Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3]†, Haiyang Cui[1,2,3]†, Jianan Canal Li[3,4], Yunan Luo[5], and Huimin Zhao[1,2,3,6]*

[1]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Carl R. Woese Institute for Genomic Biology
[3]NSF Molecule Maker Lab Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[4]Department of Computer Science, Cornell University, Ithaca, NY 14850
[5]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30308
[6]DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL 61801
*Corresponding author. Email: zhao5@illinois.edu
†These authors contributed equally to this work

Abstract: Enzyme function annotation is a fundamental challenge in protein science and numerous computational tools have been developed. However, most of these tools cannot accurately predict functional annotations such as enzyme commission (EC) number for less studied or proteins with novel or multiple functions. Herein, we present a machine learning algorithm named CLEAN (contrastive learning enabled enzyme annotation) to assign EC numbers to enzymes with better accuracy, reliability, and sensitivity than the state-of-the-art tool BLASTp. CLEAN learns an enzyme function informed sequence representation whose Euclidean distance reflects functional similarity. The transformative contrastive learning framework empowers CLEAN to confidently: i) characterize the understudied or unannotated enzymes, ii) correct the mislabeled enzymes, and iii) identify the promiscuous enzymes, which are demonstrated by systematic *in silico* and *in vitro* experiments. We expect this tool to greatly facilitate functional genomics, enzymology, and synthetic biology studies.

One-Sentence Summary: A machine learning algorithm was developed for enzyme function annotation which outperforms state-of-the-art computational tools.

Main Text: The recent genomics revolution has led to the discovery of numerous protein sequences from organisms across all branches of life. For example, UniProt Knowledgebase has cataloged approximately 190 million protein sequences. However, only less than 0.3% (~half a million) of these proteins were reviewed by human curators, out of which less than 19.4% are supported by clear experimental evidence (*1*). Consequently, protein function annotation is highly dependent on computational annotation methods. However, the study on large-scale community-based critical assessment of protein function annotation (CAFA) found that approximately 40% of the automatically annotated enzymes using existing computational tools are incorrectly annotated (*2*). Therefore, functional annotation of proteins remains an overwhelming challenge in protein science. Particularly, the inequality in protein annotation of understudied and promiscuous proteins has impeded biomedical progress and drug discovery (*3*, *4*).

Enzyme Commission (EC) system is the most well-known numerical classification scheme of enzymes, which specifies the catalytic function of an enzyme by four digits. As an experimental characterization of the function of a target enzyme is often laborious and expensive, numerous computational tools for enzyme function annotation have been developed (*1*, *5*, *6*). They include but are not limited to sequence similarity-based (*7*–*9*), homology-based (*10*, *11*), structure-based (*12*, *13*), and machine learning (ML)-based (*14*, *15*) approaches. Among them, sequence similarity-based Basic Local Alignment Search Tools for proteins (BLASTp) is the most widely used tool (*7*). However, BLASTp and other alignment tools annotate functions based solely on sequence similarity, making the prediction result less reliable when sequence similarity is low. Notably, almost all the existing ML models, such as ProteInfer (*15*) and DeepEC (*14*) are based on a multi-label classification framework and suffer from the limited and imbalanced training dataset that is common in biology. Therefore, a robust tool with better accuracy and EC coverage is required to

unlock the potential of currently uncharacterized proteins and understand the extent to which proteins are capable.

Here, we report an ML model named CLEAN (contrastive learning enabled enzyme annotation) for enzyme function prediction. CLEAN was trained on high-quality data from UniProt, taking amino acid sequence as input and outputting a list of enzyme functions (EC numbers as the example) ranked by the likelihood. To validate the accuracy and robustness of CLEAN, we performed extensively *in silico* experiments. Furthermore, we challenged CLEAN to annotate EC numbers for an in-house collected database of all uncharacterized halogenases (in total 36) followed by case studies as *in vitro* experimental validation. CLEAN significantly outperformed other EC number annotation tools, including BLASTp and state-of-the-art ML models.

Unlike previously developed ML algorithms that frame EC number prediction tasks as a multi-label classification problem, CLEAN employed a contrastive learning framework (*16*). Our training objective is to learn a representation space of enzymes where the Euclidean distance reflects the functional similarities, i.e., the amino acid sequences with the same EC number have a small Euclidean distance, while sequences with different EC numbers have a considerable distance. Contrastive losses were used to train the model with supervision (*17*, *18*). During the training process (**Fig. 1a**), each reference sequence (anchor) in the training dataset was sampled with a sequence with the same EC number (positive) and a sequence with a different EC number (negative). Aiming to facilitate training efficiency by providing the model with challenging negative samples, instead of drawing them randomly, negative sequences whose embeddings had a small Euclidean distance with the anchor were prioritized.

In the training stage, the protein representation obtained from the language model ESM-1b (*19*) was used as the input of a feedforward neural network, whose output layer produced a refined, function-aware embedding of the input protein. The learning objective is a contrastive loss function that minimizes the distance between the anchor and the positive while maximizing the distance between the anchor and the negative. When making predictions, the representation of an EC number cluster center was obtained by averaging the learned embeddings of all sequences in the training set belonging to that EC number (**Fig. 1b**). Subsequently, the pairwise distances between the query sequence and all EC number cluster centers were calculated. EC numbers of clusters that are significantly close to the query sequence are predicted as the EC numbers for the input protein (Supplementary Text 1).

The database used for model development and evaluation was a universal protein knowledgebase UniProt (*1*). Two EC selection methods were developed to predict confident EC numbers from the output ranking **(Fig. 1c)**: i) a greedy approach that prioritizes EC numbers that have the maximum separation from other EC numbers in terms of the pairwise distance to the query sequence; ii) a *p*-value-based method that identifies EC numbers that have a significantly smaller distance to the query sequence. On a train-test set split based on the clustering at 50% sequence identity, CLEAN achieved a 0.865 F1-score, a commonly used accuracy metric indicating the harmonic mean of precision and recall. Even at 10% clustering split, CLEAN reached a 0.67 F1-score. In addition, CLEAN achieved much higher performance than the baseline method using ESM-1b without contrastive learning **(Fig. S1).**

After training, the prediction performance of CLEAN was systematically investigated by comparing it to five state-of-the-art EC number annotation tools (e.g., ProteInfer (*15*), DeepEC (*14*), BLASTp, DEEPre (*20*), and ECPred (*21*)). Two independent datasets not included in any model's development were used to deliver a fair and rigorous benchmark study. The first dataset, New-392, consisted of 392 enzyme sequences covering 177 different EC numbers, containing data from Swiss-Prot released after CLEAN was trained (2022/04). The prediction scenario represented a practical situation where the labeled knowledgebase was the Swiss-Prot database, and functions of query sequences were unknown. Overall, CLEAN resulted in the highest value in various multi-label accuracy metrics, including precision (0.566), recall (0.479), and area under curve (AUC, 0.739), when compared to ProteInfer and DeepEC (**Fig. 2a**). For example, CLEAN achieved an F1-score of 0.49, while ProteInfer and DeepEC had 0.309 and 0.23, respectively.

The second independent dataset, denoted as Price-149, was a set of experimentally validated results described by Price et al. (*22*). Price-149 dataset was first curated by ProteInfer (*23*) as a challenging dataset, as the existing sequences were determined to be incorrectly or inconsistently labeled in databases like Kyoto Encyclopedia of Genes and Genomes (KEGG) by automated annotation methods. Again, CLEAN achieved the highest F1-score (0.413) in comparison with BLASTp, ProteInfer, and DeepEC (**Fig. 2b**). Notably, in this challenging task, CLEAN had a 2.5-fold higher F1-score than ProteInfer (0.163) and an almost 4-fold increase than DeepEC (0.107). The evaluations on the New-392 and Price-149 dataset proved that CLEAN is more precise and reliable than previously developed ML-models for predicting newly discovered proteins, especially the ones without known enzyme functions.

Next, we investigated why CLEAN performs better than other ML models on understudied EC numbers. We curated a validation dataset with enzymes from rare EC numbers to prove our hypothesis that compared with the multi-label classification framework, contrastive learning could better handle the imbalanced nature of EC numbers where some EC numbers have thousands of enzyme examples, and some only have very few (less than 5). In this validation dataset, each type of EC number had no more than five occurrences, and over 3000 samples were included in this dataset covering over 1000 different EC numbers. Note that ProteInfer and DeepEC were evaluated using their released pretrained models, thus, our curated validation set appeared during both models' training process. Despite this added advantage, CLEAN outperformed the both methods by achieving a 0.817 F1-score (**Fig. 2c**).

Furthermore, CLEAN's performance was analyzed based on the number of times the EC number occurred in the training set. Even at 50% clustering split based on sequence identity, where the test set and train set had a low similarity, CLEAN's performance did not drop significantly when the number of training examples was scarce **(Fig. 2d).** With the given results, the two independent datasets (New-392 and Price-149) were combined and revisited. As shown in **Fig. 2e**, the accuracy performance was studied separately based on the number of times EC numbers appeared in the training set. As expected, ProteInfer and DeepEC showed a significant bias toward popular EC numbers, limited by the classification framework. In contrast, CLEAN showed the most superiority in predicting understudied functions and maintained high accuracy regardless of the EC occurrences. The challenge posed by the biased dataset to the classification model was the lack of positive examples for understudied EC numbers. As a result, classification models can hardly learn from the limited positive examples. However, as demonstrated by analyzing triplet margin loss and Supcon-Hard loss, such a situation did not happen in CLEAN **(Fig. S2**, see more analysis in the Supplementary Material).

Next, we sought to validate the prediction accuracy of CLEAN in assigning EC numbers using halogenases as a proof-of-concept study. Halogenases have been increasingly used for biocatalytic C-H functionalization because of their excellent catalyst-controlled selectivity (*24*, *25*). Generally, small molecules with halogen atom(s) produced by halogenases have promising bioactivity and physicochemical properties, thereby offering broad application in pharmaceutical and agrochemical fields (*24*, *26*, *27*). To date, 36 incompletely annotated halogenases were identified from UniProt, covering all four types of halogenases (haloperoxidase, flavin-dependent, α-ketoglutarate (α-KG)-dependent, and S-adenosyl-methionine (SAM)-dependent halogenases, **Fig. 3a,** and **Table S2**). These halogenases were either labeled with uncharacterized and/or hypothetical proteins in UniProt or had conflicting annotations in the literature. With expert curation and experimental validations showing later, all 36 halogenases were confidentially annotated with EC numbers. Overall, CLEAN achieved much better prediction accuracy (100-86.7%, **Fig. 2f and 3a**), ranging from the first digit prediction to the fourth digit in EC number compared to the six other commonly used computational tools (e.g., ~11.1% in DeepEC and 11.1-61.1% in ProteInfer). These results demonstrate that CLEAN can distinguish enzyme functions even within the regime of similar biocatalytic reactions.

Among 36 halogenases, three enzymes named MJ1651, TTHA0338, and SsFlA showed "conflicting" functions, according to the comparison between literature (*28–30*) and the description in UniProt. Interestingly, CLEAN predicted new EC numbers in these three cases, suggesting other potential

function(s) might occur. Therefore, we performed *in vitro* experiments to validate predictions. Surprisingly, HPLC and LC-MS/MS analysis confirmed MJ1651 is *S*-Adenosyl-L-Methionine (SAM) hydrolase (EC 3.13.1.8) as CLEAN predicted, rather than chlorinase (EC 2.5.1.94) that was mislabeled in UniProt (**Fig. 3d, Figs. 3f-g** and **3m**, **Figs. S3** and **S4a-b** in SI). CLEAN also successfully annotated TTHA0338 with EC 3.13.1.8 which belongs to the DUF62 Pfam family with no known function (**Figs. 3d, 3h,** and **3n**). Notably, all other six commonly used computational tools failed to predict MJ1641 and TTHA0338, except BLASTp made a success on the target TTHA0338. These results revealed CLEAN is favorable for correcting the mislabeled enzymes and accurately identifying the understudied catalytic function. Moreover, CLEAN confidently identified the promiscuous enzyme SsFlA with three EC numbers (EC 2.5.1.63; EC 2.5.1.94; EC 3.13.1.8, **Figs. 3e**, **3i-k**, and **3o-q**). These observations confirmed CLEAN could effectively recall the defined biological activity and capture elements of enzyme promiscuity. Meanwhile, the precision of CLEAN was impressive in distinguishing the SAM-binding proteins with low sequence identity (max. 16.86%, **Figs. 3b** and **S5-S6**) but with homologous structures (**Fig. S3c**). These results suggest our sequence-based model CLEAN performed better than structure-based methods (e.g., COFACTORS (*12*, *13*)) in dealing with enzymes with similar structures but different functions.

Through systematic *in silico* and *in vitro* experimental validations, we demonstrated that CLEAN achieved much better prediction performance than six state-of-the-art tools (i.e., ProteInfer, BLASTp, DeepEC, DEEPre, COFACTOR, and ECPred). More significantly, the comprehensive performance analysis on the uncharacterized halogenase dataset indicated that CLEAN can characterize the hypothetical proteins and correct the mislabeled proteins, where most sequence-, structure-, and ML-based annotation tools predict incorrectly or are unable to produce a prediction. Identifying enzyme promiscuity is essential for improving the performance of existing enzymes (*3*, *31*), which can be effectively achieved by CLEAN (e.g., SsFlA with three functions). Unlike classification models, contrastive learning is more suitable for biological data, which is usually imbalanced/biased and scarce.

Overall, CLEAN is a powerful tool for predicting the catalytic function of query enzymes, which can greatly facilitate studies in functional genomics, enzymology, enzyme engineering, synthetic biology, metabolic engineering, and retrobiosynthesis. Moreover, the general language model representation topped with contrastive learning workflow used by CLEAN can be readily adapted to other prediction tasks not limited to enzymatic activities, such as Functional Catalogue (FunCat) and Gene Ontology (GO). The user-friendly feature of our framework allows CLEAN to be used as an independent tool in a high-throughput manner and a software component integrated into other computational platforms. Most importantly, the superior performance of CLEAN in predicting understudied proteins would greatly expand the bioinformatics toolbox, thereby laying the cornerstone for future detailed mechanistic studies.

References and Notes

1. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*. **49**, D480–D489 (2021).
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).

3. K. Hult, P. Berglund, Enzyme promiscuity: mechanism and applications. *Trends Biotech.* **25**, 231–238 (2007).
4. C. J. Jeffery, Protein moonlighting: what is it, and why is it important? *Phil. Trans. R. Soc. B*. **373**, 20160523 (2018).
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).
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).
7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology*. **215**, 403–410 (1990).
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 Bioinform*. **2011**, 1–12 (2011).
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 Research*. **25**, 3389–3402 (1997).
10. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology*. **235**, 1501–1531 (1994).
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).
12. C. Zhang, P. L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research*. **45**, W291–W299 (2017).
13. A. Roy, J. Yang, Y. Zhang, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*. **40**, W471–W477 (2012).
14. J. Y. Ryu, H. U. Kim, S. Y. Lee, Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *PNAS*. **116**, 13996–14001 (2019).
15. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, ProteInfer: deep networks for protein functional inference (2021), p. 2021.09.20.461077, , doi:10.1101/2021.09.20.461077.
16. S. Chopra, R. Hadsell, Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification" in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE, San Diego, CA, USA, 2005; http://ieeexplore.ieee.org/document/1467314/), vol. 1, pp. 539–546.
17. F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering" in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015; http://arxiv.org/abs/1503.03832), pp. 815–823.
18. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, "Supervised Contrastive Learning" (arXiv:2004.11362, arXiv, 2021), doi:10.48550/arXiv.2004.11362.

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 (2020), p. 622803.
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).
21. 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 Bioinform*. **19**, 334 (2018).
22. 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).
23. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, ProteInfer: deep networks for protein functional inference (2021), p. 2021.09.20.461077, , doi:10.1101/2021.09.20.461077.
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).
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).
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).
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).
28. K. N. Rao, S. K. Burley, S. Swaminathan, Crystal structure of a conserved protein of unknown function (MJ1651) from Methanococcus jannaschii. *Proteins: Structure, Function, and Bioinformatics*. **70**, 572–577 (2008).
29. A. S. Eustáquio, J. Härle, J. P. Noel, B. S. Moore, *S* -adenosyl-L-methionine hydrolase (adenosine-forming), a conserved bacterial and archeal protein related to SAM-dependent halogenases. *ChemBioChem*. **9**, 2215–2219 (2008).
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).
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).

Acknowledgments: We thank Dr. Hengqian Ren and Dr. Chunshuai Huang for their suggestions on the characterization of the products formed by halogenases. **Funding:** This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by U.S. National Science Foundation under grant no. 2019897 (H.Z.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the NSF. **Author contributions:** T.Y., Y.L., and H.Z. conceived of the presented idea. T.Y. implemented the computational framework. T.Y., Y.L, H.Z., and J.L. designed the *in silico* experiments. T.Y. and J.L. contributed to data preparation and carried out the *in silico* experiments. J.L. contributed to code cleaning up. H.C. and H.Z. planned the *in vitro* experiments. H.C. carried out *in vitro* experiments and data analysis. T.Y. and H.C. wrote the manuscript with input from all authors. All authors discussed the results and contributed to the final manuscript. H.Z. provided supervision and resources for this study. **Competing interests:** Authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

Supplementary Materials
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S2
References (*1–36*)

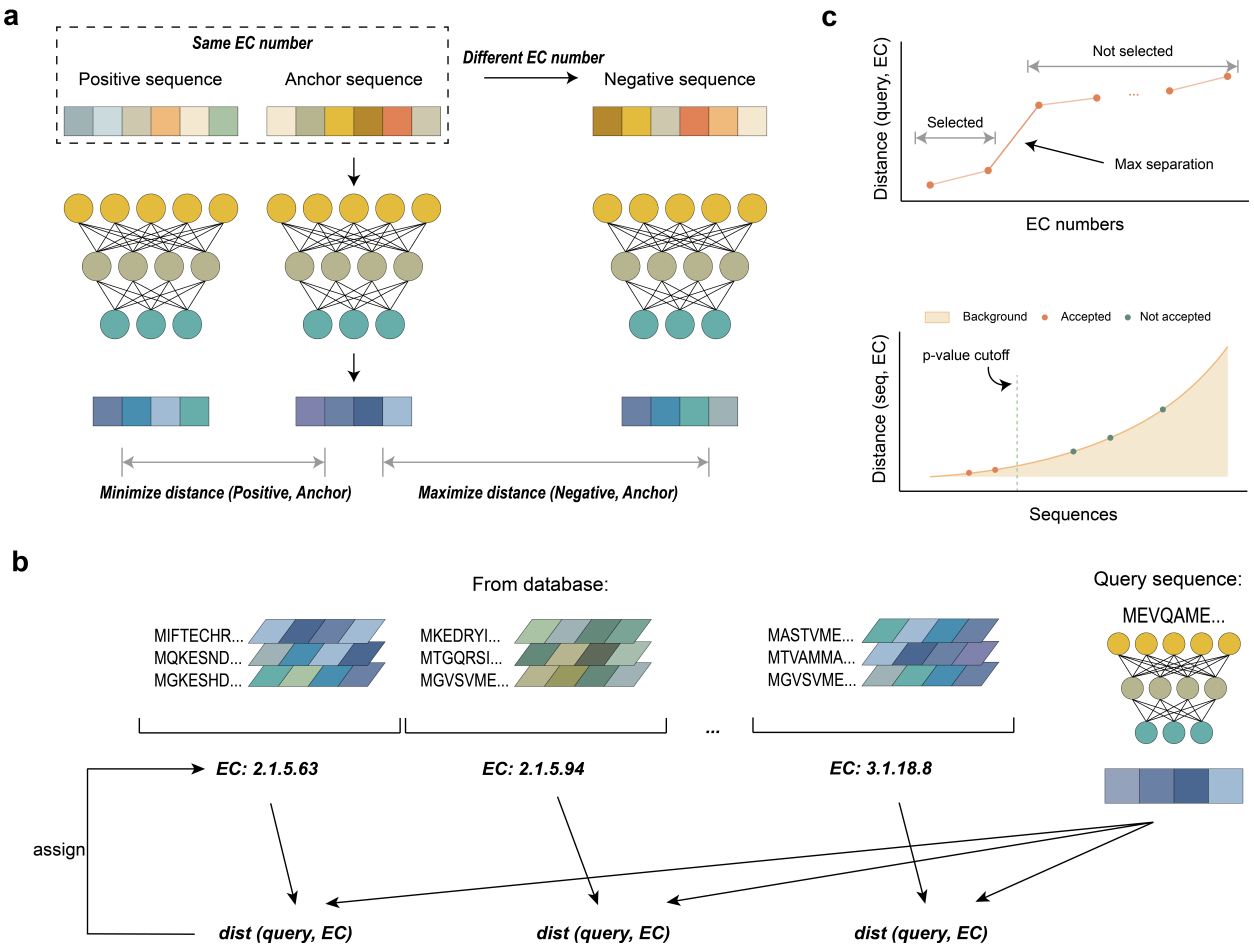

Fig. 1. The contrastive learning-based framework of CLEAN. (**a**) During training, positives and negatives are sampled based on EC numbers. The input sequences were embedded and passed through a neural network. (**b**) The representations of an EC number are obtained by averaging the representations of enzymes under this EC number. When predicting the EC number, the query sequence embedding was compared with each EC number's representation to obtain the pairwise Euclidean distance between the query sequence and each EC number. The distance reflects the similarity between EC numbers and the query sequence. (**c**) When used as a classification model, two methods, Max-separation (above) and p-value (below), were implemented to prioritize confident predictions of EC numbers from the ranking order.

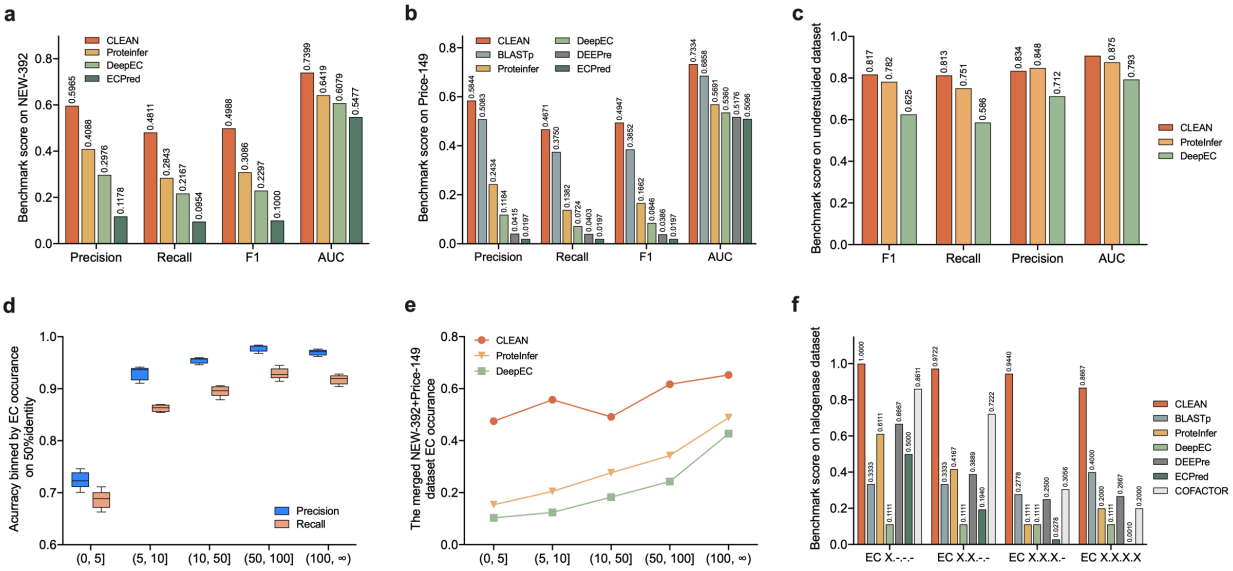

Fig. 2. **Quantitative comparison of CLEAN with the state-of-the-art EC number prediction tools.** (**a**) Evaluation of CLEAN's performance towards four multi-label accuracy metrics (F1-score, recall, precision, and AUC) examined on the NEW-392 database. AUC stands for "Area Under the receiver operating characteristic (ROC) Curve." Three top-ranked models, ProteInfer, DeepEC and ECPred were used for comparison. (**b**) Comparison of CLEAN, BLASTp, ProteInfer, and DeepEC on the Price-149 database. (**c**) Comparison of CLEAN, ProteInfer, and DeepEC on a dataset of underrepresented EC numbers. (**d**) The accuracy binned plot of CLEAN under 50% identity clustering split using the SupconH loss. Precision and recall values were binned by the number of times the EC number appeared in the training set. The box plots showed the results of 5-fold cross-validation. (**e**) Evaluation on the combined datasets of Price-149 and New-392 binned by the number of times the EC number appeared in CLEAN's training dataset. (**f**) Prediction accuracy of CLEAN on an in-house curated halogenase dataset compared to six commonly used tools (BLASTp, ProteInfer, DeepEC, DEEPre, ECPred, and COFACTOR). This dataset had good diversity covering eleven different EC numbers.

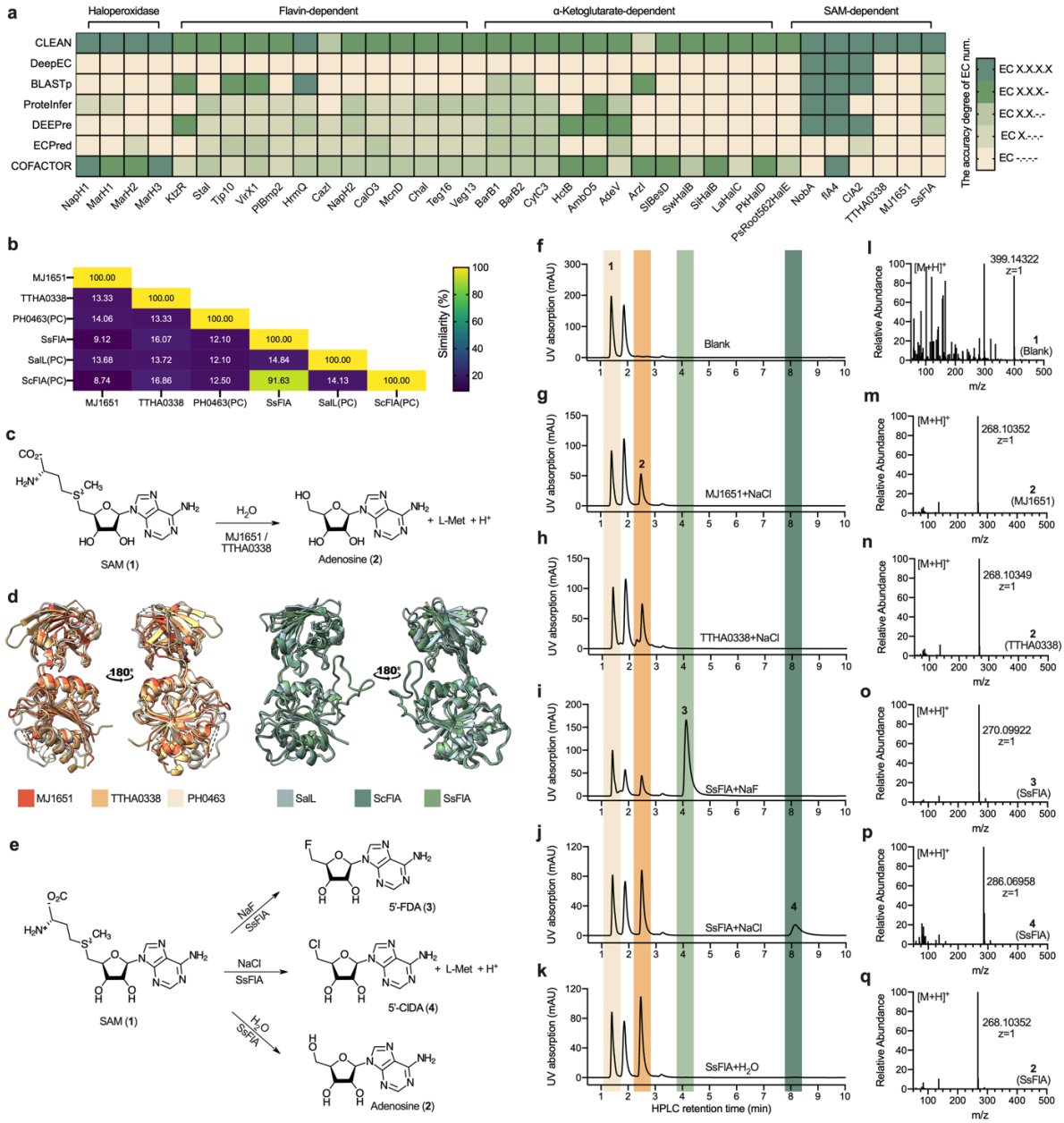

Fig. 3. **Experimental validation of CLEAN on uncharacterized halogenases.** (**a**) The accuracy degree heatmap of EC numerical ID was shown for the 36 identified halogenases; (**b**) Heatmap of sequence similarity among the uncharacterized proteins and positive control (PC) enzymes. Color bar with "viridis" color scale was in percentage (%); (**c**) The SAM hydroxide adenosyltransferase MJ1651/TTHA0338 reaction; (**d**) The structural supposition of the entire structures of uncharacterized proteins MJ1651/TTHA0338 (or SsFlA) and PC enzyme PH0463 (or SalL and ScFlA); (**e**) Nucleophilic substitution of SAM with halide ions or H_2O towards SsFlA; HPLC analysis of reaction mixtures of SAM and NaCl/NaF/H_2O with (**f**) blank, (**g**) purified MJ1651, (**h**) purified TTHA0338, and (**i-k**) purified SsFlA. The peaks of substrate SAM (**1**), product adenosine (**2**), 5'-fluoro-5'-deoxyadenosine (5'-FDA) (**3**), and 5'-chloro-5'-deoxyadenosine (5'-ClDA) (**4**) were labeled with light yellow, orange, green, and dark green, respectively, which were also aligned at the same retention time; Mass spectra of compounds obtained from the reaction mixtures: (**l**) substrate **1** in the blank reaction system, (**m**) adenosine (**2**) in MJ1651 catalyzed

reaction, (**n**) adenosine (**2**) in TTHA0338 catalyzed reaction, and (**o**) 5'-FDA (**3**), (**p**) 5'-ClDA (**4**), and (**q**) adenosine (**2**).

Science
AAAS

Supplementary Materials for

Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3]†, Haiyang Cui[1,2,3]†, Jianan Canal Li[3,4], Yunan Luo[5] and Huimin Zhao[1,2,3,6]*

[1]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Carl R. Woese Institute for Genomic Biology
[3]NSF Molecule Maker Lab Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[4]Department of Computer Science, Cornell University, Ithaca, NY 14850
[5]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30308
[6]DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL 61801,

*Corresponding author. Email: zhao5@illinois.edu
†These authors contributed equally to this work

This PDF file includes:

Materials and Methods

Contrastive losses definition in CLEAN

Two contrastive losses were employed for training: Triplet Margin Loss (*1*) 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 Marigin 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 (*2*), except that Supcon-Hard Loss fixed a 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 (*3*), 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 calling 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 calling methods. Simply calling the top-1, top-2 or top-k ranked EC numbers for all query enzymes would not be suitable because calling the top-1 EC numbers would neglect enzyme promiscuity, but calling 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-calling 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 to call of the set of EC numbers with distances close to query enzyme but far from all other EC numbers mimicking human intuition. 10/8/22 11:20:00 AM10/8/22 11:20:00 AM

***p*-value calling method**

For the *p*-value algorithm, n (e.g., n = 20,000) randomly chosen enzyme embeddings (after a forward passing of the trained model) 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 calling method

For the Max-separation algorithm, we assumed there is a background noise distance γ, such that all distances $s_{i'}$ from the incorrect set of EC numbers for the query sequence are close to γ by ε, and all distances s_i from the correct set EC_i for the query sequence are 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 sequences 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 calling 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 (*4*). 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.

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) and similarity (SIM) were calculated by SIAS with BLOSUM62 matrix. Computing the percentage of identity and similarity requires dividing identities/similarities by the length of sequence with the following equation:

$$ID|SIM_{\%} = 100 \times \frac{Identical\ |\ Similar\ Residues}{Sequence\ Length}.$$

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 by Twist Bioscience (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 was a gift from Bradley Moore (Addgene plasmid # 136422) (*5*). 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 (*6*, *7*). In detail, 5 µL of the glycerol stock strains were inoculated into 5 mL LB medium supplemented with 50 µg/mL kanamycin and pre-cultivated at 37 °C (250 rpm) overnight. After that, 5 mL of pre-culture was transferred into 1 L Erlenmeyer flask containing 150 mL TB medium supplemented with 50 µg/mL kanamycin. 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 in sodium phosphate buffer (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. S2**).

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. 0.5 mM purified enzymes MJ1651, TTHA0338, and PH04363 as positive control for S-adenosyl-L-methionine (SAM) hydrolase was 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 and mixed with 50 µL ice cold 0.2% 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). 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).

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 that can generate semantically rich representations of protein sequences, encoding their evolutionary, structural, and biophysical properties (*8*). 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. For training and testing data split, we used MMSeqs2 (*9*) 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. Notably, the clustering split was challenging because the sequence similarity between the training and testing set was decreased. The split represented the case where query enzymes had low similarities with currently annotated enzymes.

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 (*10*).

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 loss. During training, SupCon-Hard samples several negative sequences where triplet loss only samples one. SupCon-Hard was observed to have superior performance on understudied of the test queries (**Fig. S2**). Besides SupCon-Hard sampled more negatives per batch during training, it also had an intrinsic ability to balance positives/negatives (*11*). While for both SupCon-Hard 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 SupCon-Hard. SupCon-Hard weights less for the negatives far away from the anchor and weights more for those closer to the anchor. Similarly, SupCon-Hard weights more for positives near the anchor and less for those farther away. This property contributed to the SupCon-Hard'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 SupCon-Hard 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. CLEAN outperforms ESM-1b in classification

To visualize the learned embedding on the low dimension, we used t-SNE (*12*) to reduce the high dimension embedding to a two-dimension plot **(Fig. S7).** 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 (*13*). 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 EC number 3.13.1.8 (**Fig. 3c**), which was subsequently confirmed by HPLC and MS studies (**Fig. 3h** and **3n**).

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.

All three enzymes MJ1641, TTHA0338 and SsFlA were members of the DUF (domain of unknown function) 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.

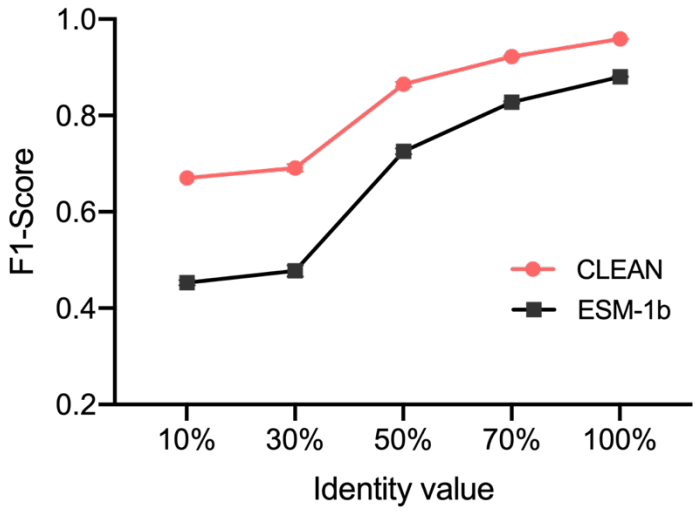

Fig. S1. The evaluation results of CLEAN on different identity clustering split under 5-fold CV (cross validated). ESM-1b was also investigated as comparison.

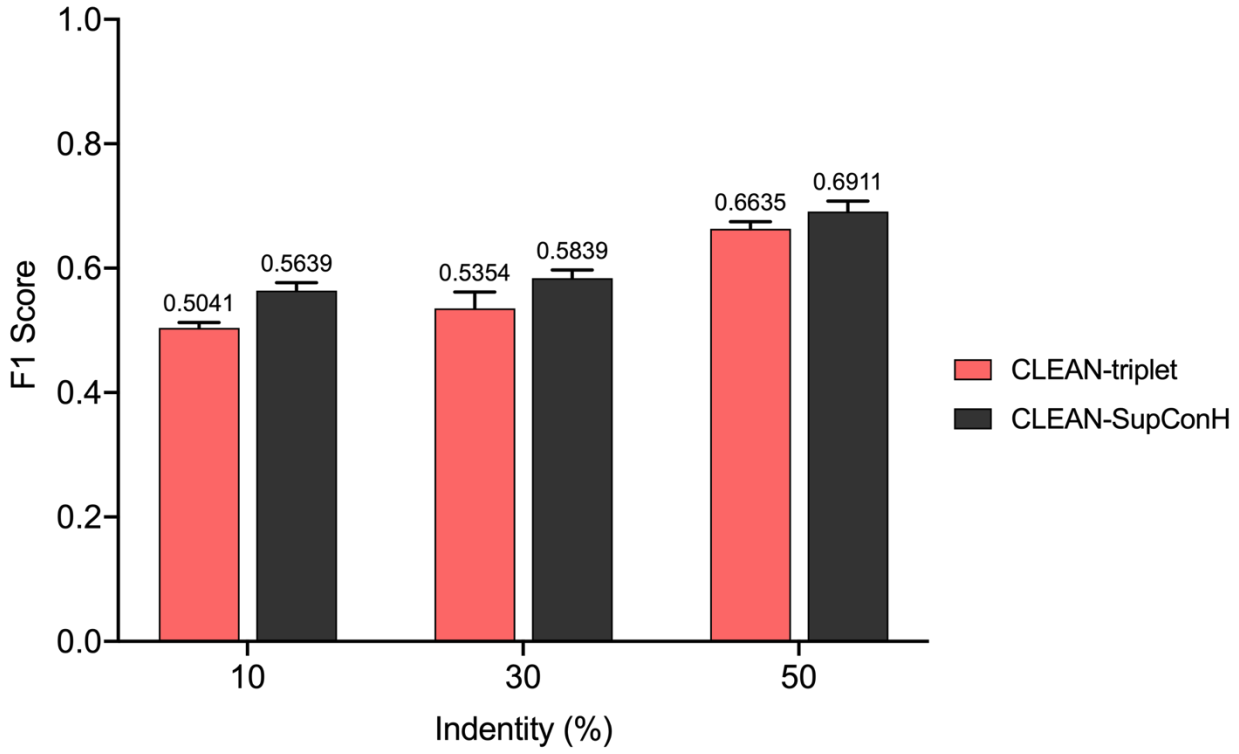

Fig. S2. Analysis of SupConH loss and triplet towards CLEAN. The validation set consists of EC numbers with no more than 5 occurrences in the training dataset. SupConH loss can further improve CLEAN's performance on understudied enzymes.

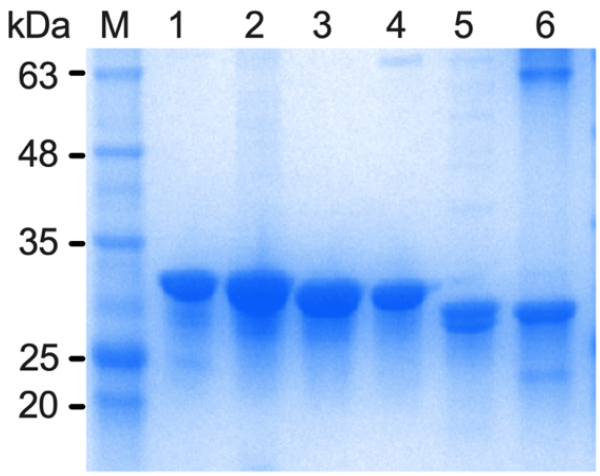

Fig. S3. SDS-PAGE analysis of the purified uncharacterized proteins and positive controls. Purified proteins were loaded onto 12% Mini-PROTEAN TGX™ Precast Gel (BIO-RAD). Lane M is a protein molecular weight marker. Lanes 1-5 are SalL (MW: 32.4 KDa), ScFlA (MW: 33.4 KDa), SsFlA (MW: 33.2 KDa), MJ1651 (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.

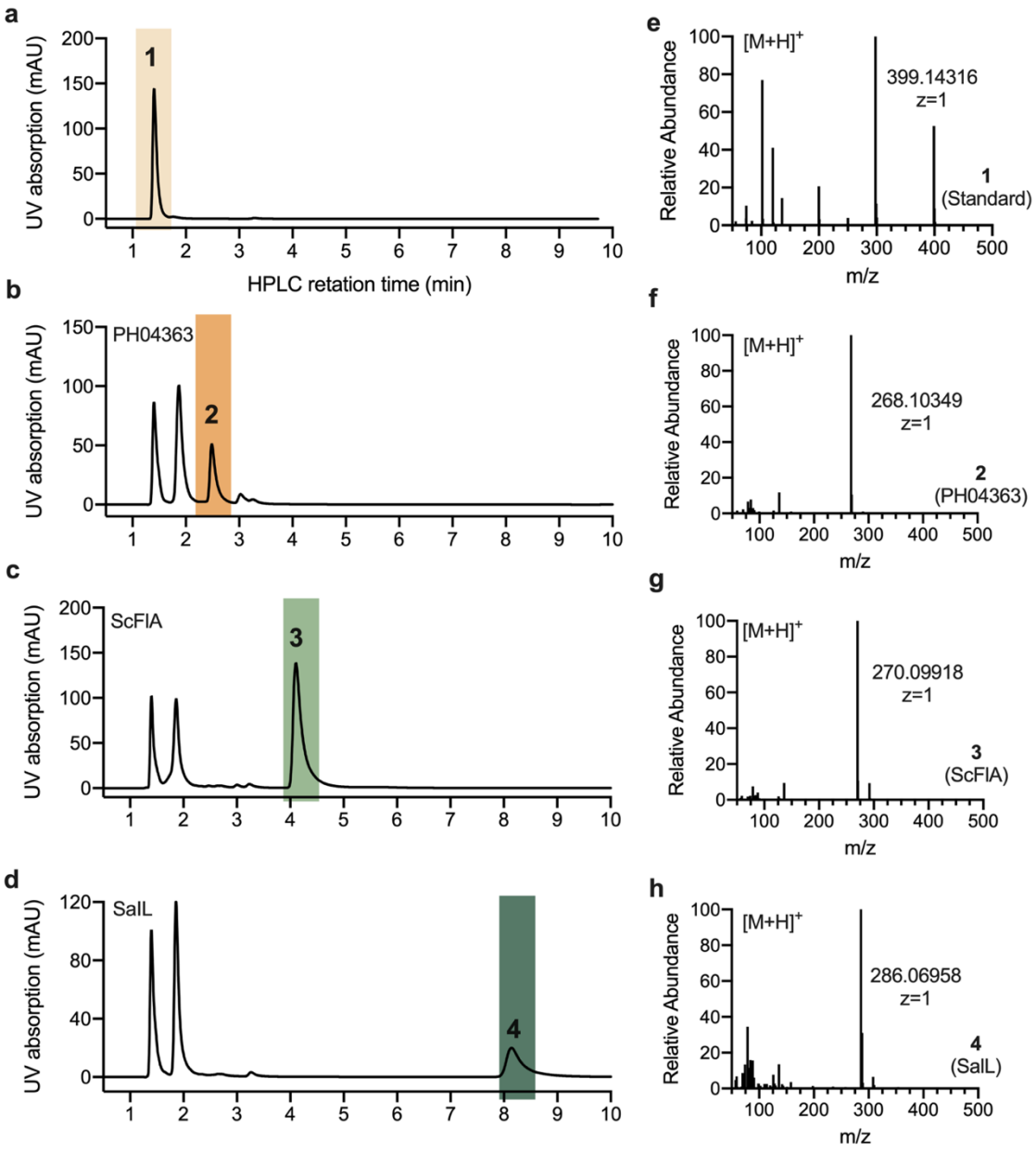

Fig. S456. Experimental validation of the CLEAN predicted EC numbers of uncharacterized halogenases. Left panel: HPLC analysis of (**a**) pure SAM (**1**) and reaction mixture of SAM with purified (**b**) PH04363+H_2O, (**c**) ScFlA+NaF, (**d**) SalL+NaCl at 30 °C for 24h. Detection was performed with 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 reaction.

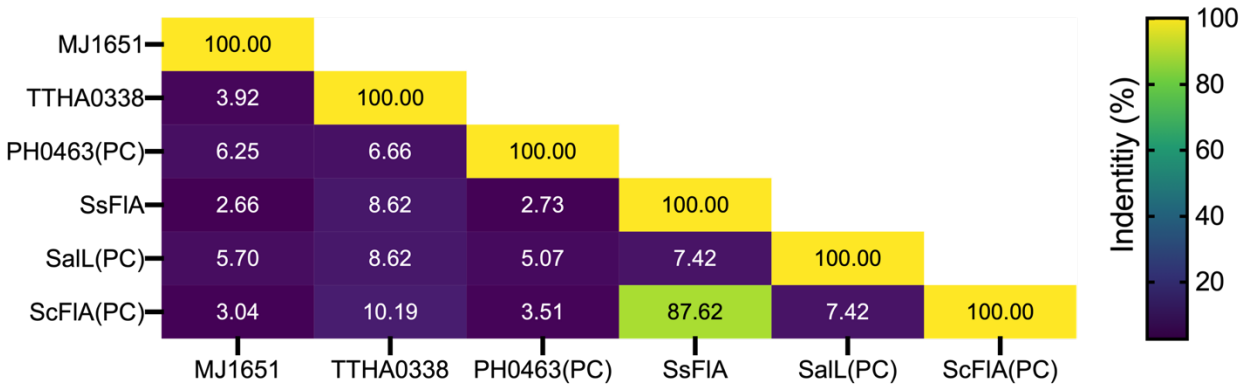

Fig. S5789. Heatmap of sequence identities among the uncharacterized proteins and positive control (PC) enzymes. Color bar with "viridis" color scale was in percentage (%).

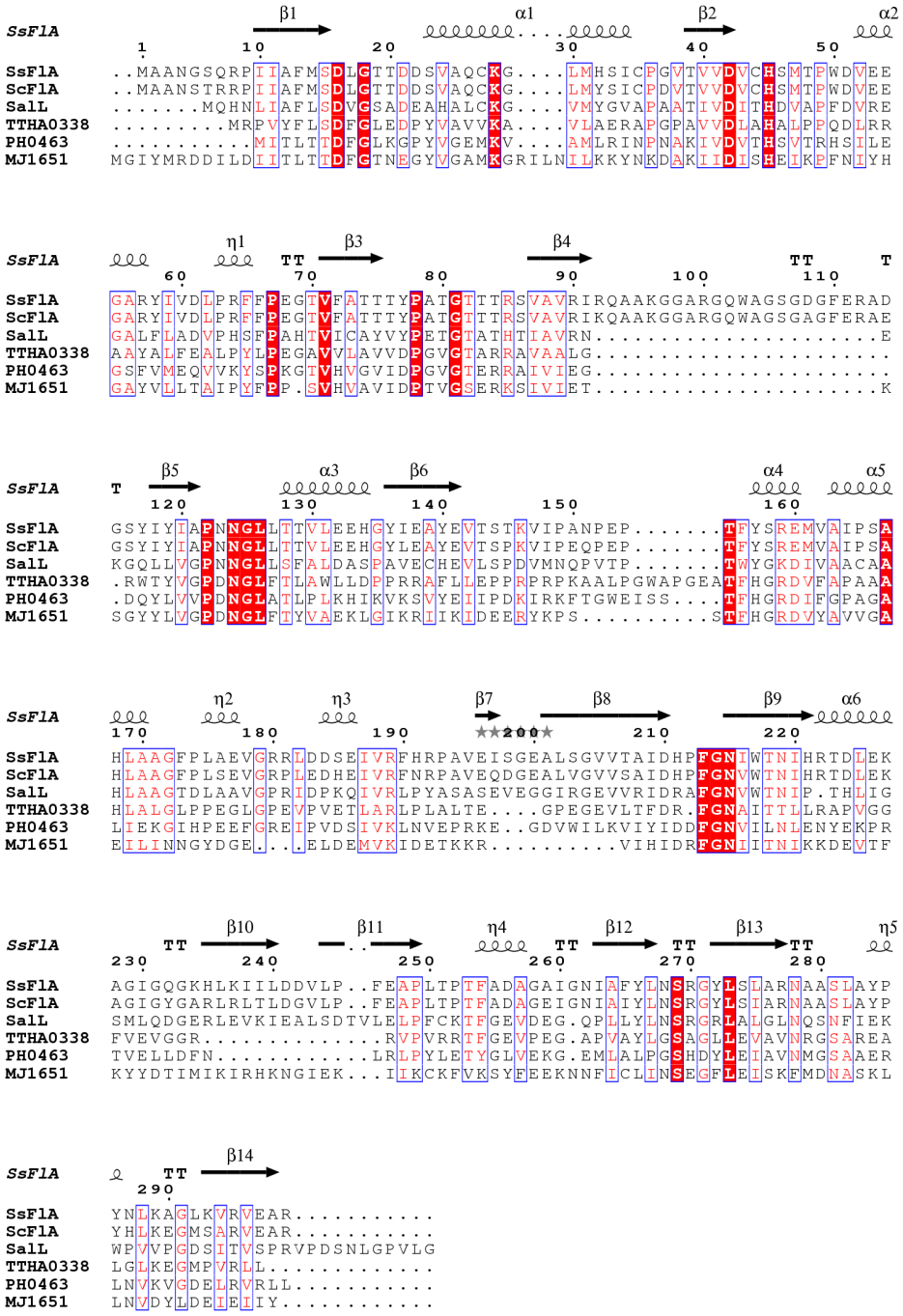

Fig. S6101112. Sequence alignment of TTHA0338, MJ1651, and SsFlA with positive control enzymes. The amino acid sequences of six SAM-dependent enzymes are aligned by ClustalW, with the secondary structural elements of SsFlA (PDB ID: 5B6I) indicated above the sequences. The highly conserved residues are labeled with red color.

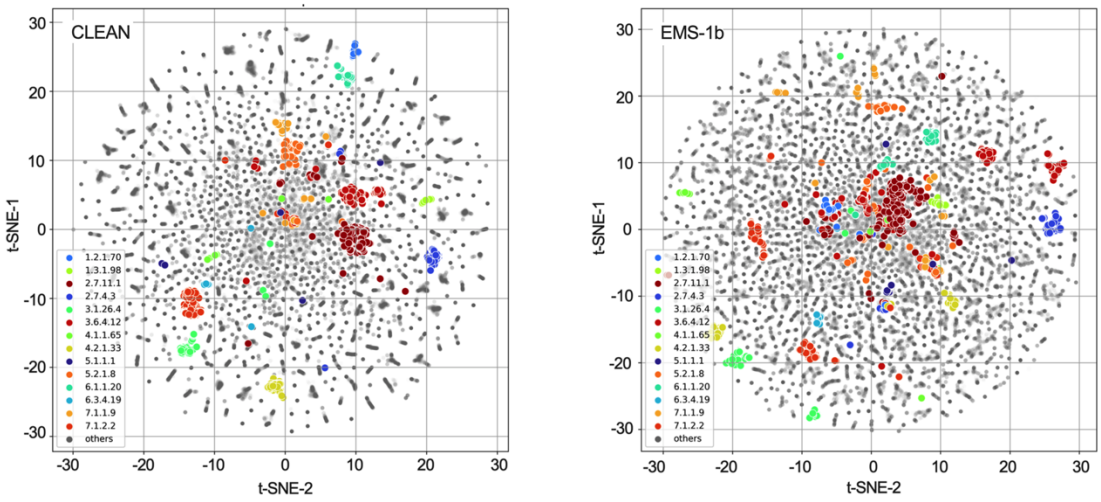

Fig. S7131415. The two-dimensional visualization of CLEAN's embedding compared with ESM-1b embedding using t-SNE. Each dot in the plot represented a single enzyme and each color represented an EC number. Several randomly selected EC numbers (~14 types) are highlighted for visualization purposes.

Table S1. Max-Separation calling method algorithm

Step	Description
1	Function MAXSEP(S)[a]
2	Let background noise distance $\gamma =$ 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 s_0, s_1, ..., s_{n-1} in sorted order.

Table S23.4. 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 (*14*)
	MarH1	A0A0F7N9T7	1.11.1.10	n.p.	unsure[f]	1.-.-.-	2.7.7.0	3.-.-.-	1.11.1.18	1.11.1.10 (*15*)
	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 (*15*)
	MarH3	A0A559V0T8	1.11.1.10	n.p.	n.p.	2.4.1.-	n.p.	4.-.-.-	1.11.1.10	1.11.1.10 (*16*)
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 (*16*)
	StaI	Q8KLM0	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*17*)
	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 (*18*)
	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.- (*19*)
	PlBmp2	A0A162BNF2	1.14.19.56	n.p.	unsure	1.-.-.-	1.3.1.101	1.3.-.-	1.3.1.-	1.14.19.- (*20*)
	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 (*21*)
	CazI	Q2GWL3	1.14.13.209	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.23	1.14.19.- (*22*)
	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.- (*23*)
	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.- (*24*)
	McnD	A4Z4I8	1.14.19.56	n.p.	unsure	1.-.-.-	1.3.7.-	1.-.-.-	1.14.13.7	1.14.19.- (*25*)
	Chal	P96557	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*26*)
	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.- (*27*)
	Veg13	B7T185	1.14.19.56	n.p.	unsure	1.14.14.-	1.3.7.11	1.-.-.-	1.14.13.7	1.14.19.- (*27*)
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.- (*28*)
	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.- (*28*)
	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.- (*29*)
	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.- (*30*)
	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.- (*31*)
	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.- (*32*)
	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.- (*33*)
	SlBesD	UPI0004CAEF9E	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	4.-.-.-	1.14.11.19	1.14.11.- (*34*)
	SwHalB	UPI000499837C	1.14.11.74	n.p.	unsure	n.p.	2.7.1.-	3.-.-.-	1.14.20.1	1.14.11.- (*34*)
	SiHalB	UPI000888A5A4	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	6.-.-.-	1.14.11.16	1.14.11.- (*34*)
	LaHalC	UPI0003675DF0	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	n.p.	1.21.3.1	1.14.11.- (*34*)
	PkHalD	UPI0005FAE1F9	1.14.11.46	n.p.	unsure	n.p.	2.3.2.-	n.p.	1.14.11.18	1.14.11.- (*34*)
	PsRoot562 HalE	UPI000702B50B	1.14.11.74	n.p.	unsure	n.p.	2.7.7.-	3.-.-.-	1.14.20.1	1.14.11.- (*34*)
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 (*35*)
	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 (*36*)
	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 (*37*)
	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.

References

1. V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks" in *Procedings of the British Machine Vision Conference 2016* (British Machine Vision Association, York, UK, 2016; http://www.bmva.org/bmvc/2016/papers/paper119/index.html), p. 119.1-119.11.

2. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, "Supervised Contrastive Learning" (arXiv:2004.11362, arXiv, 2021), doi:10.48550/arXiv.2004.11362.

3. K. Sohn, "Improved Deep Metric Learning with Multi-class N-pair Loss Objective" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016; https://papers.nips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html), vol. 29.

4. 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).

5. 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).

6. 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).

7. H. Cui, L. Eltoukhy, L. Zhang, U. Markel, K. 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).

8. 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 (2020), p. 622803.

9. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

10. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, ProteInfer: deep networks for protein functional inference (2021), p. 2021.09.20.461077, , doi:10.1101/2021.09.20.461077.

11. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, "Supervised contrastive learning" (arXiv:2004.11362, arXiv, 2021), , doi:10.48550/arXiv.2004.11362.

12. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** (2008).

13. K. N. Rao, S. K. Burley, S. Swaminathan, Crystal structure of a conserved protein of unknown function (MJ1651) from Methanococcus jannaschii. *Proteins Struct. Funct. Bioinforma.* **70**, 572–577 (2008).

14. 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).

15. L. A. Murray, S. M. 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).

16. J. R. Heemstra Jr, C. T. Walsh, Tandem action of the O2-and FADH2-dependent halogenases KtzQ and KtzR produce 6, 7-dichlorotryptophan for kutzneride assembly. *J. Am. Chem. Soc.* **130**, 14024–14025 (2008).

17. 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).

18. 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).

19. 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).

20. 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).

21. 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**, 1992–9 (2008).

22. 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).

23. 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).

24. B. R. K. Menon, D. Richmond, N. Menon, Halogenases for biosynthetic pathway engineering: Toward new routes to naturals and non-naturals. *Catal. Rev.*, 1–59 (2020).

25. 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).

26. 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).

27. 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 A*. **105**, 17273–7 (2008).

28. 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).

29. 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**, 4872–9 (2009).

30. 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** (2018), doi:10.3389/fchem.2018.00513.

31. 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 Engl*. **55**, 5780–4 (2016).

32. 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 Engl*. **59**, 9478–9484 (2020).

33. 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).

34. 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).

35. 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).

36. M. HimáTong, Fluoroacetate biosynthesis from the marine-derived bacterium Streptomyces xinghaiensis NRRL B-24674. *Org. Biomol. Chem.* **12**, 4828–4831 (2014).

37. 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).