THRUST 3

CLEAN: Enzyme function prediction using contrastive learning

Tianhao Yu, Ocean Cui, Canal Li, Yunan Luo and Huimin Zhao

Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana IL, 61801

High-quality Functional Annotation with ML

The accessibility of protein sequences in protein databases is ever growing, yet only a small fraction is functionally annotated. BLASTp and HMMs are the most widely used bioinformatics tools to label sequences. However, 1/3 bacterial proteins still cannot be annotated[1]. Many recent studies applied machine learning ML for function prediction[2,3]. Classification model's performance *decreases* with the number of examples in the training sets, a challenge for under-studied functions[4]. Our work used **contrastive learning** framework to achieve highly accurate prediction on enzyme commission (EC) number, even for under-studied functions.

Contrastive learning framework

Contrastive learning does not learn the label of inputs directly, but instead it learns the differences between samples:

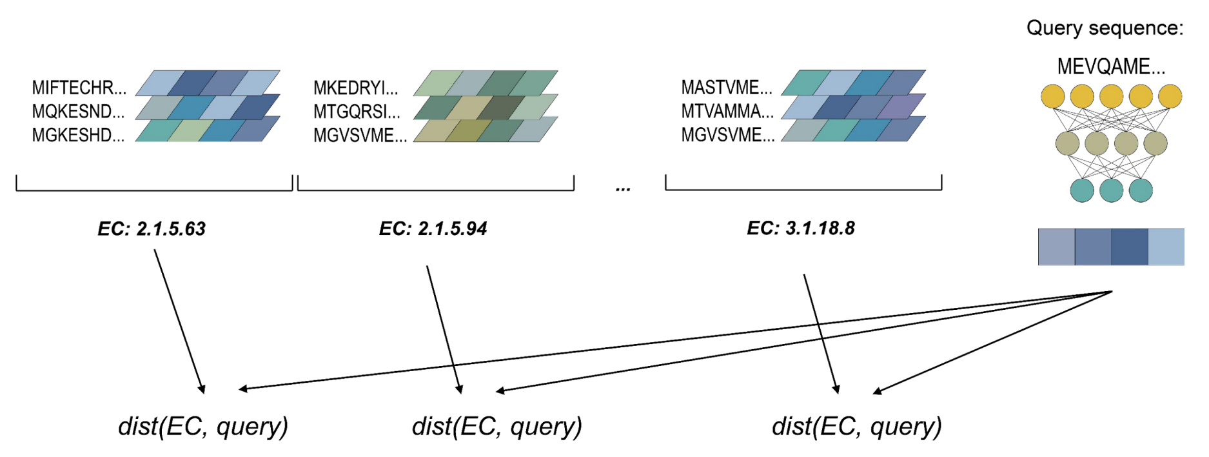
- ❑ **Minimize** the distance between sequences with the **same** function (EC)
- ❑ **Maximize** the distance between sequences with the **different** function

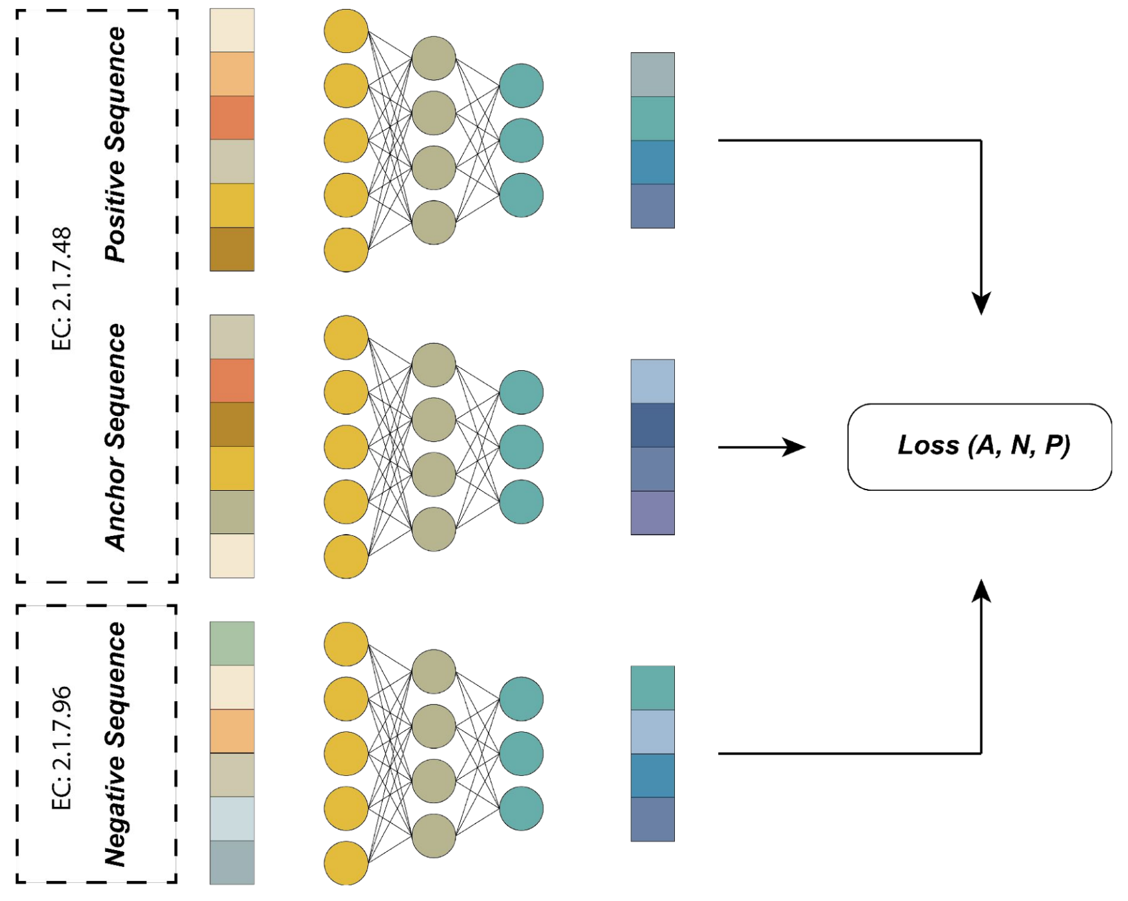
1) Triplet Margin Loss

$$\mathcal{L}^{TM} = ||z_a - z_p||_2 - ||z_a - z_n||_2 + \alpha$$

2) Supercon-Hard Loss:

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





Each EC number can be represented by **EC Cluster Center**, the average of embeddings with same EC.

We develop 2 **EC-calling** methods:

- ❖ ***p*-value**, picks random examples to rank query with threshold *p*;
- ❖ **Max-Separation**, finds maximum separation between distances

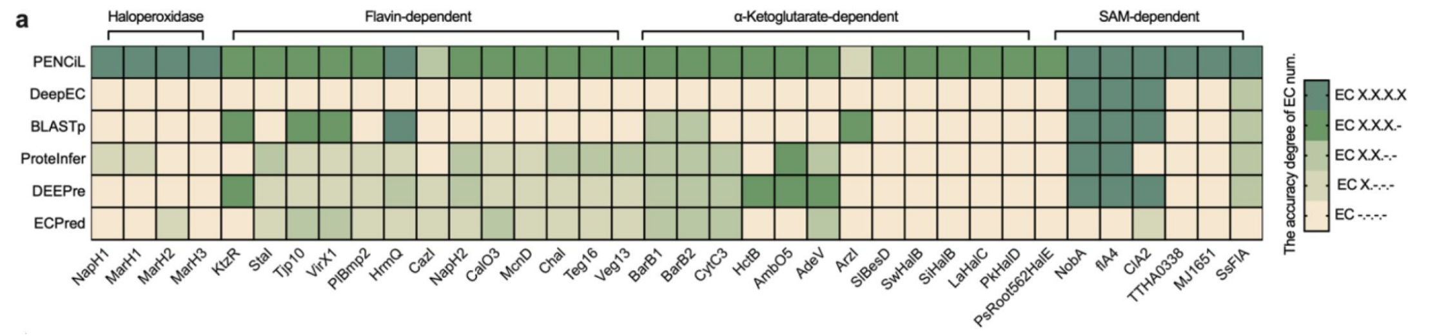
Algorithm 1 Max-Separation

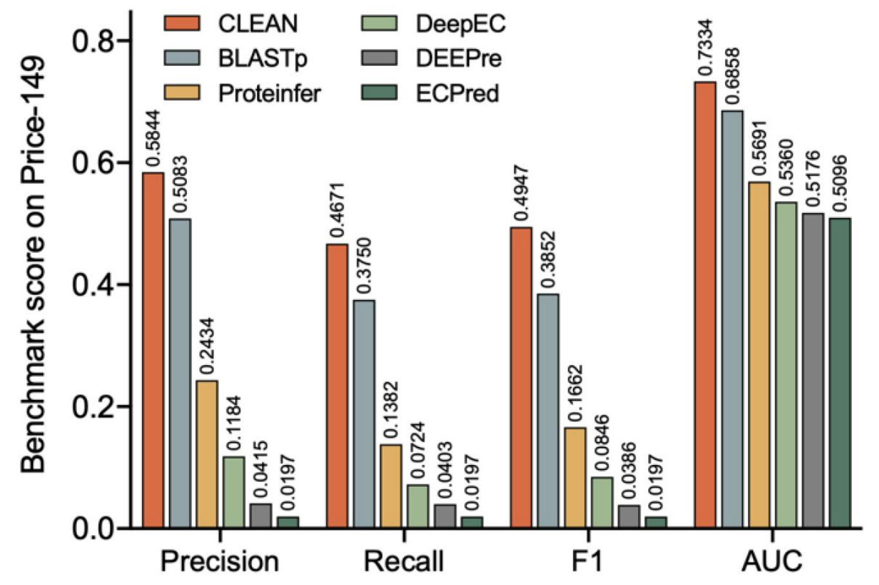
```
function MaxSep($S$)
    Require $S$ is the sequence of distances $s_0, s_1, ..., s_{n-1}$ in sorted order
    Let background noise distance $\hat{\gamma} = \text{mean}(s_1 + s_2 + ... + s_{n-1})$
    Let noise separation distances $D = d_0, ..., d_{n-1} = |s_0 - \hat{\gamma}|, ..., |s_{n-1} - \hat{\gamma}|$
    Let slope of separation curve $G = g_0, ..., g_{n-1} = |d_1 - d_0|, ..., |d_{n-1} - d_{n-2}|$
    Initialize maximum separation index $\mathfrak{i} \leftarrow 0$
    Let mean slope $\overline{g} = \text{mean}(G)$
    Let maximum separation index $\mathfrak{i} \leftarrow i$ be the first $i$ that satisfies $g_i > \overline{g}$.
    Return the correct set of EC numbers for query $\{EC_i\} = \{EC_0, ..., EC_{\mathfrak{i}}\}$
```

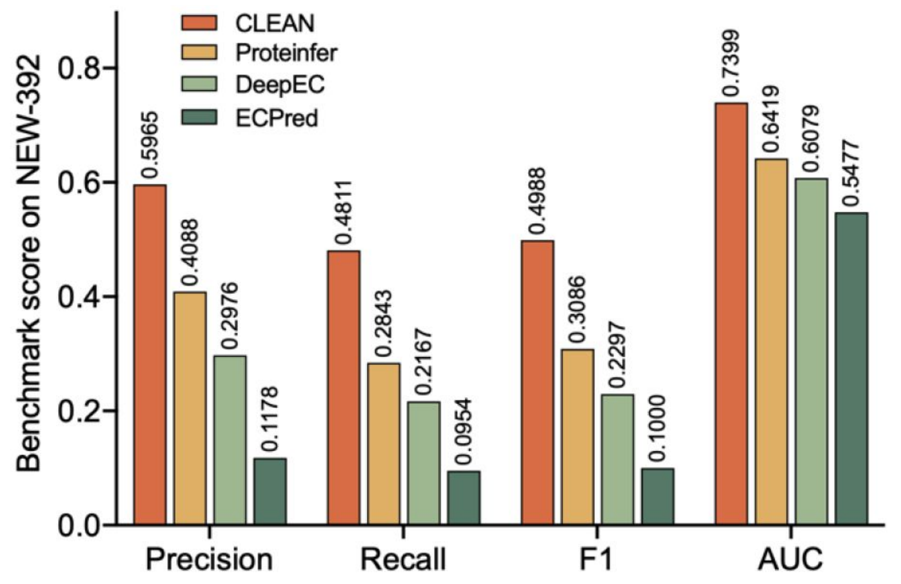
Benchmark with SOTA ML models using independent datasets

To evaluate the accuracy performance of CLEAN, two independent datasets were used to compare with two recently developed state-of-the-art ML models.

- ❖ **Price-161**[1]: curated by Price et al, where the annotations are *mislabeled or inconsistently*,
- ❖ **New-392**: recently published to Swiss-Prot dataset, unseen by any model during training,
- ❖ **Halogenases-36**: incompletely annotated halogenases

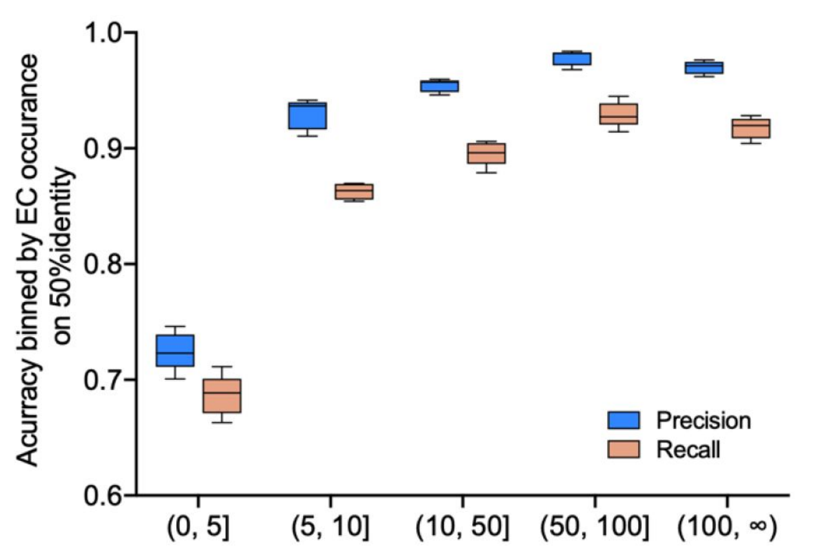


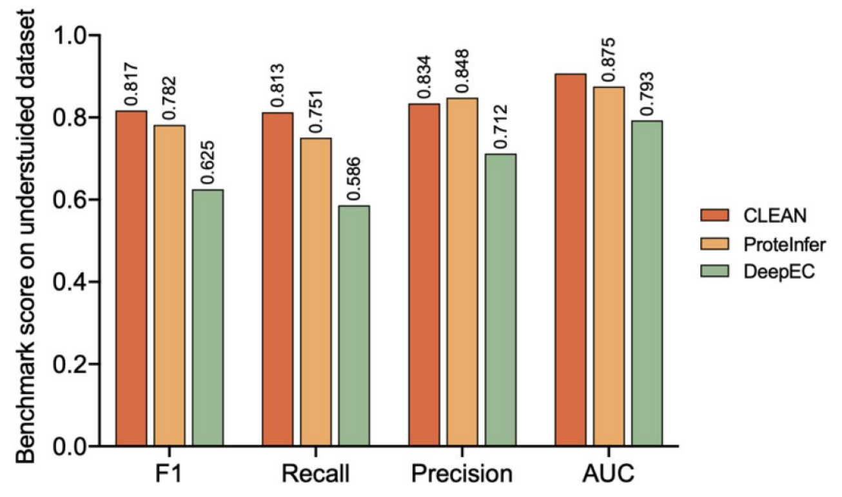


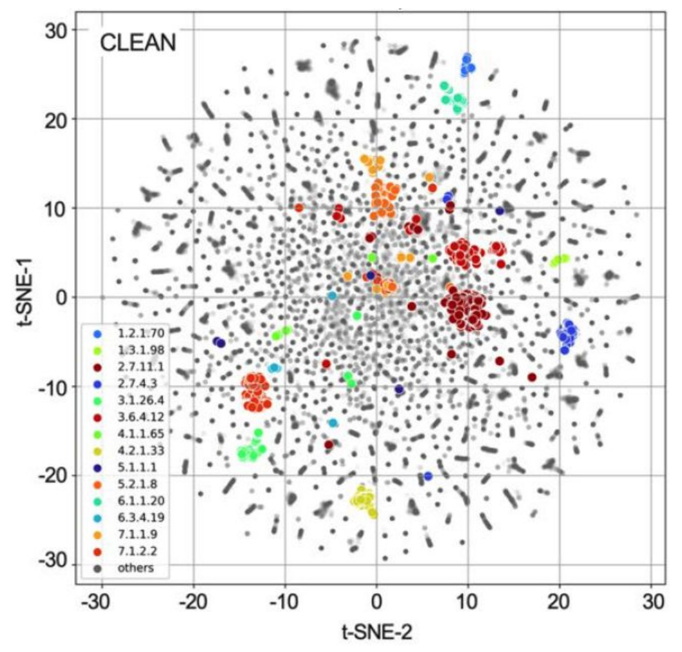


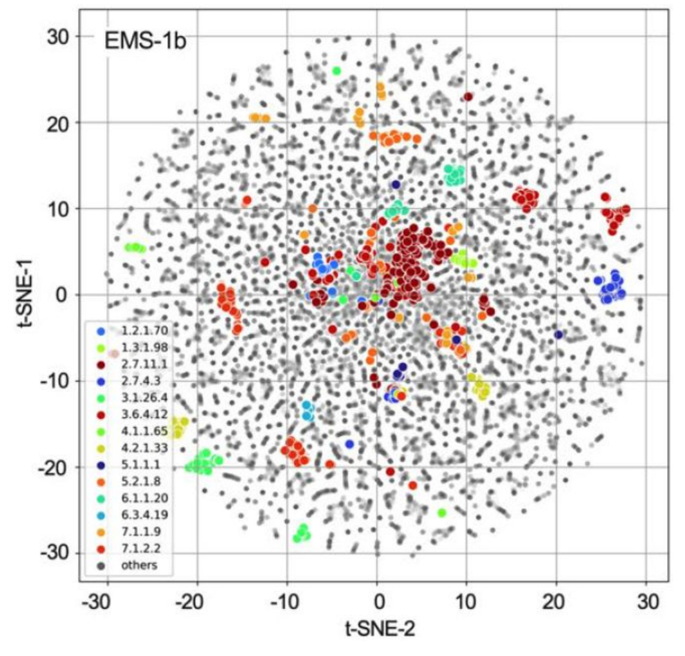
Accuracy Performance for Under-studied Enzymes

- ❑ Contrastive learning can be particularly useful for the prediction of **under-studied** enzymes.
- ❑ Contrastive learning can not only learn from **positive examples**, but also from **negative examples**.
- ❑ T-SNE dimension reduction visualizes the clustering of the embeddings after training by CLEAN.





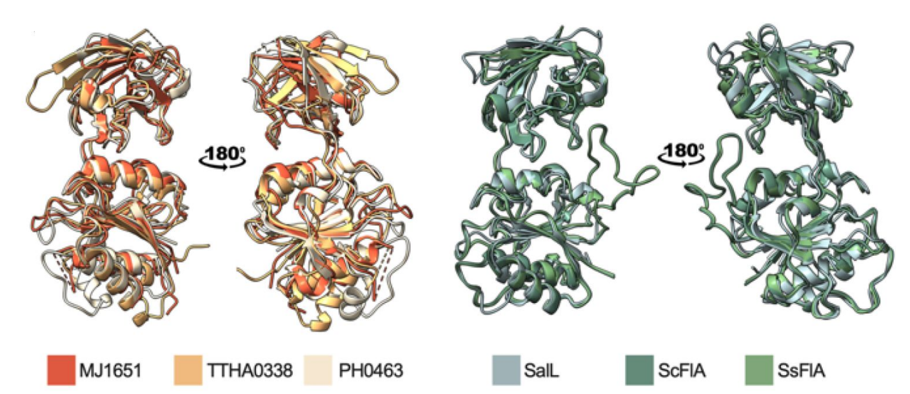


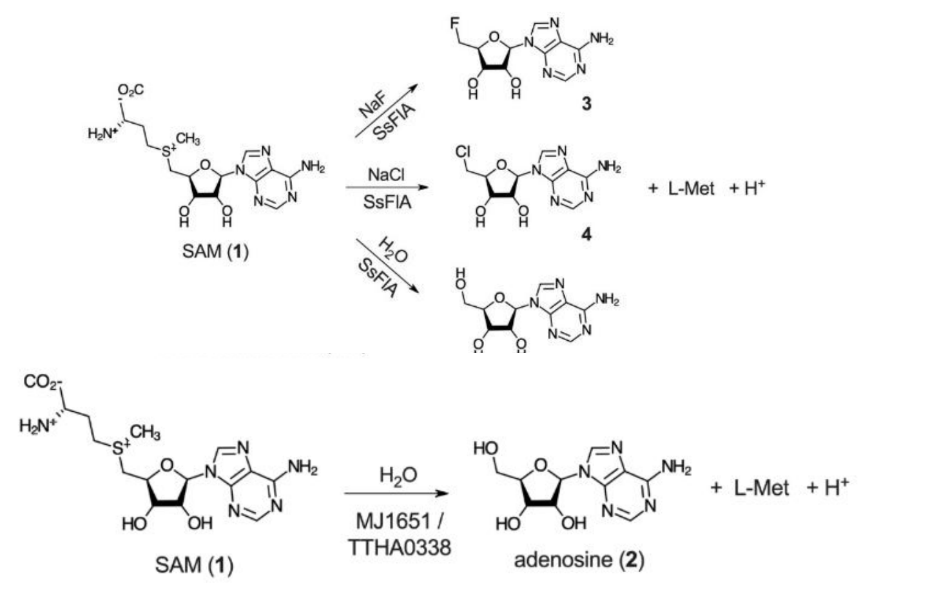


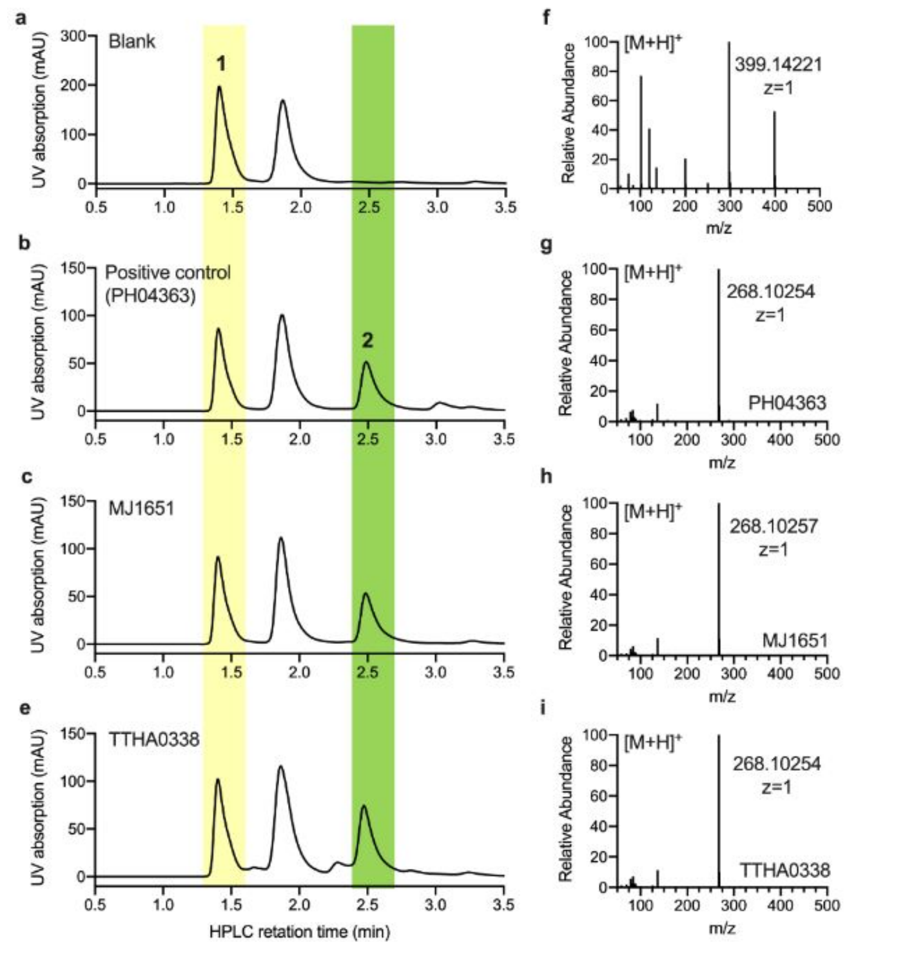
Experiment Validation Using Un- or Mis-labeled Halogenases

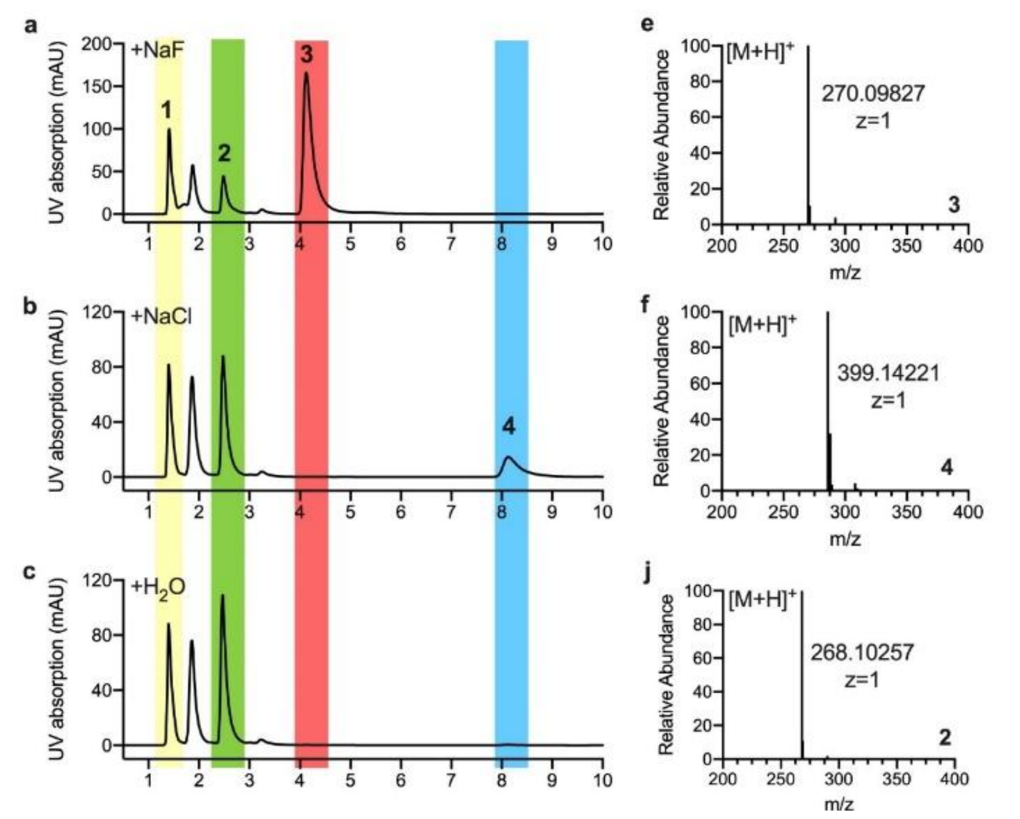
Three study cases are used here to evaluate CLEAN's prediction *in vitro*:

- ❖ **MJ1651**: Mislabeled by automatic annotation tools.
  - ➢ SAM hydrolase (**EC: 3.13.1.8**)
- ❖ **TTHA0338**: Uncharacterized protein.
  - ➢ SAM hydrolase (**EC: 3.13.1.8**)
- ❖ **SsFIA**: A promiscuous enzyme with three EC numbers.
  - ➢ SAM-dependent chlorinase, fluorinase and hydrolase
  - ➢ (**EC: 2.5.1.94, EC: 2.5.1.63, EC: 3.13.1.8**)









Reference

1. Price, M. N. et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* 557, 503–509 (2018).
2. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer: deep networks for protein functional inference. *BioRxiv* (2021).
3. Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. 116, 13996–14001 (2019).
4. Kustatscher, Georg, et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods* (2022).