FUNCTION PREDICTION

Enzyme function prediction using contrastive learning

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

Enzyme function annotation is a fundamental challenge, 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 proteins or those with previously uncharacterized functions or multiple activities. 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 compared with the state-of-the-art tool BLASTp. The contrastive learning framework empowers CLEAN to confidently (i) annotate understudied enzymes, (ii) correct mislabeled enzymes, and (iii) identify promiscuous enzymes with two or more EC numbers—functions that we demonstrate by systematic in silico and in vitro experiments. We anticipate that this tool will be widely used for predicting the functions of uncharacterized enzymes, thereby advancing many fields, such as genomics, synthetic biology, and biocatalysis.

The development of DNA sequencing technologies, and particularly genomics and metagenomics tools, has led to the discovery of numerous protein sequences from organisms across all branches of life. For example, UniProt Knowledgebase has cataloged ~190 million protein sequences. However, only <0.3% (approximately half a million) of these proteins were reviewed by human curators, out of which <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 ~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) number is the most well-known numerical classification scheme of enzymes, which specifies the catalytic function of an enzyme by four digits. Because 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. On the other hand, almost all the existing ML models, such as DeepEC (*14*) and ProteInfer (*15*), are based on a multilabel 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 to understand the range of protein functions.

In this work, we report a 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 extensive in silico experiments. Furthermore, we challenged CLEAN to annotate EC numbers for an in-house collected database of all uncharacterized halogenases (*36* in total) followed by case studies as in vitro experimental validation. CLEAN outperformed other EC number annotation tools at these tasks, including BLASTp and state-of-the-art ML models.

Model development and evaluation

Unlike previously developed ML algorithms that frame EC number prediction tasks as a multilabel classification problem, CLEAN used a contrastive learning (*16*, *17*) framework. Our training objective is to learn an embedding space of enzymes where the Euclidean distance reflects the functional similarities. The embedding refers to a numerical representation (vectors or matrices) of protein sequence that is readable by machine while still retaining the important features and information carried by the enzyme. In CLEAN's task, the amino acid sequences with the same EC number have a small Euclidean distance, whereas sequences with different EC numbers have a large distance. Contrastive losses were used to train the model with supervision (*16*, *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 with embeddings that 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, section 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 selects EC numbers that have the maximum separation (stand out) from other EC numbers in terms of the pairwise distance to the query sequence and (ii) a *P* value–based method that identifies EC numbers with statistical significance compared with background (see materials and methods). On a train-test split in which none of the enzymes in the test set share >50% identity with any enzymes in the training set, using the maximum-separation selection method, CLEAN achieved a 0.865 F1 score—a commonly used accuracy metric indicating the harmonic mean of precision and recall. Even at 10% sequence identity clustering, CLEAN reached a 0.67 F1 score. Additionally, CLEAN achieved much higher performance compared with the baseline method using ESM-1b without contrastive learning (fig. S1).

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

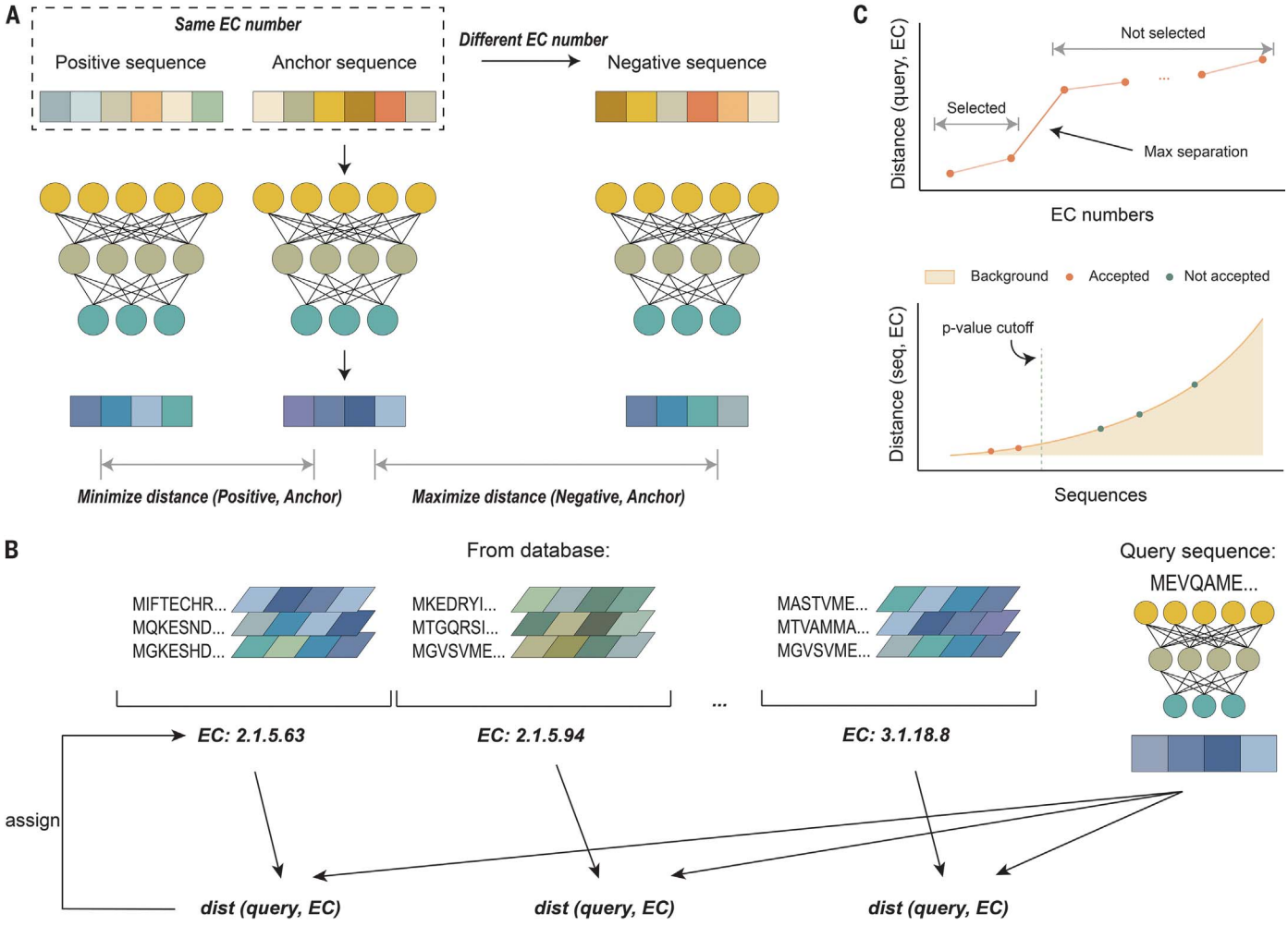

Fig. 1. The contrastive learning–based framework of CLEAN. (**A**) During training, positives and negatives were sampled on the basis of EC numbers. The input sequences were embedded and passed through a neural network. The series of squares with warm colors stands for the representation of input sequence embedded by ESM-1b. Similarly, the sequence embeddings obtained from the supervised contrastive learning neural network are illustrated by cool colors. (**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 (shown as a parallelogram with cool colors) 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, maximum separation (above) and *P* value (below), were implemented to prioritize confident predictions of EC numbers from the ranking order.

Benchmarking CLEAN with previous EC number annotation tools

After training, the prediction performance of CLEAN was systematically investigated by comparing it with six state-of-the-art EC number annotation tools [i.e., ProteInfer (*15*), DeepEC (*14*), BLASTp, DEEPre (*20*), CatFam (*21*), and ECPred (*22*)]. 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 (April 2022). 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 multilabel accuracy metrics, including precision (0.597) and recall (0.481), when compared with ProteInfer and DeepEC (Fig. 2A). Also, CLEAN achieved an F1 score of 0.499, whereas ProteInfer and DeepEC had scores of 0.309 and 0.230, respectively.

The second independent dataset, denoted as Price-149, was a set of experimentally validated results described by Price *et al.* (*23*). The Price-149 dataset was first curated by ProteInfer (*15*) as a challenging dataset because 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 (0.495) compared with BLASTp, ProteInfer, and DeepEC (Fig. 2B). Notably, in this challenging task, CLEAN had a 3.0-fold higher F1 score than ProteInfer (0.166) and an almost 5.8-fold higher score than DeepEC (0.085). The evaluations on the New-392 and Price-149 datasets demonstrate that CLEAN is more precise and reliable than previously developed ML-based models for predicting functions for newly discovered proteins, especially the ones without known enzyme functions.

Understanding CLEAN's performance on annotating understudied EC number

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 test our hypothesis that, compared with the multilabel

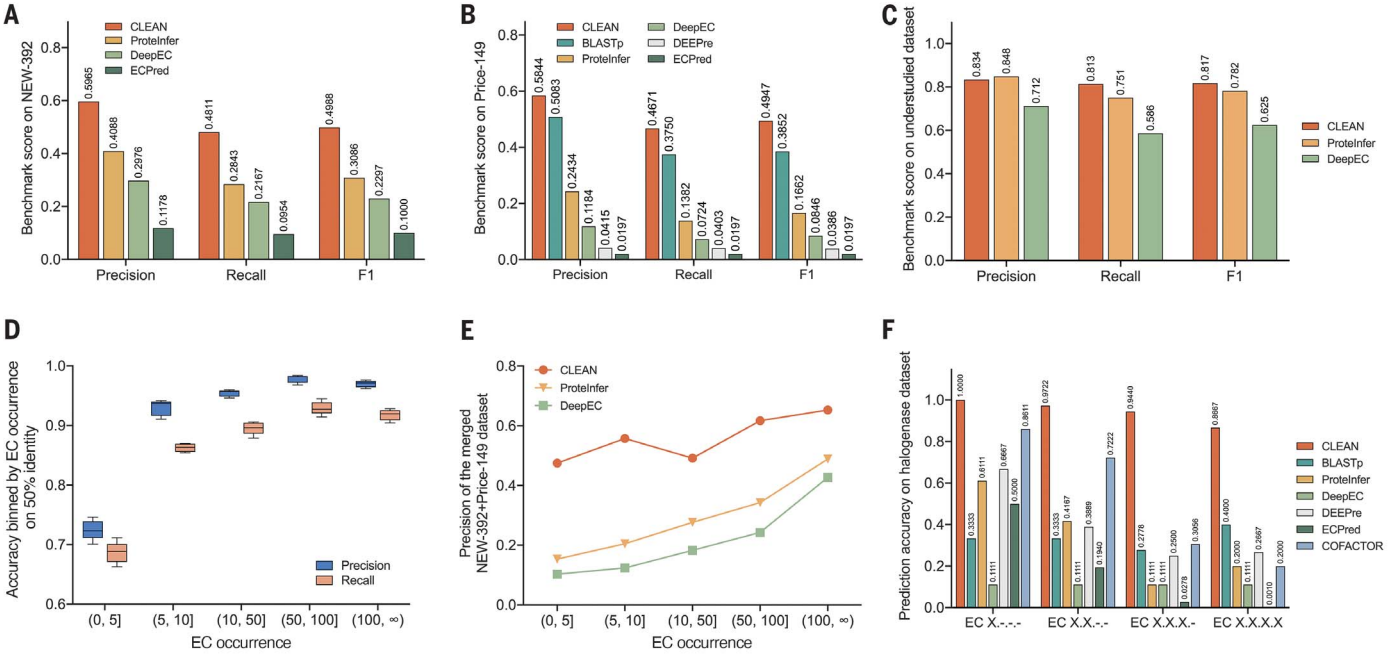

Fig. 2. Quantitative comparison of CLEAN with the state-of-the-art EC number prediction tools. (**A**) Evaluation of CLEAN's performance toward three multilabel accuracy metrics (precision, recall, and F1 score) examined on the New-392 database. Four top-ranked models, ProteInfer, DeepEC, CatFam, and ECPred, were used for comparison. (**B**) Comparison of CLEAN, BLASTp, ProteInfer, DeepEC, DEEPre, CatFam, and ECPred 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 using the test set with <50% identity to the training set evaluated with SupconH loss. Precision and recall values were binned by the number of times that the EC number appeared in the training set—i.e., the bin (0,5] means that the EC numbers occurs less than five times in the training set. The box plots show the results of fivefold cross-validation. (**E**) Evaluation on the combined datasets of Price-149 and New-392 binned by the number of times that the EC number appeared in CLEAN's training dataset. (**F**) Prediction accuracy of CLEAN on an in-house–curated halogenase dataset compared with six commonly used tools (BLASTp, ProteInfer, DeepEC, DEEPre, ECPred, and COFACTOR). This dataset had good diversity covering 11 different EC numbers.

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 five). In this validation dataset, each type of EC number had no more than five occurrences, and more than 3000 samples were included in this dataset covering more than 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. In other words, both ProteInfer and DeepEC had an advantage that both models have seen the validation dataset used in Fig. 2C during training, resulting in the acceptable 0.625 to 0.782 F1 score. Despite this added advantage, CLEAN outperformed both methods, achieving a 0.817 F1 score (Fig. 2C).

We analyzed CLEAN's performance based on the number of times that the EC number occurred in the training set. Even at 50% sequence identity clustering, where the test set and train set had a low similarity, CLEAN's performance did not drop considerably 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 that EC numbers appeared in the training set. As expected, ProteInfer and DeepEC showed a bias toward popular EC numbers, limited by the classification framework. By 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. To further analyze the hypothesis that CLEAN can leverage not only positive examples but also negative examples through contrastive learning, Supcon-Hard loss (SupconH)—a loss function that samples more negatives compared with triplet loss—was implemented (materials and methods; supplementary text, section 2; and fig. S2).

Moreover, we implemented a method to quantify the prediction result confidence. We fitted a two-component Gaussian mixture model (GMM) on the distribution of the Euclidean distances between enzyme sequence embeddings and EC number embeddings (materials and methods). Knowing the prediction confidence, researchers can make quantitative interpretations of CLEAN's prediction. The confidence quantification can also help CLEAN to avoid overprediction by reporting the third level of EC number when the confidence is low (figs. S11 to S14 and supplementary text, section 3).

Experimental validation

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 (*23*, *24*, *25*). Generally, small molecules with halogen atoms 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 have been identified from UniProt, covering all four types of halogenases [haloperoxidase, flavin-dependent, α-ketoglutarate (α-KG)–dependent, and *S*-adenosyl-L-methionine (SAM)–dependent

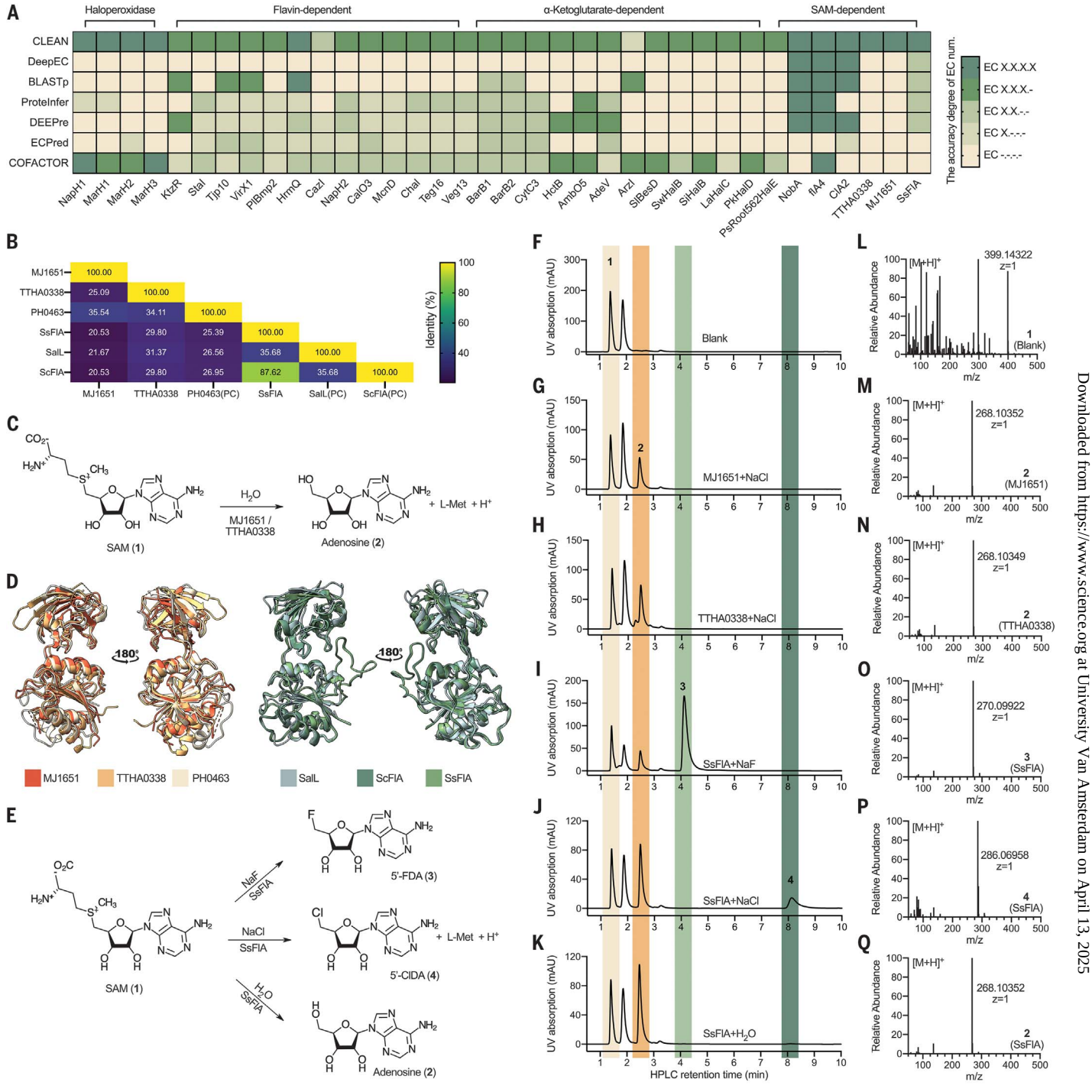

Fig. 3. Experimental validation of CLEAN on uncharacterized halogenases. (**A**) The accuracy degree heatmap of EC numerical ID was shown for the 36 identified halogenases. (**B**) Heatmap of sequence identity among the uncharacterized proteins and positive control (PC) enzymes. The color bar with the "viridis" color scale indicates percentage. (**C**) The SAM hydroxide adenosyltransferase MJ1651-TTHA0338 reaction. (**D**) Structural superposition of the three-dimensional (3D) structures of uncharacterized proteins MJ1651 [Protein Data Bank (PDB) ID: 2F4N (*28*)], TTHA0338 [PDB ID: 2CW5 (*39*)], and positive control enzyme PH0463 [PDB ID: 1WU8 (*40*)]. The same structural superposition was performed for SsFlA [PDB ID: 5B6I (*30*)], SalL [PDB ID: 2Q6O (*41*)], and ScFlA [PDB ID: 1RQR (*42*)]. The superposition shows that the 3D structures of these SAM-binding enzymes are very similar; yet, CLEAN can accurately distinguish their functions. Chain A in each crystal structure was used for structural superposition. (**E**) Nucleophilic substitution of SAM with halide ions or H_2O toward SsFlA. (**F** to **K**) HPLC analysis of reaction mixtures of SAM and NaCl/NaF/H_2O with blank (F), purified MJ1651 (G), purified TTHA0338 (H), and purified SsFlA [(I) to (K)]. The peaks of substrate SAM (**1**), product adenosine (**2**), 5′-fluoro-5′-deoxyadenosine (5′-FDA) (**3**), and 5′-chloro-5′-deoxyadenosine (5′-CIDA) (**4**) were labeled with light yellow, orange, green, and dark green, respectively, which were also aligned at the same retention time. UV, ultraviolet; mAU, milli–absorbance unit. (**L** to **Q**) Mass spectra of compounds obtained from the reaction mixtures: substrate **1** in the blank reaction system (L), adenosine (**2**) in MJ1651 catalyzed reaction (M), adenosine (**2**) in TTHA0338 catalyzed reaction (N), 5′-FDA (**3**) (O), 5′-CIDA (**4**) (P), and adenosine (**2**) (Q). m/z, mass/charge ratio.

halogenase] (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. The halogenase dataset is particularly challenging because the halogenase family is understudied, and only a limited number of halogenases are available in the database. With expert curation and experimental validations showing later, all 36 halogenases were confidentially annotated with EC numbers. Overall, CLEAN achieved much better prediction accuracy (86.7 to 100%; Fig. 2F and Fig. 3A) compared with the six other commonly used computational tools (e.g., ~11.1% in DeepEC and 11.1 to 61.1% in ProteInfer). The latter range corresponds to the prediction accuracy at different digits of EC number (from digit 1 to digit 4). These results demonstrate that CLEAN can distinguish enzyme functions even within the regime of similar biocatalytic reactions.

Among these 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. CLEAN predicted new EC numbers in these three cases, suggesting that other potential functions might occur. Therefore, we performed in vitro experiments to validate these predictions. High-performance liquid chromatography–mass spectrometry (HPLC-MS) analysis coupled with enzyme kinetic analysis confirmed that MJ1651 is SAM hydrolase (EC 3.13.1.8), as CLEAN predicted, rather than chlorinase (EC 2.5.1.94) or fluorinase (EC 2.5.1.63), as mislabeled in UniProt and by the selected computational tools used in this work (Fig. 3, C, D, F, G, and M; fig. S3; fig. S4, A and B; fig. S5A; fig. S7; and table S3). CLEAN also correctly annotated TTHA0338, which belongs to the DUF62 Pfam family with no known function, as a SAM hydrolase (Fig. 3, C, D, H, and N; figs. S5B and S7; and table S3). With the exception of BLASTp successfully predicting the target TTHA0338, all other six commonly used computational tools failed to predict MJ1641 and TTHA0338. These results revealed that CLEAN is favorable for correcting mislabeled enzymes and accurately identifying understudied catalytic functions. CLEAN also confidently identified the promiscuous enzyme SsFlA with three EC numbers (EC 2.5.1.63, EC 2.5.1.94, and EC 3.13.1.8; Fig. 3, E, I to K, and O to Q). These observations confirmed that CLEAN could effectively recall defined biological activity and capture elements of enzyme promiscuity. The precision of CLEAN is impressive in distinguishing SAM-binding proteins with homologous structures (fig. S3C) and sequence identity ranging from 20.5 to 35.7% for everything but SsFlA versus ScFlA, which is 87.6% (Fig. 3B and fig. S6). Functions of proteins with sequence identities in this range are often challenging to predict. These results suggest that our sequence-based model CLEAN performed better than structure-based methods [e.g., COFACTOR (*12*, *13*)] in dealing with enzymes with similar structures but different functions.

Discussion

Through systematic in silico and in vitro experimental validations, we have demonstrated that CLEAN achieves superior prediction performance relative to six state-of-the-art tools (i.e., ProteInfer, BLASTp, DeepEC, DEEPre, COFACTOR, and ECPred). A comprehensive analysis on an uncharacterized halogenase dataset indicated that CLEAN can characterize the hypothetical proteins and correct 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 or biased and scarce.

We believe that CLEAN will be a powerful tool for predicting the catalytic function of query enzymes, which can greatly facilitate studies in functional genomics (*32*), enzymology, enzyme engineering (*33*), synthetic biology (*34*), metabolic engineering (*35*, *36*), and retrobiosynthesis (*37*, *38*). Moreover, the general language model representation topped with the contrastive learning workflow used by CLEAN can readily be 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 as a software component integrated into other computational platforms. The superior performance of CLEAN in predicting understudied proteins should greatly expand the bioinformatics toolbox, thereby laying the cornerstone for future detailed mechanistic studies.

REFERENCES AND NOTES

1. UniProt Consortium, *Nucleic Acids Res.* **49**, D480–D489 (2021).
2. P. Radivojac *et al.*, *Nat. Methods* **10**, 221–227 (2013).
3. K. Hult, P. Berglund, *Trends Biotechnol.* **25**, 231–238 (2007).
4. C. J. Jeffery, *Phil. Trans. R. Soc. B* **373**, 20160523 (2018).
5. E. W. Sayers *et al.*, *Nucleic Acids Res.* **50**, D20–D26 (2022).
6. M. Blum *et al.*, *Nucleic Acids Res.* **49**, D344–D354 (2021).
7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403–410 (1990).
8. D. K. Desai, S. Nandi, P. K. Srivastava, A. M. Lynn, *Adv. Bioinformatics* **2011**, 743782 (2011).
9. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389–3402 (1997).
10. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, *J. Mol. Biol.* **235**, 1501–1531 (1994).
11. M. Steinegger *et al.*, *BMC Bioinformatics* **20**, 473 (2019).
12. C. Zhang, P. L. Freddolino, Y. Zhang, *Nucleic Acids Res.* **45**, W291–W299 (2017).
13. A. Roy, J. Yang, Y. Zhang, *Nucleic Acids Res.* **40**, W471–W477 (2012).
14. J. Y. Ryu, H. U. Kim, S. Y. Lee, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 13996–14001 (2019).
15. T. Sanderson, M. L. Bileschi, D. Belanger, L. J. Colwell, *eLife* **12**, e80942 (2023).
16. P. Khosla *et al.*, 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 *et al.*, *NAR Genom. Bioinform.* **4**, lqac043 (2022).
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 *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
20. Y. Li *et al.*, *Bioinformatics* **34**, 760–769 (2018).
21. C. Yu, N. Zavaljevski, V. Desai, J. Reifman, *Proteins* **74**, 449–460 (2009).
22. A. Dalkiran *et al.*, *BMC Bioinformatics* **19**, 334 (2018).
23. M. N. Price *et al.*, *Nature* **557**, 503–509 (2018).
24. C. Crowe *et al.*, *Chem. Soc. Rev.* **50**, 9443–9481 (2021).
25. K. Prakinee *et al.*, *Nat. Catal.* **5**, 534–544 (2022).
26. J. Latham, E. Brandenburger, S. A. Shepherd, B. R. K. Menon, J. Micklefield, *Chem. Rev.* **118**, 232–269 (2018).
27. V. Agarwal *et al.*, *Chem. Rev.* **117**, 5619–5674 (2017).
28. K. N. Rao, S. K. Burley, S. Swaminathan, *Proteins* **70**, 572–577 (2008).
29. A. S. Eustáquio, J. Härle, J. P. Noel, B. S. Moore, *ChemBioChem* **9**, 2215–2219 (2008).
30. H. Sun *et al.*, *Angew. Chem. Int. Ed.* **55**, 14277–14280 (2016).
31. H. Nam *et al.*, *Science* **337**, 1101–1104 (2012).
32. O. Shalem, N. E. Sanjana, F. Zhang, *Nat. Rev. Genet.* **16**, 299–311 (2015).
33. Y. Wang *et al.*, *Chem. Rev.* **121**, 12384–12444 (2021).
34. H. Zhao, *ACS Synth. Biol.* **11**, 3550 (2022).
35. X.-M. Sun, Y.-S. Xu, H. Huang, *Trends Biotechnol.* **39**, 648–650 (2021).
36. G. B. Kim, W. J. Kim, H. U. Kim, S. Y. Lee, *Curr. Opin. Biotechnol.* **64**, 1–9 (2020).
37. T. Yu *et al.*, *Nat. Catal.* **6**, 137–151 (2023).
38. D. Rother, S. Malzacher, *Nat. Catal.* **4**, 92–93 (2021).
39. S. Satoh *et al.*, 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, *Nat. Chem. Biol.* **4**, 69–74 (2008).
42. C. Dong *et al.*, *Nature* **427**, 561–565 (2004).
43. T. Yu *et al.*, Enzyme function prediction using contrastive learning, version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.7582241.

ACKNOWLEDGMENTS

We thank H. Ren and C. Huang for their suggestions on the characterization of the products formed by halogenases. We also thank for Z. Zhang, C. Huang, and Z. Pei for valuable discussion and guidance on experimental validation. **Funding:** This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by the US National Science Foundation (NSF) 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.C.L. designed the in silico experiments. T.Y. and J.C.L. contributed to data preparation and carried out the in silico experiments. J.C.L. contributed to cleaning up code. H.C. and H.Z. planned the in vitro experiments. H.C. and G.J. 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:** The authors declare that they have no competing interests. **Data and materials availability:** All data and code generated as part of this study are freely accessible in either the supplementary materials or open repositories. Code

for model development and validation are freely accessible at Zenodo (*43*) and GitHub (https://github.com/tttianhao/CLEAN). CLEAN is converted into an easy-to-use web server and made freely accessible at https://moleculemaker.org/alphasynthesis. The following datasets curated in previous publications and databases were used: Price-149 (*15*) and New-392 (*43*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf2465
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S3
References (*44*–*80*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 8 October 2022; accepted 6 March 2023
10.1126/science.adf2465