Supplementary Materials for

Enzyme function prediction using contrastive learning

Tianhao Yu *et al*.

Corresponding author: Huimin Zhao, zhao5@illinois.edu

Science **379**, 1358 (2023)
DOI: 10.1126/science.adf2465

The PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S3
References

Other Supplementary Material for this manuscript includes the following:

MDAR Reproducibility Checklist

Materials and Methods

Contrastive losses definition in CLEAN

Two contrastive losses were employed for training: Triplet Margin Loss (*44*) and Supcon-Hard Loss. Triplet Margin Loss sampled an enzyme embedding as the anchor, another enzyme embedding from the same EC class as the positive, and an enzyme embedding from a different EC class as the negative. In contrast to the Triplet Margin Loss, Supcon-Hard Loss samples multiple positives and negatives. In each epoch, we trained the model with one anchor z_e for every EC class $e \in E$. $N(e)$ was the set of hard negative mining examples with respect to the cluster center of e and for the anchor z_e. $P(e)$ was the one or the set of positive samples from the same EC class e for the anchor, followed by $A(e) = N(e) \cup P(e)$. The functions for the two losses were defined as in Equations (1) and (2):

Triplet Margin Loss:

$$\mathcal{L}^{\mathcal{TM}} = \left|\left|z_a - z_p\right|\right|_2 - \left|\left|\mathrm{z_a} - \mathrm{z_n}\right|\right|_2 + \alpha \quad (1)$$

Supcon-Hard Loss:

$$\mathcal{L}^{sup} = \sum_{e \in E} \frac{-1}{|P(e)|} \sum_{z_p \in P(e)} log \frac{\exp(z_e \cdot z_p/\tau)}{\sum_{z_a \in A(e)} \exp(z_i \cdot z_a/\tau)} \quad (2)$$

where α was the margin and set to 1 for all experiments; τ was the temperature parameter and assigned to 0.1 for all experiments. All enzyme embeddings z used in Triplet Margin Loss were the output of the trained network and unnormalized. The embeddings in Supcon-Hard Loss were L2-normalized to unit length. Supcon-Hard Loss was a modification of the supervised contrastive loss (*16*), except that Supcon-Hard Loss used a fixed number of samples per batch and used data points in the batch for negatives and positives, where in Supcon-Hard we sampled a fixed number of positives from the same EC class and hard-mined a fixed number of negatives. The presence of the normalization factor $\frac{1}{|P(e)|}$ served to remove the bias in the positives contributing to the loss. Furthermore, unlike N-Pair Loss (*45*), Supcon-Hard Loss can take an arbitrary number of positives as input, which encourages the trained network to have similar enzyme embeddings for positives z_p.

Mining hard negatives

Hard negatives referred to sequences with EC numbers different from the anchor sequence, but their embeddings were close to the anchor's embeddings on Euclidean distance. Hard negatives were helpful for contrastive learning as they were challenging cases contributing to the loss function. Therefore, hard negatives were sampled from EC numbers with a close distance to the EC number of the anchor sequences by calculating the pairwise distance of EC numbers' cluster centers.

EC selection methods

Contrastive learning essentially established a ranking model, where each EC cluster center was represented by the average of all the enzyme entries that belong to this EC number. Therefore, for any query enzyme, the correct set of EC numbers can be ranked based on Euclidean distances between all the EC cluster centers and themselves. However, because an enzyme could potentially have multiple EC numbers, there needs to be algorithms to select which of the top-ranked EC numbers are considered correct for the query enzyme. We referred to the selection algorithms as EC selection methods. Simply selecting the top-1, top-2 or top-k ranked EC numbers for all query enzymes would not be suitable because selecting the top-1 EC numbers would neglect enzyme promiscuity, but selecting more than that, the precision would drop drastically because the majority of the enzymes only have one EC number. In this work, we propose two EC selection algorithms for performance comparison and determination. The distance ranking information determined the correct cutoff for each query enzyme by 1) *p*-value and 2) Max-Separation. The *p*-value based algorithm was based on statistical significance, where a large quantity of randomly chosen enzymes from the training set were selected as the background distribution of pairwise distances for each EC number. An EC number was only called for the query enzyme if their distance was significantly smaller than random, which was determined by comparing it against the background distribution of distances using a *p*-value cutoff. By choosing different *p*-value cutoffs, users can tune the precision and recall for the prediction. A small *p*-value cutoff made the acceptance threshold tight and was favorable to precision, and a large *p*-value cutoff made the threshold more flexible and profitable to recall. Unlike the *p*-value method, Max-Separation algorithm had no tunable parameters and will give a single prediction. Max-Separation enabled calling of the set of EC numbers with distances close to query enzyme but far from all other EC numbers, mimicking human intuition.

p-value selection method

For the *p*-value algorithm, n (e.g., n = 20,000) randomly chosen enzyme embeddings (extracted from CLEAN) from the training set were used as background. With a selected *p*-value, *p* (e.g., *p* = 0.001) as the threshold, these background embeddings and *p*-value were used to determine whether an EC number should be considered statistically significant. Instead of picking the background uniformly, we weighted the probability of picking an enzyme with EC number with 1/$|EC_i|$, the inverse of the number of enzymes in that EC class. The selected backgrounds' Euclidean distances were recorded with a particular EC cluster center. In this way, a distance matrix between all EC cluster centers in the training set and all backgrounds could be obtained. When a particular query enzyme needed to call the set of EC numbers EC_i, the algorithm started with the EC number EC_0 with the smallest distance s_0, then s_0 was compared with the background's distances with EC_0's cluster center. Suppose the ranking of s_0 in the background distribution is r, EC_0 was called for the query enzyme if r/n is smaller than the *p*-value cutoff.

Max-separation selection method

For the Max-separation algorithm, we assumed there is a background noise distance γ, such that any Euclidean distance $s_{i'}$ between the query sequence and the cluster of an incorrect EC number is close to γ by ε, and any distance s_i between the query sequence and a cluster from the correct set of EC number, EC_i, is far from γ by δ, i.e.,:

$$|s_{i'} - \gamma| \leq \varepsilon,\ \ |s_i - \gamma| \leq \delta, \text{ and } \varepsilon \ll \delta$$

An example of how EC_i can be selected as follows:
The first 10 smallest distances for query q are: [5.6, 6.1, 7.5, 12.22, 12.23, 12.4, 12.5, 12.6, 12.6, 12.7], and the background noise was 12.88. Human intuition will select EC numbers corresponding to distances 5.6, 6.1 and 7.5 as the correct set of ECs for the query. Another assumption was that the sequence S = s_0, s_1, ..., s_{n-1} is long enough that at most 50% of the distances are coming from the correct ECs, that is, $|EC_i| \leq n/2$. The value i needs to be found and subject to the following requirements: 1) $|\ \varepsilon\ - \delta\ |$ is maximized, 2) the $|EC_i|$ is minimized. **Table S1** shows the detailed algorithm for selecting EC_i using Max-Separation.

Evaluation metrics

The evaluation metrics used in the study are precision score, recall score, F1-score, and area under curve (AUC). All metrics were calculated by Python package scikit-learn (*46*). To account for the multi-label setting, a weighted average was used for all studies except for the combined dataset which used a sample average. The scores were obtained by first binarizing the ground truth labels of the testing dataset by scikit-learn and then binarizing the predicted results by various models. The binarized ground truth and predicted results were used as the input according to the scikit-learn documents.

Gaussian mixture model (GMM)

N-component GMM is a probabilistic model that captures data drawn from N different Gaussian distributions with unknown parameters. The two Gaussian components in the GMM characterize the empirical distributions of embedding distances within the same EC number and across different EC numbers, respectively. Intrinsically, the GMM model tests whether the embedding distance between the query protein and the CLEAN-predicted EC number is significantly smaller than the distance between a randomly sampled protein and a randomly sampled EC number. The more significant the test is, the higher confidence score we assign to this function label predicted by CLEAN. To fit the within-EC-number Gaussian distribution of the GMM, we first randomly chose one thousand EC numbers, and for each EC number chosen, we calculated the Euclidean distances between the embedding of EC numbers and the embedding of enzyme sequences with that EC number. These distances generated one Gaussian distribution and forms one of the components in the GMM (**Fig. S11** left peak). We fit the across-EC-number distribution similarly: for each EC number chosen, the distances between EC number and enzyme sequences from a different EC number generated another Gaussian distribution and forms the other component of the GMM (**Fig. S11** right peak). Such procedure has been repeated 40 times to fit 40 GMMs in order to reduce random errors. A similar approach has been used in a previous study to quantify prediction uncertainties (*47*). The python package sklearn.mixture.GaussianMixture was used to implement the fitting of GMM and probability prediction of query samples. The density of the within-EC-number component of the GMM for the query is interpreted as the confidence of the query.

Halogenase dataset annotation and similarity analysis at the protein sequence and structure level

Besides the local version of ProteInfer and DeepEC, the online website BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi), DEEPre (http://www.cbrc.kaust.edu.sa/DEEPre/index.html), ECPred (https://ecpred.kansil.org) and COFACTOR (https://zhanggroup.org/COFACTOR/) with default parameters were used to predict EC numbers of uncharacterized halogenases. The non-redundant protein sequences (nr) database was applied for BLASTp (protein-protein BLAST). The pairwise sequence identity (ID) was calculated by SIAS with BLOSUM62 matrix (http://imed.med.ucm.es/Tools/sias.html). Computing the percentage of identity requires dividing identities by the length of sequence with the following equation:

$$ID_{\%} = 100 \ \times \frac{Identical\ Residues}{Length\ of\ Smallest\ Sequence}.$$

Chemicals and strains

All chemicals were of analytical grade or higher quality and purchased from Sigma-Aldrich (US), and Fisher Scientific (US) and used as received unless stated otherwise. The plasmids pET28a(+)-PH04363 (UniProt ID: O58212), pET28a(+)-MJ1651 (UniProt ID: Q59045), and pET28a(+)-TTHA0338 (UniProt ID: Q5SLF5), pET28a(+)-SsFlA (UniProt ID: W0W999) and pET28a(+)-ScFlA (UniProt ID: Q70GK9) were synthesized from Twist Bioscience HQ (South San Francisco, CA) with similar construction. The gene fragment was cloned into the pET28a(+)-vector via *NcoI* and *XhoI* restriction sites with appending C-terminal His$_6$-tag. The plasmid pAEM7-SalL pAEM7 (UniProt ID: A4X3Q0) was a gift from Bradley Moore (Addgene plasmid # 136422) (*41*). Lysogeny Broth (LB) medium (10 g tryptone, 5 g yeast extract, and 10 g NaCl) and Terrific Broth (TB) medium (12 g tryptone, 24 g yeast extract, 4 g glycerol, 2.31 g $KH_2PO_{4,}$ and 12.54 g K_2HPO_4) were then autoclaved at 121 °C for 20 min. In the case of LB agar 12 plates, 20 g agar was added. Chemically competent *Escherichia coli* DH5a and *E. coli* BL21-Gold (DE3) (New England BioLabs, Inc. Ipswich, MA) were used as hosts for plasmids amplification and protein expression, respectively.

Halogenases dataset setup with EC number annotation

A total of 36 halogenases incompletely annotated in UniProt but reported in the literature were identified to build up the uncharacterized halogenase dataset (**Table S2**). The EC annotations extracted from the literature were collected for all halogenases except the undetermined ones and labeled with the annotation source.

Heterologous expression and purification of experimentally validated enzymes

All three experimental validated enzymes (MJ1651, TTHA0338, and SsFlA) and three known enzymes as the positive controls (PH04363, SalL, and ScFlA) were expressed and purified (**Fig. S3**) as described elsewhere (*48*, *49*). In detail, 5 µL of the glycerol stock strains were inoculated into 5 mL LB_{KANA} medium (100 µg/mL Kanamycin) and pre-cultivated at 37 °C (250 rpm) overnight. After that, 5 mL of pre-culture were transferred into 1 L Erlenmeyer flask containing 150 mL TB_{KAN} medium. When the OD value at 600 nm reaches 0.8 after cultivation at 30 °C, 250 rpm for ca. 6 h, 0.5 mM IPTG was added to induce enzyme expression at 30 °C for 18 h. Cell pellets were harvested by centrifuging the plates at 4 °C, 3800 rpm for 15 min. Cell pellets were lysed by adding 20 mL lysis buffer (50 mM sodium phosphate, pH 7.6, 0.5 mg/mL lysozyme, 10 mM imidazole). Cells were disrupted by sonication for 10 min (10 sec. on and 5 sec. off, 50% amplitude), and debris was removed by centrifugation at 14,000 × g for 1 hour. The protein purification was performed with Protino Ni-IDA 2000 packed column following the manufacturer΄s protocol but without NaCl in buffers. Afterward, the PD-10 desalting column (GE Healthcare, Germany) was applied to remove the imidazole. The purified enzymes were stored at sodium phosphate (50 mM, pH 7.6, 5% (v/v) glycerol) in small aliquots at -80 °C, and each aliquot was used only once after thawing. SDS-PAGE verified the purified proteins, and enzyme concentration was determined with Pierce™ BCA protein assay (**Fig. S3**).

Enzymatic assays and product isolation

Enzymatic activity was assayed at 30 ° C, and products (adenosine, 5'-chloro-5'-deoxyadenosine (5′ -ClDA), and 5'-fluoro-5'-deoxyadenosine (5'-FDA)) were identified by high-performance liquid chromatography (HPLC) by comparison on their retention time with the positive control, and by mass spectroscopy. 5 µM purified enzymes MJ1651, TTHA0338, and PH04363 (a positive control for S-adenosyl-L-methionine (SAM) hydrolase) were incubated with 1 mM SAM in 50 mM sodium phosphate buffer (pH 8.0), in a final volume of 200 µL, respectively. Similar reaction condition was performed in the presence of 100 mM NaF or NaCl for the three enzymes mentioned above, SsFlA, SalL (positive control for chlorinase), and ScFlA (positive control for fluorinase), respectively. 100 µL sample was taken out after 24 h at 30 °C and mixed with 50 µL ice cold 2% formic acid to terminate the reaction by precipitating the proteins. Precipitated material was removed by centrifugation (13000 rpm 5 min 4 °C) before 5 µL portions were subjected to analytical HPLC. All HPLC analyses were carried out on an Agilent 1260, equipped with a diode array detector, using an analytical Kinetex EVO C18 100 Å LC column (5 µm, 150 × 4.6 mm) at a flow rate of 1.0 mL/min with the following elution system: solvent A (H_2O supplemented with 0.1% trifluoroacetic acid) and B (acetonitrile supplemented with 0.1% trifluoroacetic acid) and linear-gradient (ratio A/B 98/2 during 5 min, then 98/2 to 90/10 in 15 min, then 90/10 to 0/100 in 5 min), λ= 260 nm.

Product identification with high-resolution mass spectrometry

The substrate SAM (**1**), products adenosine (**2**), 5'-FDA (**3**), and 5'-ClDA (**4**) were analyzed and identified on an Thermo Scientific Liquid Chromatography Mass Spectrometry (LC-MS) by using an Hypersil GOLD™ VANQUISH™ PFP UHPLC columns (1.9 µm, 2.1 mm × 100 mm), with a flow rate of 0.4 mL/min. A gradient of acetonitrile/H_2O system (1-10%) containing 0.1% trifluoroacetic acid (TFA) was programed over 10 mins. Thermo Scientific Q Exactive was equipped with an ESI source and Orbitrap mass analyzer. Calibration was performed with Pierce LTQ Velos ESI Positive Ion Calibration Solution (ThermoFisher). And Thermo Scientific SII for Xcalibur was used to control, acquire, and interrogate data from Thermo Scientific LC-MS systems and related instruments. The Full MS-SIM was operated with the following parameters: 70,000 resolution, scan range 50 to 750 m/z, AGC target T = 3e6, Maximum IT 200 ms, and polarity positive. Data analysis was then conducted using the Qual browser application within Xcalibur software (ThermoFisher Scientific) and performed using GraphPad Prism version 9.0.2 (www.graphpad.com).

Enzyme kinetic analysis of MJ1641, TTHA0338 and SsFlA

By following the previous enzyme kinetic studies on fluorinase and chlorinase (*30*, *41*, *50*), enzymatic activity was assayed at 37 °C by monitoring adenosine or 5'-ClDA production using analytical HPLC. The MJ1641 (2 μM) was incubated at various concentrations of SAM in sodium phosphate buffer (50 mM, pH 8.0), in a final volume of 100 μL for 30 min. Similar conditions were performed for TTHA0338 (2 μM) and SsFlA (5 μM) but with 10 min and 3h, respectively. Additional saturating concentration of NaCl (100 mM) was used for measuring the enzyme kinetics of SsFlA as chlorinase. For each enzyme, the incubation time was optimized for initial rate determination by taking measurement accuracy and product detectability into account. The reaction was terminated by adding 50 μL ice cold 2% formic acid after reaction time, followed by centrifugation. 10 μL supernatant was subjected to analytical HPLC to determine the concentration of adenosine or 5'-ClDA using a standard curve. Besides Agilent 1260 HPLC system, Shimadzu Prominence UFLC system (Kyoto, Japan) with a UV/VIS Photodiode Array Detector SPD-M20A was also used for product detection with 5 μL injection. Kinetic parameters were obtained by fitting the data to Michaelis-Menten equation using GraphPad Prism. Enzyme concentration was determined by Pierce™ BCA Protein Assay Kit.

Synthesis of 5'-ClDA

Synthesis of 5'-ClDA was carried out as described elsewhere (*30*). 268 mg (1.0 mmol) adenosine was added in 163 μL (2.0 mmol) pyridine at 0 °C in ice bath, followed by adding 3.5 mL CH_3CN. Then $SOCl_2$ (370 μL, 5.0 mmol) was slowly added into the obtained suspension. Stirring (300 rpm) was continued for reaction in ice bath with warming to room temperature overnight. The resulting precipitate was filtered and dried in vacuo. The resultant white solid powder was suspended in 6 mL of $MeOH/H_2O$ with a 5:1 v/v ratio. Concentrated aqueous ammonia (500 μL) was added with stirring for 30 min at room temperature. Afterwards, the reaction was evaporated under reduced pressure. The resulting colorless solid was crystallized from water (putting on ice can accelerate the crystallization process) and lyophilized overnight to give 5'-ClDA as white powder. ^1H NMR (600 MHz, DMSO-d_6): δ 8.34 (s, 1H), 8.15 (s, 1H), 7.31 (s, 2H), 5.93 (d, J = 5.6 Hz, 1H), 5.61 (brs, 1H), 5.47 (brs, 1H), 4.75 (m, 1H), 4.22 (m, 1H), 4.09 (m, 1H), 3.95 (dd, J=11.6, 5.1 Hz, 1H), 3.84 (dd, J=11.6, 6.4 Hz, 1H). ^1H NMR (600 MHz) spectrum for 5'-ClDA (4) in DMSO-d_6 is shown in **Fig. S10.**

Code availability

The source code of CLEAN is available at https://github.com/tttianhao/CLEAN

Supplementary Text

Supplementary Text 1. ML model development and evaluation

In the training stage, raw amino acid sequences were first embedded using the pre-trained language model ESM-1b, which is a state-of-the-art algorithm that can generate semantically rich representations of protein sequences, encoding their evolutionary, structural, and biophysical properties (*19*). To preserve high-quality data, we only focused on SwissProt, an expertly reviewed portion of the UniProt. An additional filter was applied to data curation by only selecting enzymes with all four digits of EC labeled, making the total training data ~220k. During the model development, to prevent testing data being overly similar to the training data, we used MMSeqs2 (*51*) to cluster the data using various sequence identity cutoffs ranging from 10% to 70%. ”. The clustered dataset follows an 80/20 split with five-fold cross validation, in which each test set included sequences that shared no more than 100%, 70%, 50%, 30% or 10% with any sequences in the training set. Notably, the clustering split was challenging because the sequence similarity between the training and testing dataset was decreased. Clustering split rules out the possibility that the good performance of CLEAN is a result of a similar testing dataset and training dataset. The split represented the case where query enzymes had low similarities with currently annotated enzymes. Hyperparameter tuning was also performed using clustering split. The evaluation of CLEAN on independent datasets (**Fig. 2**) were trained without the use of clustering split because the testing dataset was excluded from the training process for all models. Besides precision, recall, and F1-score for evaluating the performance of CLEAN, AUC (Area Under the receiver operating characteristic Curve) were also calculated and shown in **Fig. S2a**. CLEAN has also demonstrated a better accuracy than DeepEC on a large scale dataset obtained from TrEMBL database (**Fig. S15**).

Supplementary Text 2. The performance of CLEAN on understudied functions using triplet margin loss and SupConH loss

We hypothesized that compared with the multi-label classification framework, contrastive learning could better handle the imbalanced EC numbers where some EC numbers have thousands of enzyme examples, and some only have very few (less than 5). However, these EC numbers were vital as they represent the understudied functions. The imbalanced dataset posed a challenge for multi-label classification because the model could barely learn anything from the classes lacking positive examples. For example, in the case of ProteInfer, the authors reported that the performance of understudied EC numbers was halved compared to data-abundant EC numbers (*15*).

To support our hypothesis, we not only curated a validation dataset with enzymes from rare EC numbers to demonstrate the performance of understudied functions (**Fig. 2c-f**), but also compared two different contrastive loss functions, triplet loss and a variant of supervised contrastive loss, termed SupCon-Hard (SupConH) loss. During training, SupConH samples several negative sequences where triplet loss only samples one. SupConH was observed to have superior performance on understudied of the test queries (**Fig. S2b**). Besides SupConH sampled more negatives per batch during training, it also had an intrinsic ability to balance positives/negatives (*16*). While for both SupConH and Triplet Margin losses, the negatives are hard-mined based on their distances to the EC cluster centers, not all negatives weight the same in SupConH. SupConH weights less for the negatives far away from the anchor and weights more for those closer to the anchor. Similarly, SupConH weights more for positives near the anchor and less for those farther away. This property contributed to the SupConH's better performance on less seen data, because positive examples were structurally pulled closer to the anchor and negative examples were pushed away from the anchor. The latter allowed high-quality EC cluster centers to be constructed even if few enzymes with the same EC numbers are present in the training set. This also explained why SupConH performs significantly better for the 10%, 30% and 50% sequence identity splits, since these splits produces smaller training datasets, and SupCon-Hard better unitized both the limited negative examples and limited positive examples.

Supplementary Text 3. Evaluation of CLEAN's uncertainty estimation

To test if using GMM can effectively prevent overprediction by the detection of low confidence, we first evaluated whether CLEAN is able to flag low confidence scores when it is unsure about the predictions. To this end, we constructed a test dataset such that, by design, CLEAN should have a high level of uncertainty when making predictions. Specifically, we composed an "negative control dataset" by removing a number of EC numbers from the training dataset and used as testing dataset. Under such setup, CLEAN would be impossible to make correct predictions since the true labels of the testing dataset were excluded from training. To make the task even more difficult, the removed EC numbers were selected if there are only two possible 4th level EC numbers at the 3rd level. (i.e., if a.b.c.1 and a.b.c.2 are the only two possible EC numbers for a.b.c.-, one of the a.b.c.1 and a.b.c.2 is selected at random and removed from the training dataset). If CLEAN's confidence estimation is accurate, the confidence scores associated with CLEAN's predictions for these holdout proteins should be very low since the test proteins are excluded from the training dataset of CLEAN. Our results confirmed our hypothesis: CLEAN showed extremely low confidence scores for the predictions of those test proteins, with the vast majority of predictions below a confidence score of 0.2 (**Fig. S13** left). We further constructed another test dataset for which CLEAN has over 0.95 prediction accuracy as positive control (**Fig. S13** right) and observed that the confidence scores for the predictions of those proteins were enriched to a high confidence region (>0.9). These results suggest that the confidence estimation of CLEAN is very informative and correlated with prediction accuracy.

Next, we evaluated whether the informative uncertainty quantification of CLEAN translates to improvements of the overprediction issue. To avoid overprediction, CLEAN used the estimated confidence score to guide its predictions: it only predicts the specific EC numbers to the fourth digit level when the confidence score associated with the prediction is sufficiently high, otherwise it would just report its prediction of the third-level EC number for the input protein. Superficially, we re-ran CLEAN on the combined test dataset (Price-149 and New-392) such that it predicts the EC number at the fourth level only when its associated confidence score is >0.5, otherwise it outputs the EC number at the third level. We observed that guided by confidence estimates, CLEAN successfully predicted more true positives as compared to ProteInfer and DeepEC (**Fig. S14**). These results suggest that our uncertainty quantification algorithm enabled CLEAN to achieve an adaptive prediction scheme that largely avoided overpredictions. Furthermore, on the combined test dataset (Price-149 and New-392), we further analyzed the correlation of prediction accuracy and confidence. As shown in **Fig. S12**, CLEAN's prediction accuracy goes up when the confidence is high. The result suggests that CLEAN is likely to be reliable when the confidence is high, which further demonstrates that the uncertainty quantification is indeed informative and have positive correlation with prediction accuracy.

Supplementary Text 4. CLEAN outperforms ESM-1b in classification

To visualize the learned embedding on the low dimension, we used t-SNE (*52*) to reduce the high dimension embedding to a two-dimension plot **(Fig. S9).** t-SNE can preserve the distance information when projecting high-dimensional data to low dimensions. Compared to the embedding by ESM-1b prior to contrastive learning, CLEAN's plot showed much more defined clusters than ESM-1b, and the visualization result also supported the outstanding accuracy performance of CLEAN.

Supplementary Text 4. *In vitro* experimental validation of CLEAN predicted EC numbers of MJ1651, TTHA0338, and SsFlA

Overall, CLEAN had 100% prediction accuracy on haloperoxidases and SAM-dependent halogenases covering from the first to the fourth digit of EC number. Similar results were observed for flavin-dependent (92.3%, 12/13) and α-KG-dependent halogenases (92.3%, 12/13) on the third digit. In detail, MJ1641 from *Methanocaldococcus jannashii* DSM 2661 was labeled as a chlorinase with EC number 2.5.1.- in Uniprot/Swiss-Prot database after expert curation. However, MJ1651 failed to show observable chlorinase activity (*28*). In addition, although DEEPre predicted MJ1651 as a fluorinase (EC 2.5.1.94), no peak appeared at the retention time of the expected fluorinated product 5'-FDA (3) (**Fig. S5a**), suggesting

MJ1651 lacks the fluorinase activity. In contrast, CLEAN predicted MJ1641 as a S-adenosyl-L-methionine (SAM) hydrolase with EC number 3.13.1.8 (**Fig. 3c**). The latter was firstly confirmed by comparing the retention time of product adenosine (**2**) on HPLC with positive control enzyme PH04363 (**Fig. 3f-g** and **Fig. S3**, and **S4a-b**). Moreover, the product adenosine (**2**) obtained from the reaction mixture with purified MJ1641 was identified by high-resolution mass spectrometry (MS) (**Fig. 3l-m**, and **Fig. S4e-f**), demonstrating that MJ1641 is indeed a SAM hydrolase (EC 3.13.1.8) as CLEAN predicted. Unfortunately, both BLASTp and ProteInfer failed to predict the EC number of MJ1641. TTHA0338 from *Thermus thermophilus* is a member of the DUF62 Pfam family with no known function. Also, to the best of our knowledge, no catalytic activity has been demonstrated to date. CLEAN successfully labeled the uncharacterized protein TTHA0338 as a hydrolase with EC 3.13.1.8 (**Fig. 3c**) instead of chlorinase or fluorinase, which was subsequently confirmed by HPLC and MS analysis (**Fig. 3h** and **3n, Fig. S5b**).

Benefiting from the contrastive learning, CLEAN confidently assigned three EC numbers to SsFlA (EC 2.5.1.63; EC 2.5.1.94; EC 3.13.1.8), which is different from other halogenases with single precise EC number (**Table S2**). In other words, SsFlA, a SAM-dependent fluorinase from *Streptomyces* sp. labeled in Uniprot, might have the activity of chlorinase (EC 2.5.1.94) and hydrolase (EC 3.13.1.8) regarding the forecasts of CLEAN (**Fig. 3e**). Surprisingly, these two additional enzymatic activities were verified by *in vitro* experiments with relevant substrates (NaF, NaCl, and H_2O). Both new products adenosine (**2**), 5'-ClDA (**4**) and original product 5'-FDA (**3**) were further confirmed by HPLC and MS with positive controls (e.g., chlorinase SalL and fluorinase ScFlA) (**Fig. 3i-k** and **3o-q**, **Fig. S4c-d** and **S4g-h**). The latter indicated SsFlA has promiscuous activities that can catalyze two fortuitous side reactions (chlorination and hydrolysis) in addition to its main reaction (fluorination). These observations confirmed that CLEAN can effectively recall the defined biological activity for promiscuous enzymes, in agreement with the *in silico* recall validation. In addition, we noticed that the substrate SAM (1) cannot be consumed completely after 24h for all three enzymes (MJ1651, TTHA0338, and SscFlA, **Fig. 3g-k**), which is consistent with the positive control enzymes PH04363, ScFlA and SalL (**Fig. S4**). We performed the enzyme kinetic analysis for MJ1651, TTHA0338, and SsFlA (**Table S3, Fig. S7**). The turnover numbers (k_{cat}) of MJ1651, TTHA0338, and SsFlA for converting SAM to adenosine were 0.327 min^{-1}, 2.775 min^{-1}, and 0.017 min^{-1}, respectively, suggesting that all three enzymes have considerable hydrolytic activity on SAM. These results are comparable to the previously characterized SAM hydrolase like SaDUF62 (k_{cat} = 0.5 min^{-1}) (*29*). Similar K_M values of MJ1651 (K_M = 82.6 μM), TTHA0338 (K_M = 138.0 μM) and SsFlA (K_M = 15.8 μM) were obtained compared to the positive control PH04363 (K_M = 39.2 μM) (*53*). The kinetic parameters of SsFlA with fluorinase activity were obtained from our previous study (k_{cat} = 0.22 min^{-1}, K_M = 34.6 μM) (*30*). SsFlA also had comparable SAM affinity (K_M = 10.7 μM) in the presence of chloride anion (conversion of SAM to 5'-ClDA). Its lower k_{cat} (0.008 min^{-1}) is likely because the equilibrium of the chlorination reaction lying significantly in favor of substrate over product rather than an inherent inability of the fluorinase to activate the chloride ions towards nucleophilic attack (*54*). This limitation can be solved by the two separate coupled-enzyme assays that are designed to shift the equilibrium of the reaction towards the organochlorine product as previously reported (*54*). Overall, these results suggest that the detected activity of MJ1651, TTHA0338, and SsFlA should be their biologically relevant function. Besides, as shown in **Fig. S8**, HPLC-MS studies revealed that the peak between **1** and **2** is adenine, which comes from the degradation of SAM (*55*). SAM can be hydrolyzed to adenine (**5**) and *S*-pentosylmethionine with first-order rate constants, 3 x 10^{-6} s^{-1} (*55*).

Both MJ1641 and TTHA0338 were members of the DUF62 family in the Pfam database. The reason why the six commonly used enzyme function annotation tools (i.e., BLASTp, DeepEC, ProteInfer, DEEPre, ECPred and COFACTOR) cannot accurately label the two hydrolases (MJ1641 and TTHA0338) might be because the sequence identities of both enzymes are closer to chlorinases in available databases. However, our CLEAN made the correct prediction mainly because: i) contrastive learning learns from not only positive examples but also negative examples, and ii) ESM-1b representations can reliably capture semantically rich information other than sequence similarities.

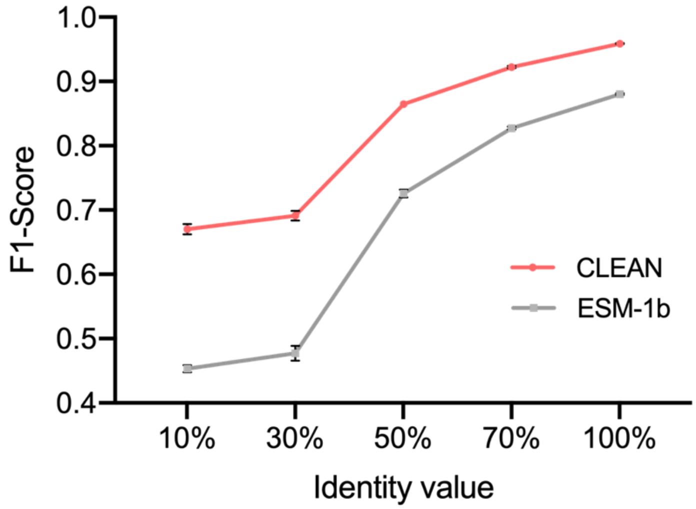


Fig. S1. The evaluation results of CLEAN on different identity clustering split under 5-fold CV (cross validated). ESM-1b was also investigated as comparison.

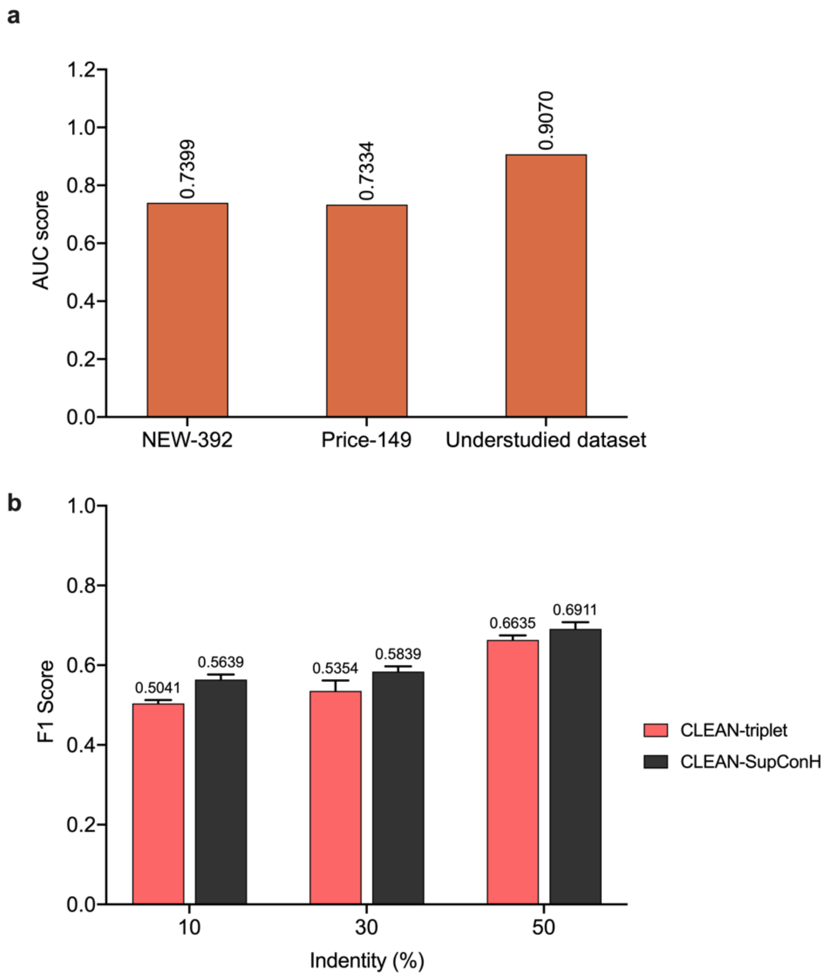


Fig. S2. (a) Evaluation of CLEAN's performance towards AUC examined on three databases. AUC stands for "Area Under the receiver operating characteristic curve." Analysis of SupConH loss and triplet towards CLEAN. The validation set consists of EC numbers with no more than 5 occurrences in the training dataset. SupConH loss can further improve CLEAN's performance on understudied enzymes.

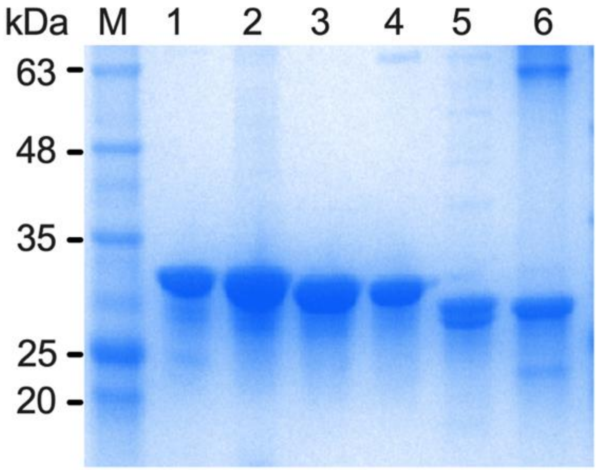


Fig. S3. SDS-PAGE analysis of the purified uncharacterized proteins and positive controls. Purified proteins were loaded onto 12% Mini-PROTEAN TGX^{TM} Precast Gel (BIO-RAD). Lane M is a protein molecular weight marker. Lanes 1-6 are SalL (MW: 32.4 KDa), ScFlA (MW: 33.4 KDa), SsFlA (MW: 33.2 KDa), MJ1651 (MW: 31.3 KDa), TTHA0338 (MW: 28.0 KDa), and PH04363 (MW: 29.7 KDa), respectively. The gel was stained with Coomassie Brilliant Blue. MW: molecular weight.

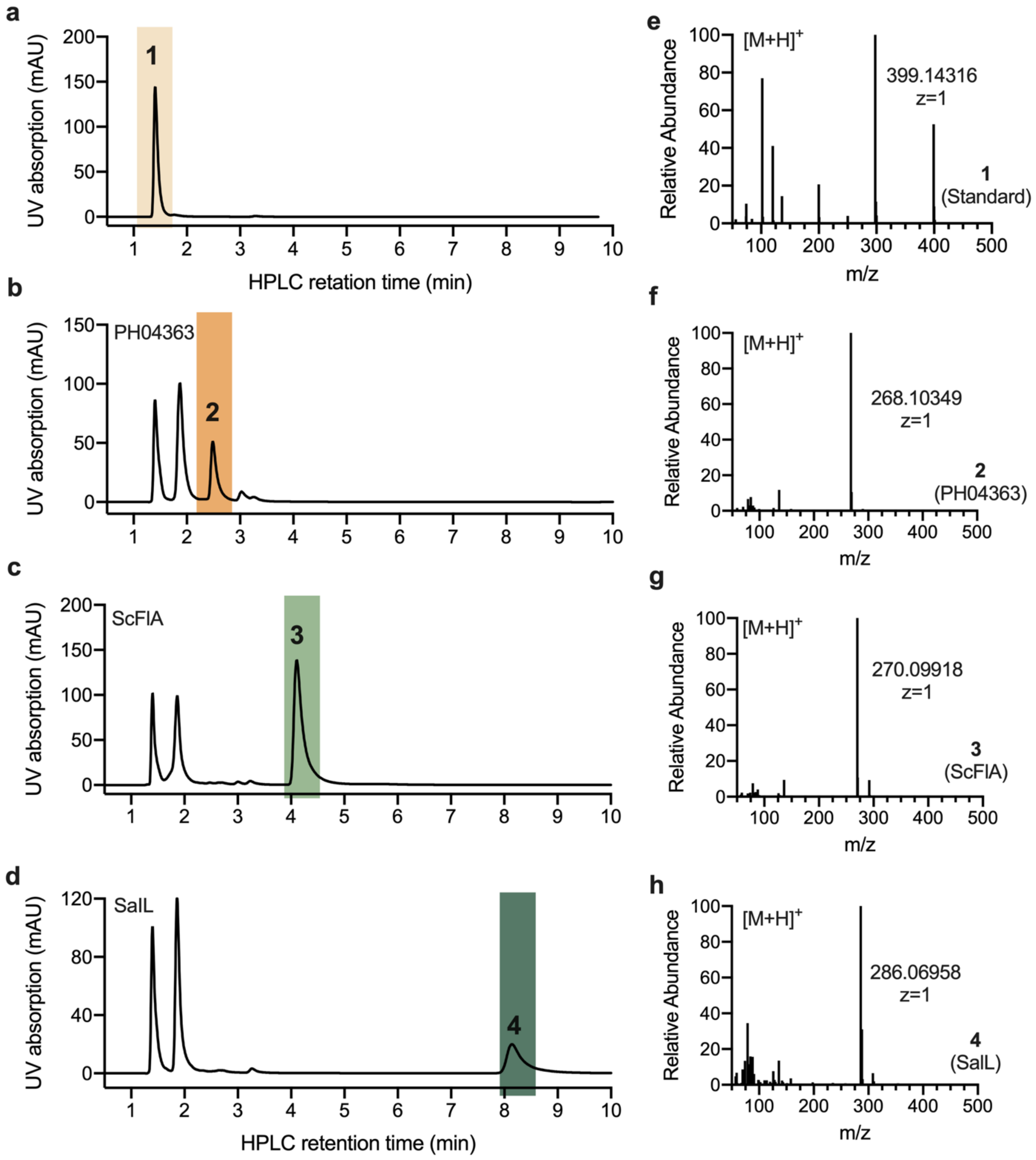


Fig. S4. Experimental validation of the CLEAN predicted EC numbers of positive control halogenases. Left panel: HPLC analysis of (**a**) pure SAM (**1**) and reaction mixture of SAM with purified (**b**) PH04363+H_2O, (**c**) ScFlA+NaF, (**d**) SalL+NaCl at 30 °C for 24h. Detection was performed with UV absorbance at 260 nm. The peaks of substrate SAM (**1**), product adenosine (**2**), 5'-FDA (**3**), and 5'-ClDA (**4**) were labeled with light yellow, orange, green, and dark green, respectively, which are also aligned vertically based on retention time; Right panel: Mass spectra of compounds obtained from various reaction mixtures: (**e**) **1** standard, (**f**) **2** in PH04363 reaction, (**g**) **3** in ScFlA reaction, and (**h**) **4** in SalL reactions.

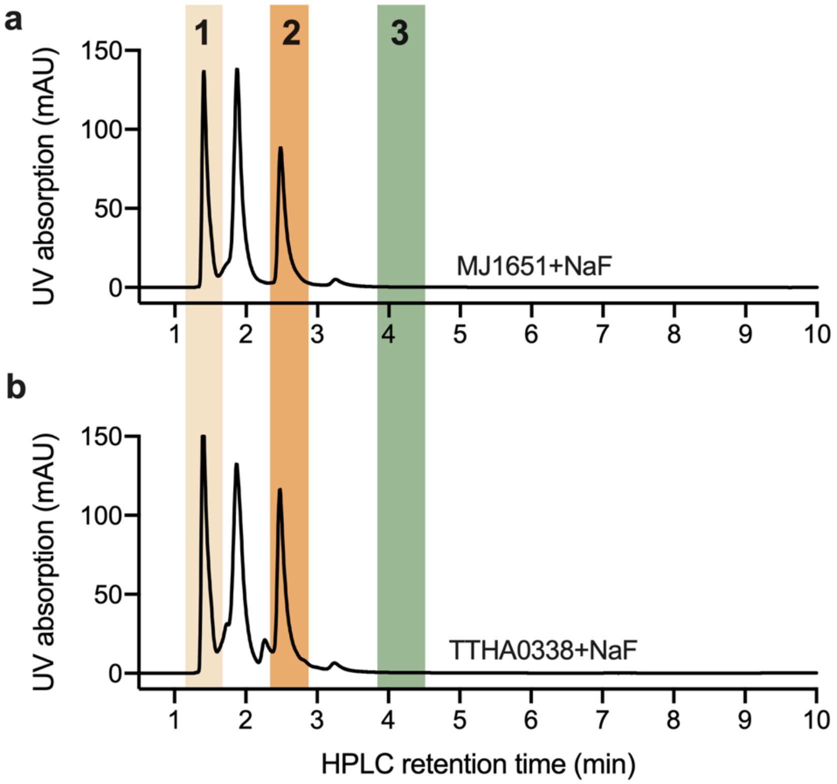
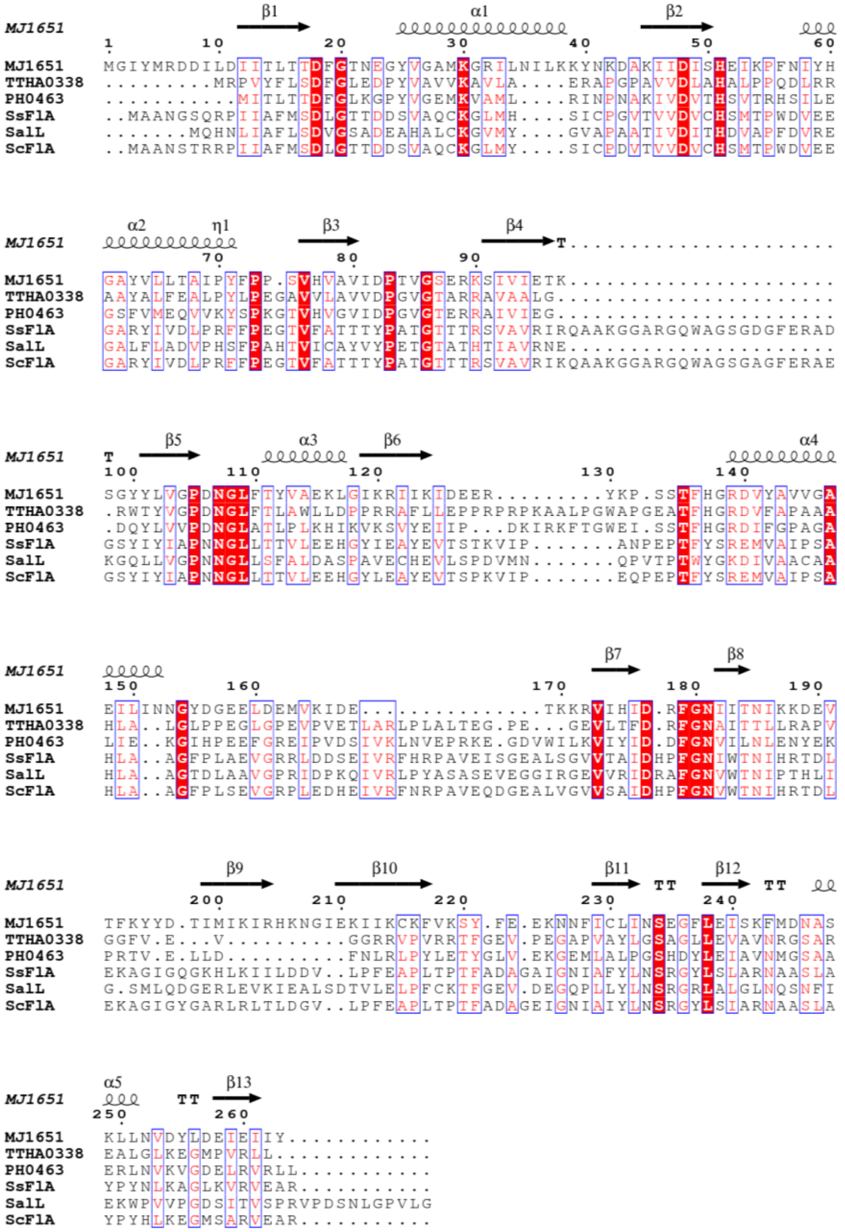


Fig. S5. HPLC analysis of reaction mixtures of SAM with purified enzymes: (**a**) MJ1651+NaF and (**b**) TTHA0338+NaF at 30 °C for 24 h. Detection was performed with UV absorbance at 260 nm. The peaks of substrate SAM (**1**) and product adenosine (**2**) were labeled with light yellow and orange, respectively, which were also aligned vertically based on the retention time. In addition, the peak of the expected fluorinated product 5'-FDA (**3**) was aligned vertically based on its retention time and labeled with green.



```
MJ1651     1        10        20        30        40        50        60
MJ1651     MGIYMRDDILDIITLTTDFGTNEGYVGAMKGRILNILKKYNKDAKIIDISHEIKPFNIYH
TTHA0338   .........MRPVYFLSDFGLEDPYVAVVKAVLA....ERAPGPAVVDLAHALPPQDLRR
PH0463     ...........MITLTTDFGLKGPYVGEMKVAML....RINPNAKIVDVTHSVTRHSILE
SsFlA      ..MAANGSQRPIIAFMSDLGTTDDSVAQCKGLMH....SICPGVTVVDVCHSMTPWDVEE
SalL       .......MQHNLIAFLSDVGSADEAHALCKGVMY....GVAPAATIVDITHDVAPFDVRE
ScFlA      ..MAANSTRRPIIAFMSDLGTTDDSVAQCKGLMY....SICPDVTVVDVCHSMTPWDVEE

MJ1651                                     T.......................
                    70        80        90
MJ1651     GAYVLLTAIPYFPP.SVHVAVIDPTVGSERKSIVIETK......................
TTHA0338   AAYALFEALPYLPEGAVVLAVVDPGVGTARRAVAALG.......................
PH0463     GSFVMEQVVKYSPKGTVHVGVIDPGVGTERRAIVIEG.......................
SsFlA      GARYIVDLPRFFPEGTVFATTTYPATGTTTRSVAVRIRQAAKGGARGQWAGSGDGFERAD
SalL       GALFLADVPHSFPAHTVICAYVYPETGTATHTIAVRNE......................
ScFlA      GARYIVDLPRFFPEGTVFATTTYPATGTTTRSVAVRIKQAAKGGARGQWAGSGAGFERAE

MJ1651     T
           100       110       120       130       140
MJ1651     SGYYLVGPDNGLFTYVAEKLGIKRIIKIDEER.........YKP.SSTFHGRDVYAVVGA
TTHA0338   .RWTYVGPDNGLFTLAWLLDPPRRAFLLEPPRPRPKAALPGWAPGEATFHGRDVFAPAAA
PH0463     .DQYLVVPDNGLATLPLKHIKVKSVYEIIP...DKIRKFTGWEI.SSTFHGRDIFGPAGA
SsFlA      GSYIYIAPNNGLLTTVLEEHGYIEAYEVTSTKVIP.......ANPEPTFYSREMVAIPSA
SalL       KGQLLVGPNNGLLSFALDASPAVECHEVLSPDVMN.......QPVTPTWYGKDIVAACAA
ScFlA      GSYIYIAPNNGLLTTVLEEHGYLEAYEVTSPKVIP.......EQPEPTFYSREMVAIPSA

MJ1651
           150       160                170       180       190
MJ1651     EILINNGYDGEELDEMVKIDE...............TKKRVIHID.RFGNIITNIKKDEV
TTHA0338   HLA..LGLPPEGLGPEVPVETLARLPLALTEG.PE...GEVLTFD.RFGNAITTLLRAPV
PH0463     LIE..KGIHPEEFGREIPVDSIVKLNVEPRKE.GDVWILKVIYID.DFGNVILNLENYEK
SsFlA      HLA..AGFPLAEVGRRLDDSEIVRFHRPAVEISGEALSGVVTAIDHPFGNIWTNIHRTDL
SalL       HLA..AGTDLAAVGPRIDPKQIVRLPYASASEVEGGIRGEVVRIDRAFGNVWTNIPTHLI
ScFlA      HLA..AGFPLSEVGRPLEDHEIVRFNRPAVEQDGEALVGVVSAIDHPFGNVWTNIHRTDL

MJ1651
                    200       210       220          230       240
MJ1651     TFKYYD.TIMIKIRHKNGIEKIIKCKFVKSY.FE.EKNNFICLINSEGFLEISKFMDNAS
TTHA0338   GGFV.E...V..........GGRRVPVRRTFGEV.PEGAPVAYLGSAGLLEVAVNRGSAR
PH0463     PRTV.E.LLD..........FNLRLPYLETYGLV.EKGEMLALPGSHDYLEIAVNMGSAA
SsFlA      EKAGIGQGKHLKIILDDV..LPFEAPLTPTFADAGAIGNIAFYLNSRGYLSLARNAASLA
SalL       G.SMLQDGERLEVKIEALSDTVLELPFCKTFGEV.DEGQPLLYLNSRGRLALGLNQSNFI
ScFlA      EKAGIGYGARLRLTLDGV..LPFEAPLTPTFADAGEIGNIAIYLNSRGYLSIARNAASLA

MJ1651
          250       260
MJ1651     KLLNVDYLDEIEIIY............
TTHA0338   EALGLKEGMPVRLL.............
PH0463     ERLNVKVGDELRVRLL...........
SsFlA      YPYNLKAGLKVRVEAR...........
SalL       EKWPVVPGDSITVSPRVPDSNLGPVLG
ScFlA      YPYHLKEGMSARVEAR...........
```

Fig. S6. Sequence alignment of TTHA0338, MJ1651, and SsFlA with positive control enzymes. The amino acid sequences of six SAM-dependent enzymes are aligned by Clustal Omega. Known secondary structure elements from MJ1651 (PDB ID: 2F4N(*28*)) are displayed for all aligned sequences. The identical and similar residues are labeled with red color in blue box. ENDscript (*56*) was used for figure generation.

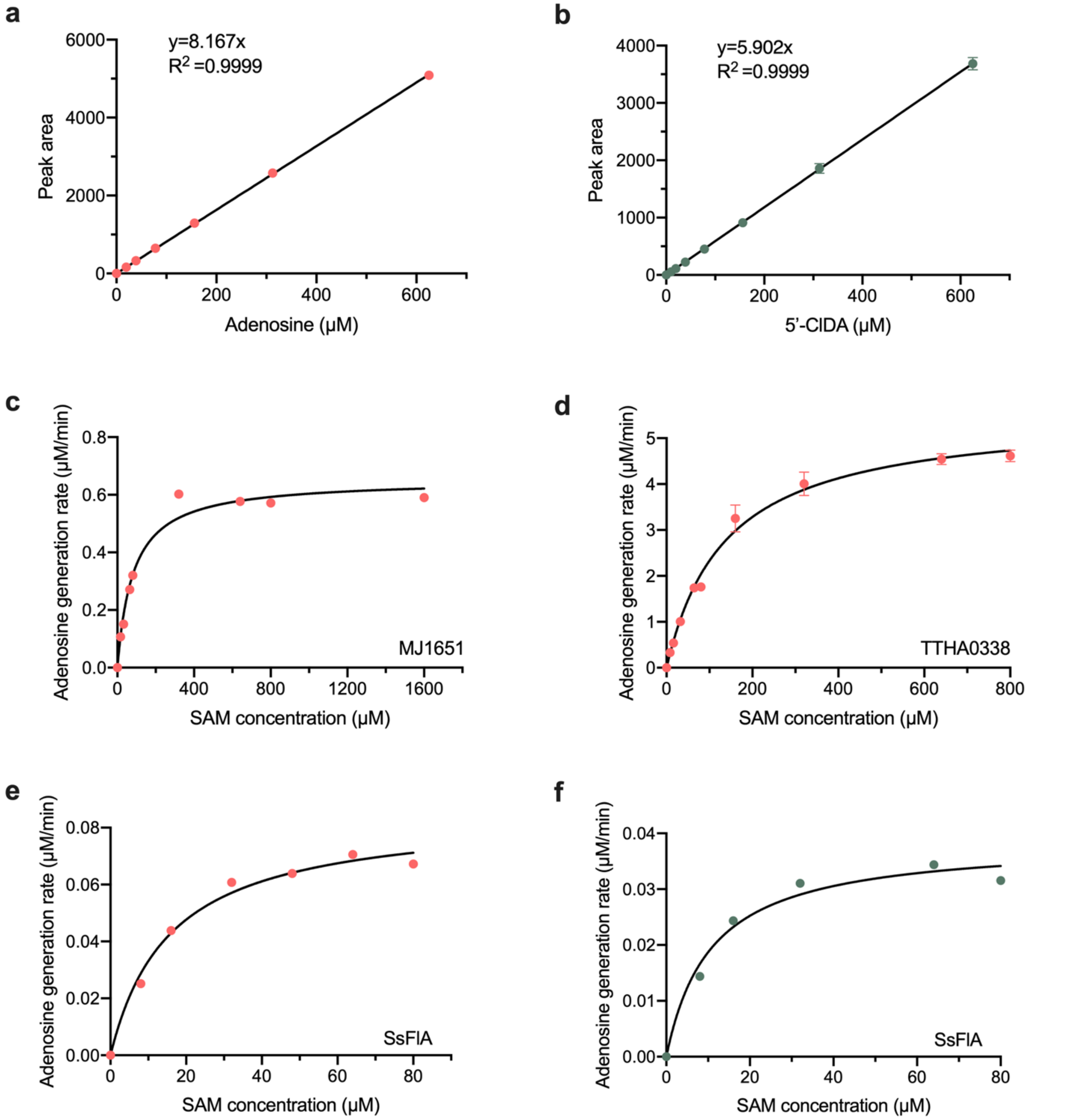


Fig. S7. Enzyme kinetic analysis of TTHA0338, MJ1651, and SsFlA. (**a, b**) Calibration curves of adenosine and 5'-ClDA by HPLC (Agilent 1260, 10uL injection), respectively. (**c-e**) Kinetic assays of MJ1651, TTHA0338 and SsFlA for conversion of SAM to adenosine, respectively. Assays contain varying concentrations of SAM in 50 mM sodium phosphate buffer (pH 8.0). (**f**) Kinetic assay of SsFlA for conversion of SAM to 5'-ClDA. Assays contain 100 mM NaCl and varying concentrations of SAM in 50 mM sodium phosphate buffer (pH 8.0). All the experiments were performed at least in duplicate. Since the variance of the experimental results is very small, the error bars might not be shown obviously.

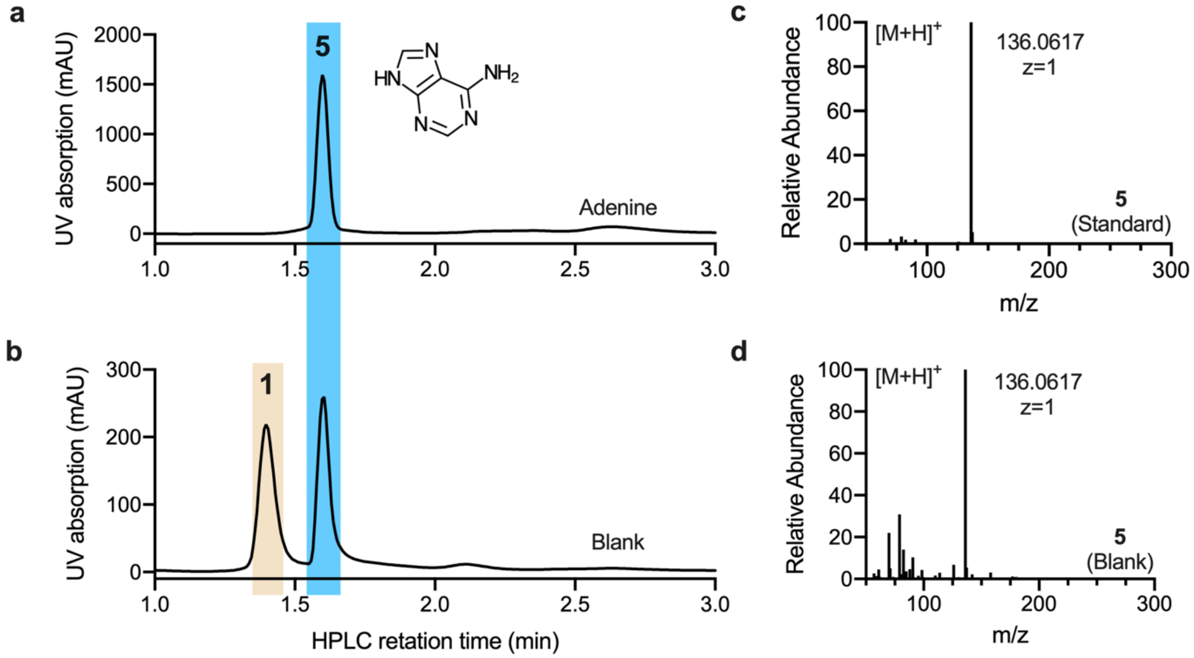


Fig. S8. Experimental validation of the adenine (**5**) degraded from SAM (**1**) during reaction. Left panel: HPLC analysis of (**a**) standard adenine (**5**) and (**b**) reaction mixtures of SAM and NaCl with blank (no enzyme). Detection was performed with UV absorbance at 260 nm. The peaks of substrate SAM (**1**) and adenine (**5**) were labeled with light yellow and blue, respectively, which are also aligned vertically based on retention time. Figure inset shows the chemical structure of adenine (**5**); Right panel: Mass spectra of compounds obtained from (**c**) **5** standard and (**d**) **5** in blank. A modified HPLC method with short retention time was used with a linear-gradient flow rate at 0-3 min (1.0 mL/min to 0.3 mL/min). Elution system: solvent A (H_2O supplemented with 0.1% trifluoroacetic acid) and B (acetonitrile supplemented with 0.1% trifluoroacetic acid) and linear-gradient (ratio A/B 98/2 during 5 min, then 98/2 to 0/100 in 1 min), flow rate after 3 min: 0.3 mL/min.

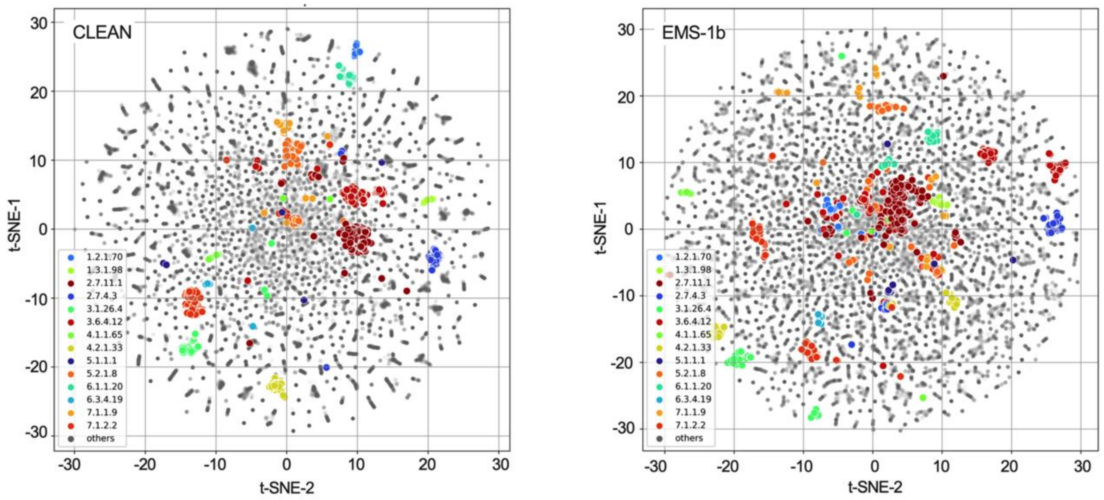


Fig. S9. The two-dimensional visualization of CLEAN's embedding compared with ESM-1b embedding using t-SNE. Each dot in the plot represented a single enzyme and each color represented an EC number. Several randomly selected EC numbers (~14 types) are highlighted for visualization purposes.

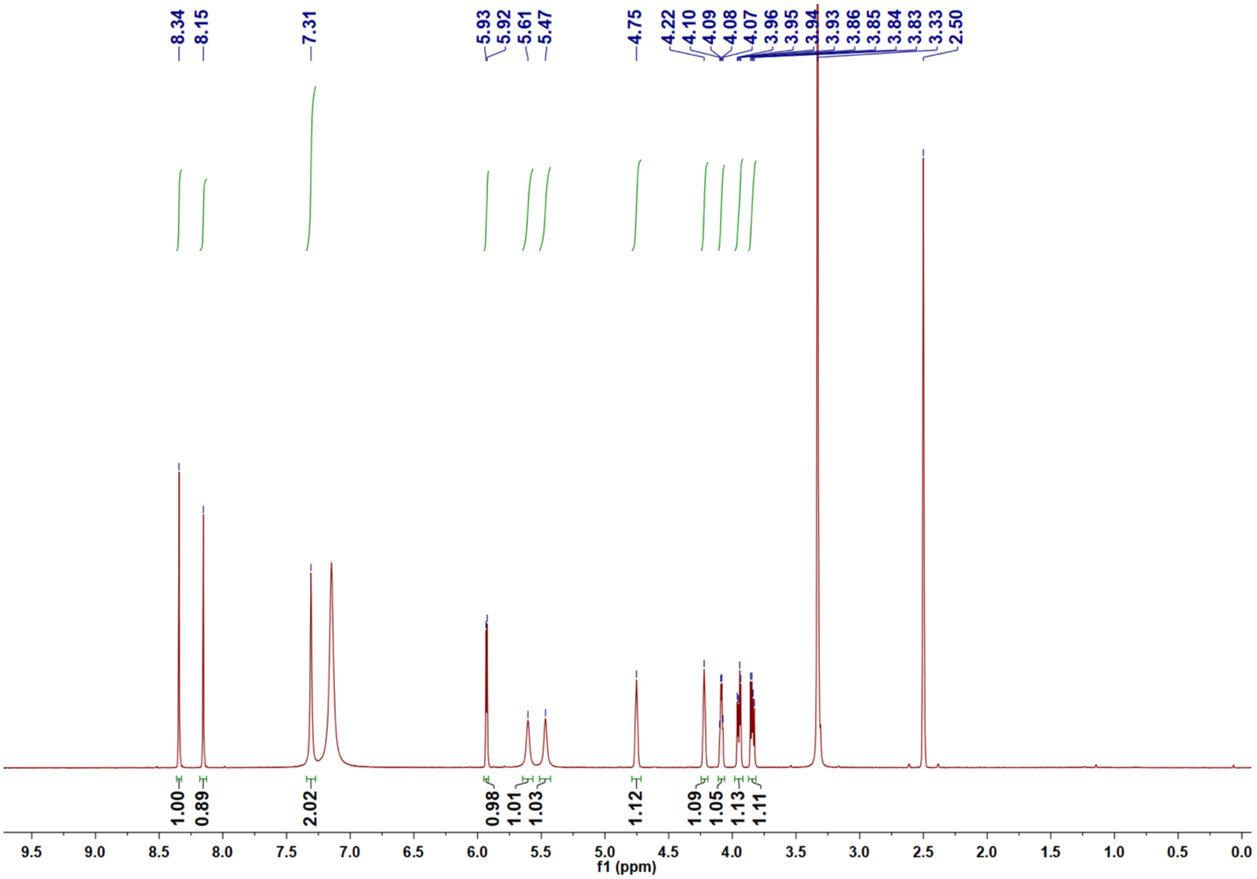


Fig. S10. ^{1}H NMR (600 MHz) spectrum for 5'-ClDA (**4**) in DMSO-d_6.

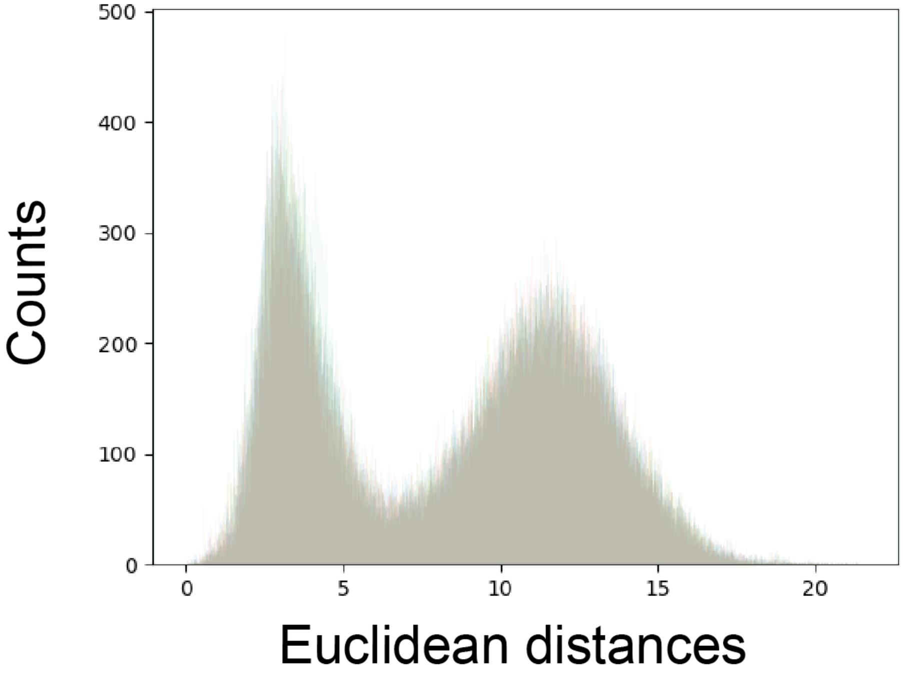


Fig. S11. The distribution of Euclidean distances between the embeddings of EC numbers and individual enzyme sequences shown by histogram, where x-axis are the distances and y-axis are the counts. The left peak is formed by matching EC numbers and enzymes, and the right peak is formed by mismatched EC numbers and enzymes.

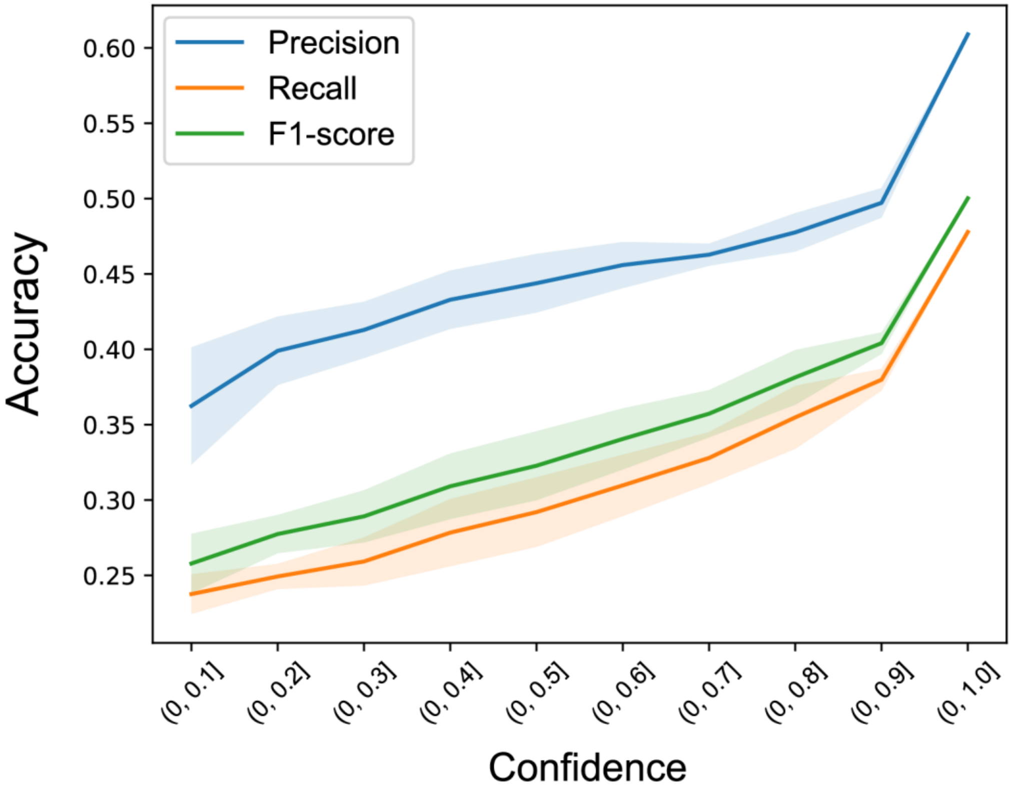


Fig. S12. The confidence vs cumulative prediction accuracy. Error bar is created by repeating using 40 independently fitted GMMs.

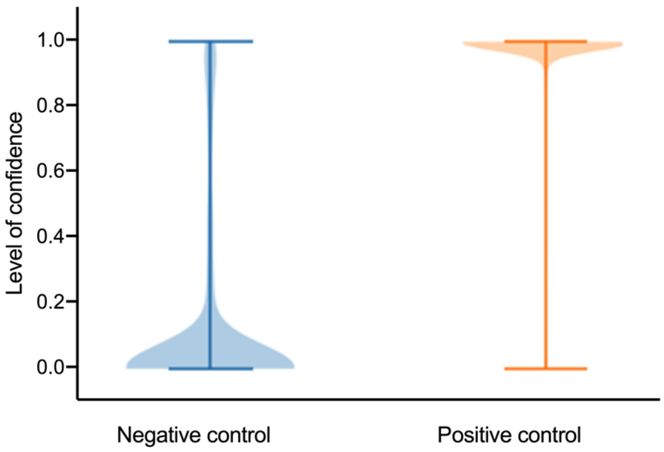


Fig. S13. Comparison of confidence distribution of extremely low accuracy dataset (left) and high accuracy dataset (right)

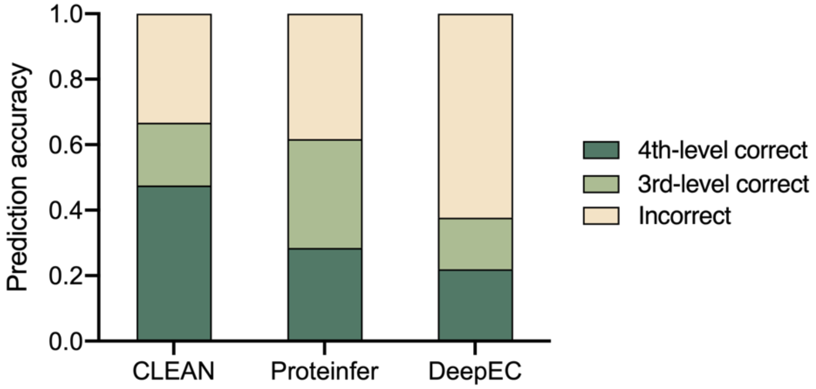


Fig. S14. The fraction of 4th level EC number accuracy and 3rd level EC number accuracy of the combined dataset of Price-149 and New-392.

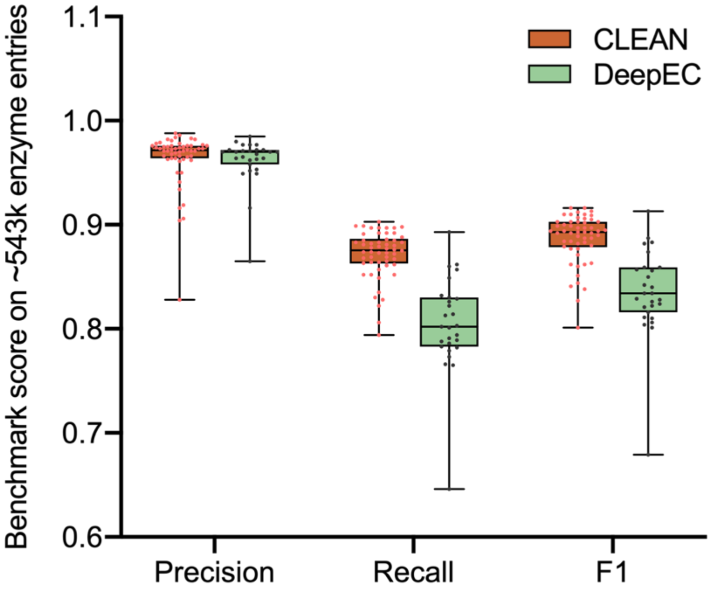


Fig. S15. The comparison of CLEAN and DeepEC on large (~543k enzyme entries obtained from TrEMBL) dataset. The entries were selected only if their annotation score is at least 4 out of 5 to ensure labels' quality. Due to the consideration of computing time, the whole dataset was broken down into smaller ones each containing 10k entries. Each point in the box plot was obtained from each split dataset of 10k entries.

Table S1. Max-Separation selection method algorithm

Step	Description
1	Function MAXSEP(S)[a]
2	Let background noise distance $\gamma = \text{mean}(s_1 + s_2 + \cdots + s_{n-1})$
3	Let noise separation distances $D = d_0, \ldots, d_{n-1} = \lvert s_0 - \gamma \rvert, \ldots, \lvert s_{n-1} - \gamma \rvert$
4	Let slope of separation curve $G = g_0, \ldots, g_{n-1} = \lvert d_1 - d_0 \rvert, \ldots, \lvert d_{n-1} - d_{n-2} \rvert$
5	Initialize maximum separation index $i \leftarrow 0$
6	Let mean slope $\overline{g} = \text{mean}(G)$
7	Let maximum separation index $i \leftarrow i'$ be the first i that satisfies $g_i > \overline{g}$
8	Return the correct set of EC numbers for query $\{EC_i\} = \{EC_0, \ldots, EC_i\}$

[a] S is defined as the sequence of distances between the query sequence and each EC number cluster s_0, s_1, ..., s_{n-1} in sorted order.

Table S2. Comparison of the prediction performance of various EC number prediction tools using the uncharacterized halogenases in Uniprot.

Type of halogenase	Name	Uniprot ID	EC Number Annotation							
			CLEAN	DeepEC	BLASTp[a]	ProteInfer[b]	DEEPre[c]	ECPred[c]	COFACTOR[c]	Manual curation[d,h]
Haloperoxidase	NapH1	A7KH27	1.11.1.10	n.p.[e]	6.1.1.19	1.-.-.-	2.7.1.0	3.-.-.-	1.11.1.10	1.11.1.10 (*57*)
	MarH1	A0A0F7N9T7	1.11.1.10	n.p.	unsure[f]	1.-.-.-	2.7.7.0	3.-.-.-	1.11.1.18	1.11.1.10 (*58*)
	MarH2	A0A559V0A1	1.11.1.10	n.p.	n.p.	3.1.-.-	2.7.7.0	1.13.-.-	1.11.1.18	1.11.1.10 (*58*)
	MarH3	A0A559V0T8	1.11.1.10	n.p.	n.p.	2.4.1.-	n.p.	4.-.-.-	1.11.1.10	1.11.1.10 (*59*)
Flavin-dependent	KtzR	A8CF74	1.14.19.9	n.p.	1.14.19.9	n.p.	1.14.19.9	6.-.-.-	1.5.5.1	1.14.19.59 (*59*)
	StaI	Q8KLM0	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*60*)
	Tjp10	A0A6H0DY41	1.14.19.9	n.p.	1.14.19.9	1.1.-.-	1.3.7.11	1.14.-.-	1.5.5.1	1.14.19.58 (*61*)
	VirX1	M4SKV1	1.14.19.9	n.p.	1.14.19.9	1.-.-.-	1.3.7.9	1.14.-.-	1.14.13.2	1.14.19.- (*62*)
	PlBmp2	A0A162BNF2	1.14.19.56	n.p.	unsure	1.-.-.-	1.3.1.101	1.3.-.-	1.3.1.-	1.14.19.- (*63*)
	HrmQ	C1IHU5	1.14.19.56	n.p.	1.14.19.56	1.-.-.-	1.14.13.11	1.3.-.-	1.14.13.2	1.14.19.56 (*64*)
	CazI	Q2GWL3	1.14.13.209	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.23	1.14.19.- (*65*)
	NapH2	A7KH29	1.14.19.56	n.p.	unsure	1.14.14.-	1.14.13.11	1.5.3.-	1.14.13.23	1.14.19.- (*66*)
	CalO3	Q8KND5	1.14.19.56	n.p.	unsure	1.3.-.-	1.3.7.101	1.14.14.-	1.14.13.7	1.14.19.- (*67*)
	McnD	A4Z4I8	1.14.19.56	n.p.	unsure	1.-.-.-	1.3.7.-	1.-.-.-	1.14.13.7	1.14.19.- (*68*)
	Chal	P96557	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*69*)
	Teg16	B7T1E1	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.3.-.-	1.14.13.23	1.14.19.- (*70*)
	Veg13	B7T185	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*70*)
α-Ketoglutarate-dependent	BarB1	Q8GAQ9	1.14.20.15	n.p.	1.14.11.18	1.14.11.-	1.14.11.18	1.14.11.-	1.14.11.-	1.14.20.- (*71*)
	BarB2	Q8GAQ8	1.14.20.15	n.p.	1.14.11.18	1.14.-.-	1.14.11.18	1.14.-.-	1.14.11.27	1.14.20.- (*71*)
	CytC3	D0VX22	1.14.20.15	n.p.	n.p.	1.14.11.-	1.14.11.18	1.14.11.-	1.14.11.16	1.14.20.- (*72*)
	HctB	Q1EDB4	1.14.11.18	n.p.	n.p.	n.p.	1.14.11.41	n.p.	1.14.11.18	1.14.11.- (*73*)
	AmbO5	A0A1L1YPD7	1.14.11.74	n.p.	5.5.1.14	1.14.11.-	1.14.11.-	4.-.-.-	1.14.11.16	1.14.11.- (*74*)
	AdeV	A0A1U8X168	1.14.11.26	n.p.	2.1.1.33	1.14.-.-	1.14.11.13	1.14.11.-	1.21.3.1	1.14.11.- (*75*)
	ArzI	UPI000360A9BD	1.3.7.6	n.p.	1.14.11.73	n.p.	1.14.11.-	n.p.	1.14.11.27	1.14.11.- (*76*)
	SlBesD	UPI0004CAEF9E	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	4.-.-.-	1.14.11.19	1.14.11.- (*77*)
	SwHalB	UPI000499837C	1.14.11.74	n.p.	unsure	n.p.	2.7.1.-	3.-.-.-	1.14.20.1	1.14.11.- (*77*)
	SiHalB	UPI000888A5A4	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	6.-.-.-	1.14.11.16	1.14.11.- (*77*)
	LaHalC	UPI0003675DF0	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	n.p.	1.21.3.1	1.14.11.- (*77*)
	PkHalD	UPI0005FAE1F9	1.14.11.46	n.p.	unsure	n.p.	2.3.2.-	n.p.	1.14.11.18	1.14.11.- (*77*)
	PsRoot562 HalE	UPI000702B50B	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	3.-.-.-	1.14.20.1	1.14.11.- (*77*)
SAM-dependent	NobA	W8JNL4	2.5.1.63	2.5.1.63	2.5.1.63	2.5.1.63	2.5.1.63	1.-.-.-	3.1.1.1	2.5.1.63 (*50*)
	flA4	A0A068VNW5	2.5.1.63	2.5.1.63	2.5.1.63	2.5.1.63	2.5.1.63	1.-.-.-	2.5.1.63	2.5.1.63 (*78*)
	ClA2	A0A2T0T269	2.5.1.94	2.5.1.94	2.5.1.94	4.2.1.51	2.5.1.94	2.-.-.-	3.4.14.10	2.5.1.94 (*79*)
	TTHA0338	Q5SLF5	3.13.1.8	n.p.	3.13.1.8	n.p.	2.5.1.94	2.-.-.-	2.5.1.63	3.13.1.8[g]
	MJ1651	Q59045	3.13.1.8	n.p.	n.p.	2.5.1.-	2.5.1.94	2.5.1.-	2.5.1.63	3.13.1.8[g]

	SsFlA	W0W999	2.5.1.63;
			2.5.1.94;
			3.13.1.8
				2.5.1.63	2.5.1.63	2.5.1.63	2.5.1.63	3.-.-.-	1.17.99.2	2.5.1.63[g];
										2.5.1.94[g];
										3.13.1.8[g]

[a]The BLASTp algorithm with standard database UniProKB/Swiss-Prot was applied for sequence alignment. The EC number was extracted from the top one result.

[b]Prediction was made by the code version of Proteinfer.

[c]Prediction was made by online server (DEEPre: http://www.cbrc.kaust.edu.sa/DEEPre/index.html; ECPred: https://ecpred.kansil.org; COFACTOR: https://zhanggroup.org/COFACTOR/, and structure was obtained from I-TASSE).

[d]Manual curation for obtaining the EC number was performed by biochemists according to the extracted knowledge from literatures.

[e]Not predictable is abbreviated with n.p., which means no results can be obtained based on current methods.

[f]Instead of obtaining EC number, only a protein name was provided.

[g]The result of EC number was obtained in this study.

[h]Although CLEAN was able to predict the fourth digit of EC number on flavin-dependent and α-ketoglutarate-dependent halogenases, only the third digit can be confidentially annotated based on the manual curation from reported studies.

Table S3. Enzyme kinetic analysis of MJ1651, TTHA0338 and SsFlA.

Enzyme	EC 3.13.1.8[a] (SAM hydrolase)		EC 2.5.1.63 (Fluorinase)		EC 2.5.1.94[b] (Chlorinase)	
	K_M (μM) [c]	k_{cat} (min^{-1})	K_M (μM) [d]	k_{cat} (min^{-1})	K_M (μM) [e]	k_{cat} (min^{-1})
MJ1651	82.6	0.327	-	-	-	-
TTHA0338	138.0	2.775	-	-	-	-
SsFlA[f]	15.8	0.017	34.6 (*30*)	0.22 (*30*)	10.7	0.008

[a] Assays contain various concentrations of SAM.

[b] Assays contain 100 mM NaCl and various concentrations of SAM.

[c] K_M refers to SAM K_M measured by conversion of SAM to adenosine.

[d] K_M refers to SAM K_M measured by conversion of SAM to 5'-FDA (*30*).

[e] K_M refers to SAM K_M measured by conversion of SAM to 5'-ClDA.

[f] SsFlA is also named as FlA1 or FlA^{M37} in previous studies (*30*, *80*).

References and Notes

1. UniProt Consortium, UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021). doi:10.1093/nar/gkaa1100 Medline

2. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Šmuc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, I. Friedberg, A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013). doi:10.1038/nmeth.2340 Medline

3. K. Hult, P. Berglund, Enzyme promiscuity: Mechanism and applications. *Trends Biotechnol.* **25**, 231–238 (2007). doi:10.1016/j.tibtech.2007.03.002 Medline

4. C. J. Jeffery, Protein moonlighting: What is it, and why is it important? *Phil. Trans. R. Soc. B* **373**, 20160523 (2018). doi:10.1098/rstb.2016.0523 Medline

5. E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, S. T. Sherry, Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022). doi:10.1093/nar/gkab1112 Medline

6. M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, R. D. Finn, The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021). doi:10.1093/nar/gkaa977 Medline

7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2 Medline

8. D. K. Desai, S. Nandi, P. K. Srivastava, A. M. Lynn, ModEnzA: Accurate identification of metabolic enzymes using function specific profile HMMs with optimised discrimination threshold and modified emission probabilities. *Adv. Bioinformatics* **2011**, 743782 (2011). doi:10.1155/2011/743782 Medline

9. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). doi:10.1093/nar/25.17.3389 Medline

10. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994). doi:10.1006/jmbi.1994.1104 Medline

11. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). doi:10.1186/s12859-019-3019-7 Medline

12. C. Zhang, P. L. Freddolino, Y. Zhang, COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017). doi:10.1093/nar/gkx366 Medline

13. A. Roy, J. Yang, Y. Zhang, COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40**, W471–W477 (2012). doi:10.1093/nar/gks372 Medline

14. J. Y. Ryu, H. U. Kim, S. Y. Lee, Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 13996–14001 (2019). doi:10.1073/pnas.1821905116 Medline

15. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, ProteInfer, deep neural networks for protein functional inference. *eLife* **12**, e80942 (2023). doi:10.7554/eLife.80942 Medline

16. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), pp. 18661–18673.

17. M. Heinzinger, M. Littmann, I. Sillitoe, N. Bordin, C. Orengo, B. Rost, Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinform.* **4**, lqac043 (2022). doi:10.1093/nargab/lqac043 Medline

18. F. Schroff, D. Kalenichenko, J. Philbin, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 815–823.

19. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021). doi:10.1073/pnas.2016239118 Medline

20. Y. Li, S. Wang, R. Umarov, B. Xie, M. Fan, L. Li, X. Gao, DEEPre: Sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* **34**, 760–769 (2018). doi:10.1093/bioinformatics/btx680 Medline

21. C. Yu, N. Zavaljevski, V. Desai, J. Reifman, Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases. *Proteins* **74**, 449–460 (2009). doi:10.1002/prot.22167 Medline

22. A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay, T. Doğan, ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* **19**, 334 (2018). doi:10.1186/s12859-018-2368-y Medline

23. M. N. Price, K. M. Wetmore, R. J. Waters, M. Callaghan, J. Ray, H. Liu, J. V. Kuehl, R. A. Melnyk, J. S. Lamson, Y. Suh, H. K. Carlson, Z. Esquivel, H. Sadeeshkumar, R. Chakraborty, G. M. Zane, B. E. Rubin, J. D. Wall, A. Visel, J. Bristow, M. J. Blow, A. P. Arkin, A. M. Deutschbauer, Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018). doi:10.1038/s41586-018-0124-0 Medline

24. C. Crowe, S. Molyneux, S. V. Sharma, Y. Zhang, D. S. Gkotsi, H. Connaris, R. J. M. Goss, Halogenases: A palette of emerging opportunities for synthetic biology-synthetic chemistry and C-H functionalisation. *Chem. Soc. Rev.* **50**, 9443–9481 (2021). doi:10.1039/D0CS01551B Medline

25. K. Prakinee, A. Phintha, S. Visitsatthawong, N. Lawan, J. Sucharitakul, C. Kantiwiriyawanitch, J. Damborsky, P. Chitnumsub, K.-H. van Pée, P. Chaiyen, Mechanism-guided tunnel engineering to increase the efficiency of a flavin-dependent halogenase. *Nat. Catal.* **5**, 534–544 (2022). doi:10.1038/s41929-022-00800-8

26. J. Latham, E. Brandenburger, S. A. Shepherd, B. R. K. Menon, J. Micklefield, Development of halogenase enzymes for use in synthesis. *Chem. Rev.* **118**, 232–269 (2018). doi:10.1021/acs.chemrev.7b00032 Medline

27. V. Agarwal, Z. D. Miles, J. M. Winter, A. S. Eustáquio, A. A. El Gamal, B. S. Moore, Enzymatic halogenation and dehalogenation reactions: Pervasive and mechanistically diverse. *Chem. Rev.* **117**, 5619–5674 (2017). doi:10.1021/acs.chemrev.6b00571 Medline

28. K. N. Rao, S. K. Burley, S. Swaminathan, Crystal structure of a conserved protein of unknown function (MJ1651) from *Methanococcus jannaschii*. *Proteins* **70**, 572–577 (2008). doi:10.1002/prot.21646 Medline

29. A. S. Eustáquio, J. Härle, J. P. Noel, B. S. Moore, S-Adenosyl-L-methionine hydrolase (adenosine-forming), a conserved bacterial and archaeal protein related to SAM-dependent halogenases. *ChemBioChem* **9**, 2215–2219 (2008). doi:10.1002/cbic.200800341 Medline

30. H. Sun, W. L. Yeo, Y. H. Lim, X. Chew, D. J. Smith, B. Xue, K. P. Chan, R. C. Robinson, E. G. Robins, H. Zhao, E. L. Ang, Directed evolution of a fluorinase for improved fluorination efficiency with a non-native substrate. *Angew. Chem. Int. Ed.* **55**, 14277–14280 (2016). doi:10.1002/anie.201606722 Medline

31. H. Nam, N. E. Lewis, J. A. Lerman, D.-H. Lee, R. L. Chang, D. Kim, B. O. Palsson, Network context and selection in the evolution to enzyme specificity. *Science* **337**, 1101–1104 (2012). doi:10.1126/science.1216861 Medline

32. O. Shalem, N. E. Sanjana, F. Zhang, High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015). doi:10.1038/nrg3899 Medline

33. Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, H. Zhao, Directed evolution: Methodologies and applications. *Chem. Rev.* **121**, 12384–12444 (2021). doi:10.1021/acs.chemrev.1c00260 Medline

34. H. Zhao, The ever-expanding boundaries of synthetic biology. *ACS Synth. Biol.* **11**, 3550 (2022). doi:10.1021/acssynbio.2c00582 Medline

35. X.-M. Sun, Y.-S. Xu, H. Huang, Thraustochytrid cell factories for producing lipid compounds. *Trends Biotechnol.* **39**, 648–650 (2021). doi:10.1016/j.tibtech.2020.10.008 Medline

36. G. B. Kim, W. J. Kim, H. U. Kim, S. Y. Lee, Machine learning applications in systems metabolic engineering. *Curr. Opin. Biotechnol.* **64**, 1–9 (2020). doi:10.1016/j.copbio.2019.08.010 Medline

37. T. Yu, A. G. Boob, M. J. Volk, X. Liu, H. Cui, H. Zhao, Machine learning-enabled retrobiosynthesis of molecules. *Nat. Catal.* **6**, 137–151 (2023). doi:10.1038/s41929-022-00909-w

38. D. Rother, S. Malzacher, Computer-aided enzymatic retrosynthesis. *Nat. Catal.* **4**, 92–93 (2021). doi:10.1038/s41929-021-00582-5

39. S. Satoh, M. Yao, Y. Kousumi, A. Ebihara, K. Matsumoto, A. Okamoto, I. Tanaka, S. Yokoyama, S. Kuramitsu, RIKEN Structural Genomics/Proteomics Initiative (RSGI), Crystal Structure of the Conserved Hypothetical Protein TTHA1091 from *Thermus thermophilus* HB8 (PDB, Entry 1VGG, 2004); http://doi.org/10.2210/pdb1vgg/pdb.

40. K. Shimizu, N. Kunishima, RIKEN Structural Genomics/Proteomics Initiative (RSGI), Crystal structure of project PH0463 from *Pyrococcus horikoshii* OT3 (PDB, Entry 1WU8, 2005); http://doi.org/10.2210/pdb1wu8/pdb.

41. A. S. Eustáquio, F. Pojer, J. P. Noel, B. S. Moore, Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nat. Chem. Biol.* **4**, 69–74 (2008). doi:10.1038/nchembio.2007.56 Medline

42. C. Dong, F. Huang, H. Deng, C. Schaffrath, J. B. Spencer, D. O'Hagan, J. H. Naismith, Crystal structure and mechanism of a bacterial fluorinating enzyme. *Nature* **427**, 561–565 (2004). doi:10.1038/nature02280 Medline

43. T. Yu, H. Cui, C. Li, Y. Luo, G. Jiang, H. Zhao, Enzyme function prediction using contrastive learning, version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.7582241.

44. V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, in *Procedings of the British Machine Vision Conference 2016*, R. C. Wilson, E. R. Hancock, W. A. P. Smith, Eds. (BMVA, 2016), pp. 119.1–119.11.

45. K. Sohn, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, Inc., 2016), pp. 1857–1865.

46. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

47. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, D. S. Marks, Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021). doi:10.1038/s41586-021-04043-8 Medline

48. H. Cui, S. Pramanik, K.-E. Jaeger, M. D. Davari, U. Schwaneberg, CompassR-guided recombination unlocks design principles to stabilize lipases in ILs with minimal experimental efforts. *Green Chem.* **23**, 3474–3486 (2021). doi:10.1039/D1GC00763G

49. H. Cui, L. Eltoukhy, L. Zhang, U. Markel, K.-E. Jaeger, M. D. Davari, U. Schwaneberg, Less unfavorable salt bridges on the enzyme surface result in more organic cosolvent resistance. *Angew. Chem. Int. Ed.* **60**, 11448–11456 (2021). doi:10.1002/anie.202101642 Medline

50. H. Deng, L. Ma, N. Bandaranayaka, Z. Qin, G. Mann, K. Kyeremeh, Y. Yu, T. Shepherd, J. H. Naismith, D. O'Hagan, Identification of fluorinases from *Streptomyces* sp MA37, *Norcardia brasiliensis*, and *Actinoplanes* sp N902-109 by genome mining. *ChemBioChem* **15**, 364–368 (2014). doi:10.1002/cbic.201300732 Medline

51. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017). doi:10.1038/nbt.3988 Medline

52. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

53. H. Deng, C. H. Botting, J. T. G. Hamilton, R. J. M. Russell, D. O'Hagan, *S*-adenosyl-L-methionine:hydroxide adenosyltransferase: A SAM enzyme. *Angew. Chem. Int. Ed.* **47**, 5357–5361 (2008). doi:10.1002/anie.200800794 Medline

54. H. Deng, S. L. Cobb, A. R. McEwan, R. P. McGlinchey, J. H. Naismith, D. O'Hagan, D. A. Robinson, J. B. Spencer, The fluorinase from *Streptomyces cattleya* is also a chlorinase. *Angew. Chem. Int. Ed.* **45**, 759–762 (2006). doi:10.1002/anie.200503582 Medline

55. J. L. Hoffman, Chromatographic analysis of the chiral and covalent instability of S-adenosyl-L-methionine. *Biochemistry* **25**, 4444–4449 (1986). doi:10.1021/bi00363a041 Medline

56. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014). doi:10.1093/nar/gku316 Medline

57. P. Bernhardt, T. Okino, J. M. Winter, A. Miyanaga, B. S. Moore, A stereoselective vanadium-dependent chloroperoxidase in bacterial antibiotic biosynthesis. *J. Am. Chem. Soc.* **133**, 4268–4270 (2011). doi:10.1021/ja201088k Medline

58. L. A. M. Murray, S. M. K. McKinnie, B. S. Moore, J. H. George, Meroterpenoid natural products from *Streptomyces* bacteria - the evolution of chemoenzymatic syntheses. *Nat. Prod. Rep.* **37**, 1334–1366 (2020). doi:10.1039/D0NP00018C Medline

59. J. R. Heemstra Jr., C. T. Walsh, Tandem action of the O_2- and $FADH_2$-dependent halogenases KtzQ and KtzR produce 6,7-dichlorotryptophan for kutzneride assembly. *J. Am. Chem. Soc.* **130**, 14024–14025 (2008). doi:10.1021/ja806467a Medline

60. T. P. Cardoso, L. A. de Sá, P. D. S. Bury, S. M. Chavez-Pacheco, M. V. B. Dias, Cloning, expression, purification and biophysical analysis of two putative halogenases from the

glycopeptide A47,934 gene cluster of *Streptomyces toyocaensis*. *Protein Expr. Purif.* **132**, 9–18 (2017). doi:10.1016/j.pep.2017.01.001 Medline

61. T. Chilczuk, T. F. Schäberle, S. Vahdati, U. Mettal, M. El Omari, H. Enke, M. Wiese, G. M. König, T. H. J. Niedermeyer, Halogenation-guided chemical screening provides insight into tjipanazole biosynthesis by the Cyanobacterium *Fischerella ambigua*. *ChemBioChem* **21**, 2170–2177 (2020). doi:10.1002/cbic.202000025 Medline

62. D. S. Gkotsi, H. Ludewig, S. V. Sharma, J. A. Connolly, J. Dhaliwal, Y. Wang, W. P. Unsworth, R. J. K. Taylor, M. M. W. McLachlan, S. Shanahan, J. H. Naismith, R. J. M. Goss, A marine viral halogenase that iodinates diverse substrates. *Nat. Chem.* **11**, 1091–1097 (2019). doi:10.1038/s41557-019-0349-z Medline

63. V. Agarwal, A. A. El Gamal, K. Yamanaka, D. Poth, R. D. Kersten, M. Schorn, E. E. Allen, B. S. Moore, Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat. Chem. Biol.* **10**, 640–647 (2014). doi:10.1038/nchembio.1564 Medline

64. L. Heide, L. Westrich, C. Anderle, B. Gust, B. Kammerer, J. Piel, Use of a halogenase of hormaomycin biosynthesis for formation of new clorobiocin analogues with 5-chloropyrrole moieties. *ChemBioChem* **9**, 1992–1999 (2008). doi:10.1002/cbic.200800186 Medline

65. M. Sato, J. M. Winter, S. Kishimoto, H. Noguchi, Y. Tang, K. Watanabe, Combinatorial generation of chemical diversity by redox enzymes in chaetoviridin biosynthesis. *Org. Lett.* **18**, 1446–1449 (2016). doi:10.1021/acs.orglett.6b00380 Medline

66. J. M. Winter, M. C. Moffitt, E. Zazopoulos, J. B. McAlpine, P. C. Dorrestein, B. S. Moore, Molecular basis for chloronium-mediated meroterpene cyclization: Cloning, sequencing, and heterologous expression of the napyradiomycin biosynthetic gene cluster. *J. Biol. Chem.* **282**, 16362–16368 (2007). doi:10.1074/jbc.M611046200 Medline

67. B. R. K. Menon, D. Richmond, N. Menon, Halogenases for biosynthetic pathway engineering: Toward new routes to naturals and non-naturals. *Catal. Rev.* **64**, 533–591 (2020). doi:10.1080/01614940.2020.1823788

68. L. Xu, T. Han, M. Ge, L. Zhu, X. Qian, Discovery of the new plant growth-regulating compound LYXLF2 based on manipulating the halogenase in *Amycolatopsis orientalis*. *Curr. Microbiol.* **73**, 335–340 (2016). doi:10.1007/s00284-016-1052-6 Medline

69. W.-Y. Wang, S.-B. Yang, Y.-J. Wu, X.-F. Shen, S.-X. Chen, Enhancement of A82846B yield and proportion by overexpressing the halogenase gene in *Amycolatopsis orientalis* SIPI18099. *Appl. Microbiol. Biotechnol.* **102**, 5635–5643 (2018). doi:10.1007/s00253-018-8983-8 Medline

70. J. J. Banik, S. F. Brady, Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17273–17277 (2008). doi:10.1073/pnas.0807564105 Medline

71. D. P. Galonić, F. H. Vaillancourt, C. T. Walsh, Halogenation of unactivated carbon centers in natural product biosynthesis: Trichlorination of leucine during barbamide biosynthesis. *J. Am. Chem. Soc.* **128**, 3900–3901 (2006). doi:10.1021/ja060151n Medline

72. C. Wong, D. G. Fujimori, C. T. Walsh, C. L. Drennan, Structural analysis of an open active site conformation of nonheme iron halogenase CytC3. *J. Am. Chem. Soc.* **131**, 4872–4879 (2009). doi:10.1021/ja8097355 Medline

73. A. Timmins, N. J. Fowler, J. Warwicker, G. D. Straganz, S. P. de Visser, Does substrate positioning affect the selectivity and reactivity in the hectochlorin biosynthesis halogenase? *Front Chem.* **6**, 513 (2018). doi:10.3389/fchem.2018.00513 Medline

74. M. L. Hillwig, Q. Zhu, K. Ittiamornkul, X. Liu, Discovery of a promiscuous non-heme iron halogenase in ambiguine alkaloid biogenesis: Implication for an evolvable enzyme family for late-stage halogenation of aliphatic carbons in small molecules. *Angew. Chem. Int. Ed.* **55**, 5780–5784 (2016). doi:10.1002/anie.201601447 Medline

75. C. Zhao, S. Yan, Q. Li, H. Zhu, Z. Zhong, Y. Ye, Z. Deng, Y. Zhang, An Fe^{2+}- and α-ketoglutarate-dependent halogenase acts on nucleotide substrates. *Angew. Chem. Int. Ed.* **59**, 9478–9484 (2020). doi:10.1002/anie.201914994 Medline

76. P. Moosmann, R. Ueoka, M. Gugger, J. Piel, Aranazoles: Extensively chlorinated nonribosomal peptide–polyketide hybrids from the Cyanobacterium *Fischerella* sp. PCC 9339. *Org. Lett.* **20**, 5238–5241 (2018). doi:10.1021/acs.orglett.8b02193 Medline

77. M. E. Neugebauer, K. H. Sumida, J. G. Pelton, J. L. McMurry, J. A. Marchand, M. C. Y. Chang, A family of radical halogenases for the engineering of amino-acid-based products. *Nat. Chem. Biol.* **15**, 1009–1016 (2019). doi:10.1038/s41589-019-0355-x Medline

78. S. Huang, L. Ma, M. H. Tong, Y. Yu, D. O'Hagan, H. Deng, Fluoroacetate biosynthesis from the marine-derived bacterium *Streptomyces xinghaiensis* NRRL B-24674. *Org. Biomol. Chem.* **12**, 4828–4831 (2014). doi:10.1039/C4OB00970C Medline

79. H. Sun, H. Zhao, E. L. Ang, A coupled chlorinase-fluorinase system with a high efficiency of *trans*-halogenation and a shared substrate tolerance. *Chem. Commun.* **54**, 9458–9461 (2018). doi:10.1039/C8CC04436H Medline

80. I. Pardo, D. Bednar, P. Calero, D. C. Volke, J. Damborský, P. I. Nikel, A nonconventional archaeal fluorinase identified by in silico mining for enhanced fluorine biocatalysis. *ACS Catal.* **12**, 6570–6577 (2022). doi:10.1021/acscatal.2c01184 Medline