainly manifested in lower pairwise positional epistasis (compare Figure 5f vs Figure 3c) and, by extension, higher predictive power of statistical fitness scores for LazDEF (Figure 5g,h). Compared to LazBF, analysis of the substrate preferences for LazDEF through the prism of $\log_2$Y* values was more meaningful. Consistent with the prior studies,[47,53] LazDEF primarily relied on amino acids in positions −1 and +1 to discriminate its substrates, preferring small (Gly/Ser/Ala) amino acids on either side of the modification site and strongly disfavored Asp/Glu anywhere in the CP (Figure S15a). However, we note that even in this case, *S* scores could not reliably predict the fitness of nearly half of the total substrate space: 46% of all library 6C6 substrates had *S* scores between −3 and 2 where statistical predictions can be inaccurate (Figure 5h). Accordingly, numerous exceptions to the aforementioned substrate preferences were apparent (Table S5). For example, LazDEF readily modified validation peptide dVP31 (cyclodehydration efficiency: 0.89) which contained Asp adjacent to the modification site (Figure S16).

CONCLUSIONS

Our study demonstrates that mRNA display can produce large amounts of clean, labeled data for supervised deep learning applications. The platform relies on a differential chemical reactivity of enzyme substrates and their reaction products. An mRNA display scheme can be constructed so long as either species can be chemoselectively biotinylated (in this study, reaction products for LazBF and unreactive substrates for LazDEF). We believe that the plethora of contemporary bioconjugation techniques[67,68] will aid the development of analogous workflows for PTM enzymes catalyzing diverse chemical reactions.

Further, we found that highly accurate models of enzymatic substrate preferences of two PTM enzymes catalyzing different reactions can be constructed using a unified pipeline. The resulting classifier models could be employed for quantitative assessment of reaction yields, prediction of modification sites, and to analyze the influence of individual amino acids on the overall substrate fitness. The deep learning workflow proved superior to traditional machine learning methods and to statistical enrichment metrics, commonly used for analysis of high-throughput enzyme-profiling data. Combined, these advances have allowed us to dissect the catalytic preferences of a Ser dehydratase and a YcaO cyclodehydratase, which uncovered unusually complex substrate fitness landscapes in both cases. We believe that the LazBF and LazDEF models will facilitate lactazole bioengineering and, more generally, that the developed platform will foster the study of catalysis by promiscuous PTM enzymes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.2c00223.

Experimental procedures; supplementary figures (PDF)
Supplementary tables (XLSX)
Transparent Peer Review report available (PDF)

AUTHOR INFORMATION

Corresponding Authors

Hiroaki Suga − *Department of Chemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan;* orcid.org/0000-0002-5298-9186; Email: hsuga@chem.s.u-tokyo.ac.jp

Alexander A. Vinogradov − *Department of Chemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan;* orcid.org/0000-0002-8899-0533; Email: a_vin@chem.s.u-tokyo.ac.jp

Authors

Jun Shi Chang − *Department of Chemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan*

Hiroyasu Onaka − *Department of Biotechnology, Graduate School of Agricultural and Life Sciences and Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan*

Yuki Goto − *Department of Chemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan;* orcid.org/0000-0003-4317-0790

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.2c00223

Notes

The authors declare no competing financial interest.

Source Python code for data analysis, trained model weights for LazBF and LazDEF, and instructions to reproduce the work can be found at https://github.com/avngrdv/mRNA-display-deep-learning. NGS data sets were deposited to DDBJ (Accession No. DRA013287). Other data are available from the corresponding authors upon reasonable request.

ACKNOWLEDGMENTS

We thank Yuchen Zhang, Ethan Evans, and Adam Beattie for stimulating scientific discussions related to this work. This work was supported by KAKENHI (JP20K15407 to A.A.V.; JP16H06444 to H.S., Y.G., and H.O.; JP20H05618 to H.S.; and JP20H02866 and JP19K22243 to Y.G.) from the Japan Society for the Promotion of Science.

REFERENCES

(1) Arnison, P. G.; Bibb, M. J.; Bierbaum, G.; Bowers, A. A.; Bugni, T. S.; Bulaj, G.; Camarero, J. A.; Campopiano, D. J.; Challis, G. L.; Clardy, J.; et al. Ribosomally Synthesized and Post-Translationally Modified

Peptide Natural Products: Overview and Recommendations for a Universal Nomenclature. *Nat. Prod. Rep.* **2013**, *30*, 108−160.
(2) Montalbán-López, M.; Scott, T. A.; Ramesh, S.; Rahman, I. R.; van Heel, A. J.; Viel, J. H.; Bandarian, V.; Dittmann, E.; Genilloud, O.; Goto, Y.; et al. New Developments in RiPP Discovery, Enzymology and Engineering. *Nat. Prod. Rep.* **2021**, *38*, 130−239.
(3) Repka, L. M.; Chekan, J. R.; Nair, S. K.; van der Donk, W. A. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem. Rev.* **2017**, *117*, 5457−5520.
(4) Hegemann, J. D.; Süssmuth, R. D. Matters of Class: Coming of Age of Class III and IV Lanthipeptides. *RSC Chem. Biol.* **2020**, *1*, 110−127.
(5) Sivonen, K.; Leikoski, N.; Fewer, D. P.; Jokela, J. Cyanobactins-Ribosomal Cyclic Peptides Produced by Cyanobacteria. *Appl. Microbiol. Biotechnol.* **2010**, *86* (5), 1213−1225.
(6) Li, B.; Sher, D.; Kelly, L.; Shi, Y.; Huang, K.; Knerr, P. J.; Joewono, I.; Rusch, D.; Chisholm, S. W.; van der Donk, W. A. Catalytic Promiscuity in the Biosynthesis of Cyclic Peptide Secondary Metabolites in Planktonic Marine Cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 10430−10435.
(7) Donia, M. S.; Ravel, J.; Schmidt, E. W. A Global Assembly Line for Cyanobactins. *Nat. Chem. Biol.* **2008**, *4*, 341−343.
(8) Yang, X.; Lennard, K. R.; He, C.; Walker, M. C.; Ball, A. T.; Doigneaux, C.; Tavassoli, A.; van der Donk, W. A. A Lanthipeptide Library Used to Identify a Protein-Protein Interaction Inhibitor. *Nat. Chem. Biol.* **2018**, *14*, 375−380.
(9) Schmitt, S.; Montalbán-López, M.; Peterhoff, D.; Deng, J.; Wagner, R.; Held, M.; Kuipers, O. P.; Panke, S. Analysis of Modular Bioengineered Antimicrobial Lanthipeptides at Nanoliter Scale. *Nat. Chem. Biol.* **2019**, *15*, 437−443.
(10) Cebrián, R.; Macia-Valero, A.; Jati, A. P.; Kuipers, O. P. Design and Expression of Specific Hybrid Lantibiotics Active Against Pathogenic Clostridium Spp. *Front. Microbiol.* **2019**, *10*, 2154.
(11) Urban, J. H.; Moosmeier, M. A.; Aumuller, T.; Thein, M.; Bosma, T.; Rink, R.; Groth, K.; Zulley, M.; Siegers, K.; Tissot, K.; Moll, G. N.; Prassler, J. Phage Display and Selection of Lanthipeptides on the Carboxy-Terminus of the Gene-3 minor Coat Protein. *Nat. Commun.* **2017**, *8*, 1500.
(12) Song, I.; Kim, Y.; Yu, J.; Go, S. Y.; Lee, H. G.; Song, W. J.; Kim, S. Molecular Mechanism Underlying Substrate Recognition of the Peptide Macrocyclase PsnB. *Nat. Chem. Biol.* **2021**, *17*, 1123−1131.
(13) Miller, F. S.; Crone, K. K.; Jensen, M. R.; Shaw, S.; Harcombe, W. R.; Elias, M. H.; Freeman, M. F. Conformational Rearrangements Enable Iterative Backbone N-Methylation in RiPP Biosynthesis. *Nat. Commun.* **2021**, *12*, 5355.
(14) Le, T.; Jeanne Dit Fouque, K.; Santos-Fernandez, M.; Navo, C. D.; Jiménez-Osés, G.; Sarksian, R.; Fernandez-Lima, F. A.; van der Donk, W. A. Substrate Sequence Controls Regioselectivity of Lanthionine Formation by ProcM. *J. Am. Chem. Soc.* **2021**, *143*, 18733−18743.
(15) Tang, W.; Jiménez-Osés, G.; Houk, K. N.; van der Donk, W. A. Substrate Control in Stereoselective Lanthionine Biosynthesis. *Nat. Chem.* **2015**, *7*, 57−64.
(16) Habibi, Y.; Uggowitzer, K. A.; Issak, H.; Thibodeaux, C. J. Insights into the Dynamic Structural Properties of a Lanthipeptide Synthetase Using Hydrogen-Deuterium Exchange Mass Spectrometry. *J. Am. Chem. Soc.* **2019**, *141*, 14661−14672.
(17) Dong, S. H.; Liu, A.; Mahanta, N.; Mitchell, D. A.; Nair, S. K. Mechanistic Basis for Ribosomal Peptide Backbone Modifications. *ACS Cent. Sci.* **2019**, *5*, 842−851.
(18) Beltrao, P.; Bork, P.; Krogan, N. J.; Van Noort, V. Evolution and Functional Cross-Talk of Protein Post-Translational Modifications. *Mol. Syst. Biol.* **2013**, *9*, 714.
(19) Tompa, P.; Davey, N. E.; Gibson, T. J.; Babu, M. M. A Million Peptide Motifs for the Molecular Biologist. *Mol. Cell* **2014**, *55*, 161−169.
(20) Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. Post-Translational Modification: Nature's Escape from Genetic Imprisonment and the Basis for Dynamic Information Encoding. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2012**, *4*, 565−583.
(21) Gosalia, D. N.; Salisbury, C. M.; Ellman, J. A.; Diamond, S. L. High Throughput Substrate Specificity Profiling of Serine and Cysteine Proteases Using Solution-Phase Fluorogenic Peptide Microarrays. *Mol. Cell. Proteomics* **2005**, *4*, 626−636.
(22) Meyer, N. O.; O'Donoghue, A. J.; Schulze-Gahmen, U.; Ravalin, M.; Moss, S. M.; Winter, M. B.; Knudsen, G. M.; Craik, C. S. Multiplex Substrate Profiling by Mass Spectrometry for Kinases as a Method for Revealing Quantitative Substrate Motifs. *Anal. Chem.* **2017**, *89*, 4550−4558.
(23) Matthews, D. J.; Goodman, L. J.; Gorman, C. M.; Wells, J. A. A Survey of Furin Substrate Specificity Using Substrate Phage Display. *Protein Sci.* **1994**, *3*, 1197−1205.
(24) Si, Y.; Kretsch, A. M.; Daigh, L. M.; Burk, M. J.; Mitchell, D. A. Cell-Free Biosynthesis to Evaluate Lasso Peptide Formation and Enzyme-Substrate Tolerance. *J. Am. Chem. Soc.* **2021**, *143*, 5917−5927.
(25) Ruffner, D. E.; Schmidt, E. W.; Heemstra, J. R. Assessing the Combinatorial Potential of the RiPP Cyanobactin Tru Pathway. *ACS Synth. Biol.* **2015**, *4*, 482−492.
(26) Kightlinger, W.; Lin, L.; Rosztoczy, M.; Li, W.; Delisa, M. P.; Mrksich, M.; Jewett, M. C. Design of Glycosylation Sites by Rapid Synthesis and Analysis of Glycosyltransferases. *Nat. Chem. Biol.* **2018**, *14*, 627−635.
(27) Ge, Y.; Czekster, C. M.; Miller, O. K.; Botting, C. H.; Schwarz-Linek, U.; Naismith, J. H. Insights into the Mechanism of the Cyanobactin Heterocyclase Enzyme. *Biochemistry* **2019**, *58*, 2125−2132.
(28) Eckhard, U.; Huesgen, P. F.; Schilling, O.; Bellac, C. L.; Butler, G. S.; Cox, J. H.; Dufour, A.; Goebeler, V.; Kappelhoff, R.; Keller, U. a. d.; Klein, T.; Lange, P. F.; Marino, G.; Morrison, C. J.; Prudova, A.; Rodriguez, D.; Starr, A. E.; Wang, Y.; Overall, C. M. Active Site Specificity Profiling of the Matrix Metalloproteinase Family: Proteomic Identification of 4300 Cleavage Sites by Nine MMPs Explored with Structural and Synthetic Peptide Cleavage Analyses. *Matrix Biol.* **2016**, *49*, 37−60.
(29) Kutil, Z.; Skultetyova, L.; Rauh, D.; Meleshin, M.; Snajdr, I.; Novakova, Z.; Mikesova, J.; Pavlicek, J.; Hadzima, M.; Baranova, P.; et al. The Unraveling of Substrate Specificity of Histone Deacetylase 6 Domains Using Acetylome Peptide Microarrays and Peptide Libraries. *FASEB J.* **2019**, *33*, 4035−4045.
(30) Rauh, D.; Fischer, F.; Gertz, M.; Lakshminarasimhan, M.; Bergbrede, T.; Aladini, F.; Kambach, C.; Becker, C. F. W.; Zerweck, J.; Schutkowski, M.; Steegborn, C. An Acetylome Peptide Microarray Reveals Specificities and Deacetylation Substrates for All Human Sirtuin Isoforms. *Nat. Commun.* **2013**, *4*, 2327.
(31) Wooderchak, W. L.; Zang, T.; Zhou, Z. S.; Acuña, M.; Tahara, S. M.; Hevel, J. M. Substrate Profiling of PRMT1 Reveals Amino Acid Sequences That Extend beyond the "RGG" Paradigm. *Biochemistry* **2008**, *47*, 9456−9466.
(32) Young, T. S.; Dorrestein, P. C.; Walsh, C. T. Codon Randomization for Rapid Exploration of Chemical Space in Thiopeptide Antibiotic Variants. *Chem. Biol.* **2012**, *19*, 1600−1610.
(33) Tran, H. L.; Lexa, K. W.; Julien, O.; Young, T. S.; Walsh, C. T.; Jacobson, M. P.; Wells, J. A. Structure-Activity Relationship and Molecular Mechanics Reveal the Importance of Ring Entropy in the Biosynthesis and Activity of a Natural Product. *J. Am. Chem. Soc.* **2017**, *139*, 2541−2544.
(34) Bheda, P.; Jing, H.; Wolberger, C.; Lin, H. The Substrate Specificity of Sirtuins. *Annu. Rev. Biochem.* **2016**, *85*, 405−429.
(35) Pethe, M. A.; Rubenstein, A. B.; Khare, S. D. Data-Driven Supervised Learning of a Viral Protease Specificity Landscape from Deep Sequencing and Molecular Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 168−176.
(36) Zhou, J.; Li, S.; Leung, K. K.; O'Donovan, B.; Zou, J. Y.; DeRisi, J. L.; Wells, J. A. Deep Profiling of Protease Substrate Specificity Enabled by Dual Random and Scanned Human Proteome Substrate Phage Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 25464−25475.

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824

(37) Kretz, C. A.; Tomberg, K.; Van Esbroeck, A.; Yee, A.; Ginsburg, D. High Throughput Protease Profiling Comprehensively Defines Active Site Specificity for Thrombin and ADAMTS13. *Sci. Rep.* **2018**, *8*, 2788.
(38) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188−1190.
(39) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.
(40) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.
(41) Smith, B. C.; Settles, B.; Hallows, W. C.; Craven, M. W.; Denu, J. M. SIRT3 Substrate Specificity Determined by Peptide Arrays and Machine Learning. *ACS Chem. Biol.* **2011**, *6*, 146−157.
(42) Sorimachi, H.; Mamitsuka, H.; Ono, Y. Understanding the Substrate Specificity of Conventional Calpains. *Biol. Chem.* **2012**, *393*, 853−871.
(43) Yu, K.; Zhang, Q.; Liu, Z.; Du, Y.; Gao, X.; Zhao, Q.; Cheng, H.; Li, X.; Liu, Z. X. Deep Learning Based Prediction of Reversible HAT/HDAC-Specific Lysine Acetylation. *Brief. Bioinform.* **2020**, *21*, 1798−1805.
(44) Wang, D.; Zeng, S.; Xu, C.; Qiu, W.; Liang, Y.; Joshi, T.; Xu, D. MusiteDeep: A Deep-Learning Framework for General and Kinase-Specific Phosphorylation Site Prediction. *Bioinformatics* **2017**, *33*, 3909−3916.
(45) Li, F.; Chen, J.; Leier, A.; Marquez-Lago, T.; Liu, Q.; Wang, Y.; Revote, J.; Smith, A. I.; Akutsu, T.; Webb, G. I.; Kurgan, L.; Song, J. DeepCleave: A Deep Learning Predictor for Caspase and Matrix Metalloprotease Substrates and Cleavage Sites. *Bioinformatics* **2020**, *36*, 1057−1065.
(46) Fleming, S. R.; Himes, P. M.; Ghodge, S. V.; Goto, Y.; Suga, H.; Bowers, A. A. Exploring the Post-Translational Enzymology of PaaA by mRNA Display. *J. Am. Chem. Soc.* **2020**, *142*, 5024−5028.
(47) Vinogradov, A. A.; Nagai, E.; Chang, J. S.; Narumi, K.; Onaka, H.; Goto, Y.; Suga, H. Accurate Broadcasting of Substrate Fitness for Lactazole Biosynthetic Pathway from Reactivity-Profiling MRNA Display. *J. Am. Chem. Soc.* **2020**, *142*, 20329−20334.
(48) Kamalinia, G.; Grindel, B. J.; Takahashi, T. T.; Millward, S. W.; Roberts, R. W. Directing Evolution of Novel Ligands by mRNA Display. *Chem. Soc. Rev.* **2021**, *50*, 9055−9103.
(49) Huang, Y.; Wiedmann, M. M.; Suga, H. RNA Display Methods for the Discovery of Bioactive Macrocycles. *Chem. Rev.* **2019**, *119*, 10360−10391.
(50) Hayashi, S.; Ozaki, T.; Asamizu, S.; Ikeda, H.; Omura, S.; Oku, N.; Igarashi, Y.; Tomoda, H.; Onaka, H. Genome Mining Reveals a Minimum Gene Set for the Biosynthesis of 32-Membered Macrocyclic Thiopeptides Lactazoles. *Chem. Biol.* **2014**, *21*, 679−688.
(51) Vinogradov, A. A.; Suga, H. Introduction to Thiopeptides: Biological Activity, Biosynthesis, and Strategies for Functional Reprogramming. *Cell Chem. Biol.* **2020**, *27*, 1032−1051.
(52) Vinogradov, A. A.; Shimomura, M.; Goto, Y.; Ozaki, T.; Asamizu, S.; Sugai, Y.; Suga, H.; Onaka, H. Minimal Lactazole Scaffold for in Vitro Thiopeptide Bioengineering. *Nat. Commun.* **2020**, *11*, 2272.
(53) Vinogradov, A. A.; Shimomura, M.; Kano, N.; Goto, Y.; Onaka, H.; Suga, H. Promiscuous Enzymes Cooperate at the Substrate Level En Route to Lactazole A. *J. Am. Chem. Soc.* **2020**, *142*, 13886−13897.
(54) Ortega, M. A.; Hao, Y.; Zhang, Q.; Walker, M. C.; van der Donk, W. A.; Nair, S. K. Structure and Mechanism of the tRNA-Dependent Lantibiotic Dehydratase NisB. *Nature* **2015**, *517*, 509−512.
(55) Ozaki, T.; Kurokawa, Y.; Hayashi, S.; Oku, N.; Asamizu, S.; Igarashi, Y.; Onaka, H. Insights into the Biosynthesis of Dehydroalaninesin Goadsporin. *ChemBioChem.* **2016**, *17*, 218−223.
(56) Vinogradov, A. A.; Nagano, M.; Goto, Y.; Suga, H. Site-Specific Nonenzymatic Peptide S/O-Glutamylation Reveals the Extent of Substrate Promiscuity in Glutamate Elimination Domains. *J. Am. Chem. Soc.* **2021**, *143*, 13358−13369.
(57) Bogart, J. W.; Bowers, A. A. Dehydroamino Acids: Chemical Multi-Tools for Late-Stage Diversification. *Org. Biomol. Chem.* **2019**, *17*, 3653−3669.
(58) Goto, Y.; Katoh, T.; Suga, H. Flexizymes for Genetic Code Reprogramming. *Nat. Protoc.* **2011**, *6*, 779−790.
(59) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(60) Mohapatra, S.; Hartrampf, N.; Poskus, M.; Loas, A.; Gómez-Bombarelli, R.; Pentelute, B. L. Deep Learning for Prediction and Optimization of Fast-Flow Peptide Synthesis. *ACS Cent. Sci.* **2020**, *6*, 2277−2286.
(61) Schissel, C. K.; Mohapatra, S.; Wolfe, J. M.; Fadzen, C. M.; Bellovoda, K.; Wu, C. L.; Wood, J. A.; Malmberg, A. B.; Loas, A.; Gómez-Bombarelli, R.; Pentelute, B. L. Deep Learning to Design Nuclear-Targeting Abiotic Miniproteins. *Nat. Chem.* **2021**, *13*, 992−1000.
(62) Atchley, W. R.; Zhao, J.; Fernandes, A. D.; Drüke, T. Solving the Protein Sequence Metric Problem. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6395−6400.
(63) Georgiev, A. G. Interpretable Numerical Descriptors of Amino Acid Space. *J. Comput. Biol.* **2009**, *16*, 703−723.
(64) Hu, T.; Chitnis, N.; Monos, D.; Dinh, A. Next-Generation Sequencing Technologies: An Overview. *Hum. Immunol.* **2021**, *82*, 801−811.
(65) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR, 2017; Vol. *70*, pp 3319−3328
(66) Burkhart, B. J.; Schwalen, C. J.; Mann, G.; Naismith, J. H.; Mitchell, D. A. YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function. *Chem. Rev.* **2017**, *117*, 5389−5456.
(67) Koniev, O.; Wagner, A. Developments and Recent Advancements in the Field of Endogenous Amino Acid Selective Bond Forming Reactions for Bioconjugation. *Chem. Soc. Rev.* **2015**, *44*, 5495−5551.
(68) McKay, C. S.; Finn, M. G. Click Chemistry in Complex Mixtures: Bioorthogonal Bioconjugation. *Chem. Biol.* **2014**, *21*, 1075−1101.

https://doi.org/10.1021/acscentsci.2c00223
ACS Cent. Sci. 2022, 8, 814−824