Enzyme function prediction using contrastive learning (by Tianhao Yu et al.)

Presenter: Ding Bai, MBZUAI

July 13th, 2023

Outline

Enzyme Activities

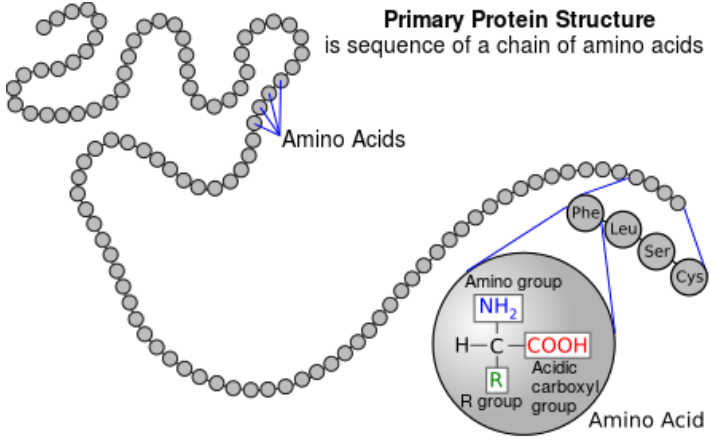
- Enzymes: biological catalyst proteins that accelerate chemical reactions.
- By DNA sequencing technologies (genomics and metagenomics), we identify many enzyme protein sequences (amino acid sequences).
- However:
  - Only a minuscule portion ($< 0.3\%$) of these proteins have been reviewed by human experts,
  - An even smaller subset ($< 19.4\%$) of them are supported by concrete experimental evidence.
- Need computational methods for function annotation
  - But $\approx 40\%$ automated annotations are incorrect.
  - Especially concerning understudied and promiscuous proteins.

Enzyme function annotation

- Enzyme commission (EC) number: the most well-known numerical classification scheme of enzymes.
- A system which specifies the catalytic function of an enzyme by 4 digits;
- Every EC number is associated with a recommended name for the corresponding enzyme-catalyzed reaction;
  - E.g. the code "EC 3.4.11.4" components indicate the following groups of enzymes:
  - EC 3 enzymes are hydrolases enzymes (enzymes that use water to break up some other molecule)
  - EC 3.4 are hydrolases that act on peptide bonds.
  - EC 3.4.11 are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide.
  - EC 3.4.11.4 are those that cleave off the amino-terminal end from a tripeptide.

Input and Output of this problem

- Input: Enzyme Protein sequences and structures (for this paper CLEAN, only sequences)



- Output: prediction of Enzyme Commission (EC) numbers (tree-like labels)

Current enzyme function annotation methods

- Experiments to decide the function of a target enzyme: laborious and expensive;
- Numerous computational tools.
- Similarity–based, homology-based, structure-based, and machine learning–based approaches.
- Sequence similarity–based Basic Local Alignment Search Tools for proteins (BLASTp) widely used;
  - Methods based solely on sequence similarity, making the prediction result less reliable when sequence similarity is low.
- Almost all the existing ML models (DeepEC and ProteInfer) are based on a multilabel classification framework;
  - Suffering from the limited and imbalanced training dataset that is common in biology.

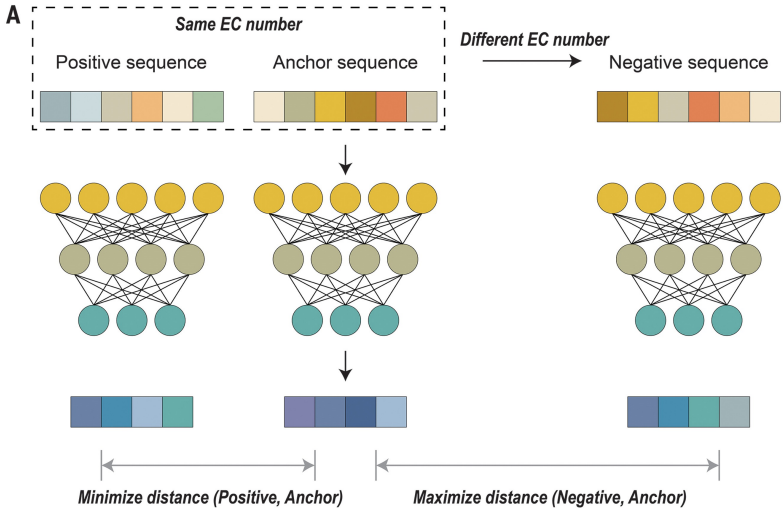
Contrastive learning–enabled enzyme annotation

- CLEAN predicts enzyme function from amino acid sequences.
  - Input: amino acid sequences without structure;
  - Output: a list of enzyme functions (EC numbers as the example) ranked by the likelihood.
- Trained on high-quality data from UniProt.
- Experiments:
  - In silico experiments;
  - In-house database of uncharacterized halogenases
  - In vitro experimental validation.
- Compared to other tools, including BLASTp and other advanced machine learning models, CLEAN exhibited superior performance in enzyme function annotation tasks.

Contrastive learning–enabled enzyme annotation (CLEAN)

- Aim to develop an embedding space for enzymes where functional similarities are represented by Euclidean distances.
- Similar enzyme sequences (with the same EC number) are close together, while dissimilar sequences are further apart.
- Contrastive losses for supervised training, where each enzyme sequence (anchor) was compared to both a similar (positive) and dissimilar (negative) sequence.
- Prioritizing negative samples that were still closely located to the anchor in the embedding space,
  - Not random. Improved training efficiency by
  - Which provides a more challenging comparison

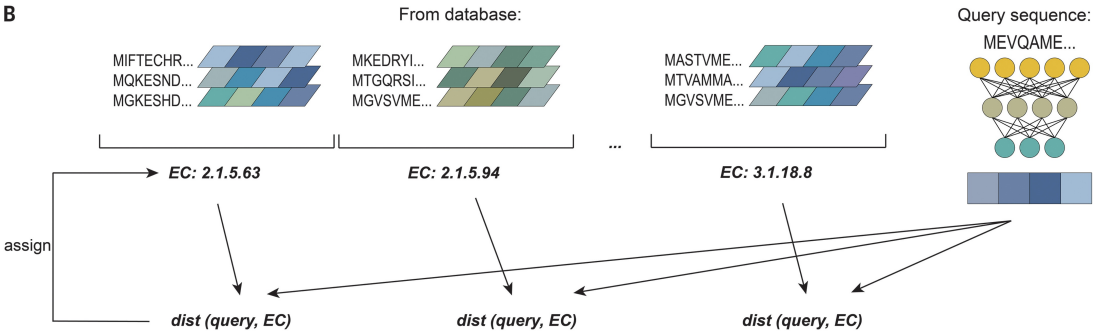
CLEAN: training structure



CLEAN: training process and backbone

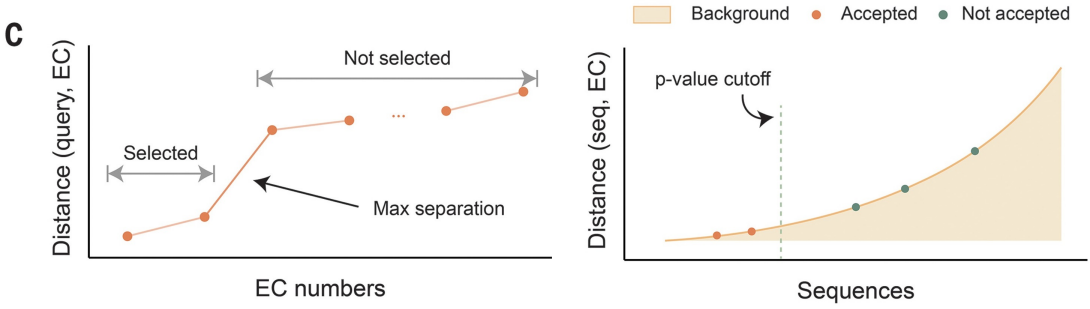
- Positive and negative samples are selected based on their EC numbers.
- The input sequence representations are embedded by **ESM-1b**.
  - A deep contextual language model on 250 million protein sequences.
  - Based on transformers. 33 layers. 650M parameters.
  - Input only the protein sequences, no structure information.
- These input sequences were then processed through a neural network.
- Output layer produced a refined, function-aware embedding of the input protein
  - Which is the function representation of input enzyme.

CLEAN: prediction



- The representations of an EC number are obtained by averaging the representations of enzymes under this EC number.
- The query sequence embedding is compared with each EC number's embedding to obtain the pairwise Euclidean distance between the query and each EC number.
- Distance: the similarity between EC numbers and the query sequence.

CLEAN: classification



- When used as a classification model, two methods:
  - 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.
  - A P value–based method that identifies EC numbers with statistical significance compared with background.

CLEAN: evaluation on the UniProt dataset

- On a 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. With 10%-split: 0.67 F1.
- CLEAN achieved much higher performance compared with the baseline method using ESM-1b without contrastive learning.

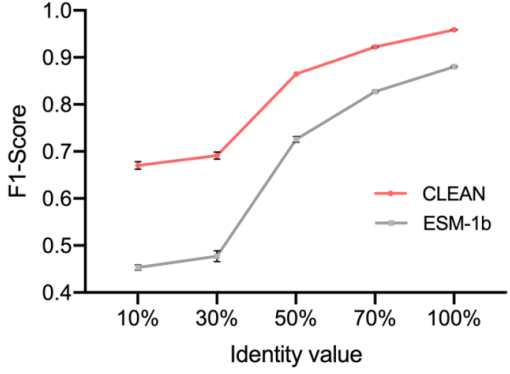
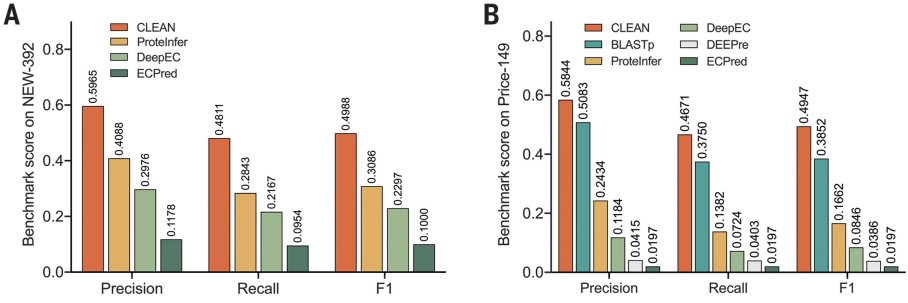


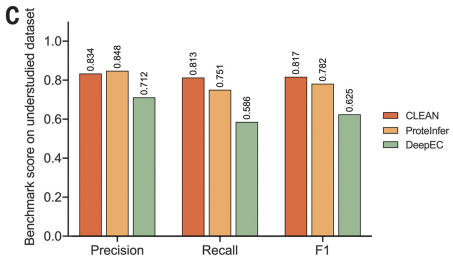
Fig. S1. The evaluation results of CLEAN on different identity clustering split under 5-fold CV (cross validated). ESM-1b was also investigated as comparison.

Benchmarking CLEAN with previous EC number annotation tools



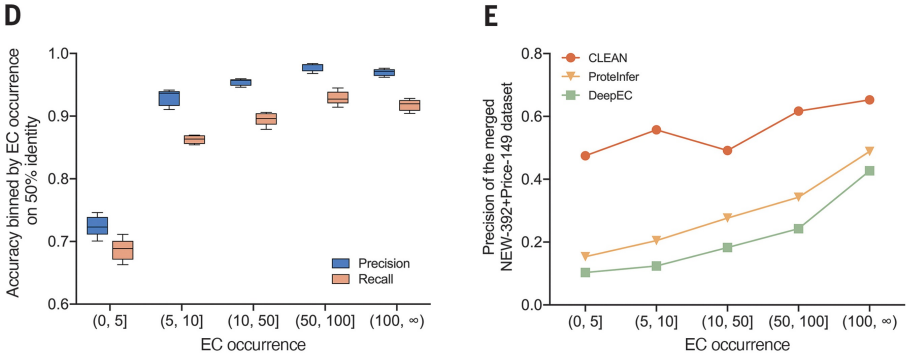
- Two independent datasets not included in any model's development.
- New-392: A practical situation.
  - 392 enzyme sequences covering 177 different EC numbers;
  - Labeled knowledgebase Swiss-Prot database, unknown functions of query sequences.
- Price-149: first curated by ProteInfer as a challenging dataset.
  - 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.

CLEAN's performance on annotating understudied EC number



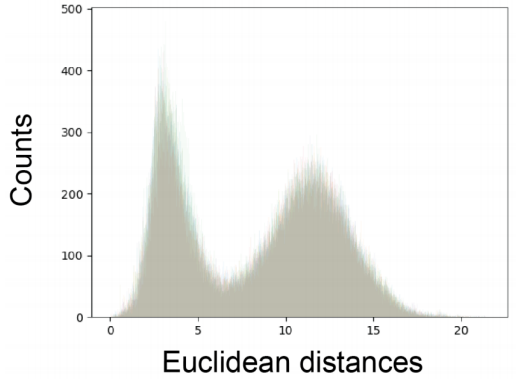
- On a validation dataset with enzymes from EC numbers rarely shown in the training set.
- ProteInfer and DeepEC had an advantage that both models have seen the validation dataset; CLEAN has not.
- Contrastive learning could better handle the imbalanced nature of EC numbers.
  - Where some EC numbers have many ($> 1k$) enzyme examples and some only have very few (< 5).

CLEAN's performance on annotating understudied EC number



- (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.
  - 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 in the similar way.

CLEAN's annotation confidence



- A method to quantify the prediction result confidence was implemented.
- Using 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.

Avoid overprediction

- The confidence quantification can also help CLEAN to avoid overprediction by reporting the third level of EC number when the confidence is low.

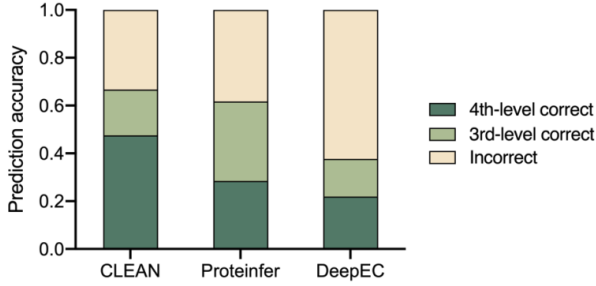
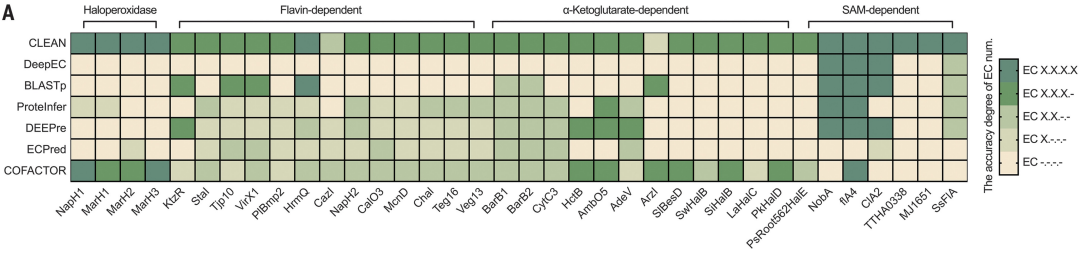


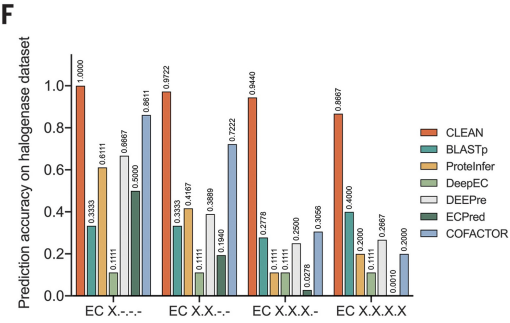
Fig. S14. The fraction of 4th level EC number accuracy and 3rd level EC number accuracy of the combined dataset of Price-149 and New-392.

Halogenases as a test case

- Halogenases, known for their catalyst-controlled selectivity, are increasingly used for biocatalytic C-H functionalization.
- The small molecules they produce, containing halogen atoms, have notable bioactivity and physicochemical properties, making them valuable in pharmaceutical and agrochemical fields.
- There are 36 partially annotated halogenases identified from UniProt, representing all four types of halogenases.



Halogenases as a test case

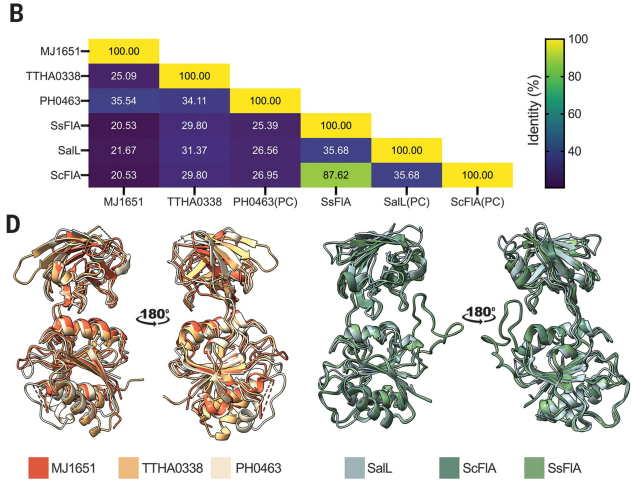


- Halogenase: particularly challenging because of the understudied halogenase family and the limited number of available halogenases.
- Result: CLEAN achieved much better prediction accuracy and it can distinguish enzyme functions even within the regime of similar biocatalytic reactions.

Mislabeling in UniProt

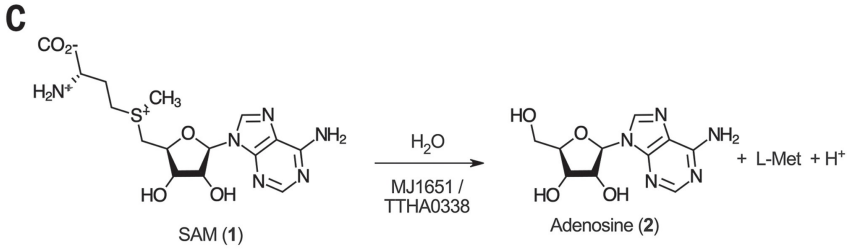
- Among these 36 halogenases, 3 enzymes named **MJ1651**, **TTHA0338**, and **SsFlA** showed conflicting functions according to the comparison between literature and the description in UniProt.
- CLEAN predicted new EC numbers in these 3 cases, suggesting that other potential functions might occur.
- Therefore, in vitro experiments to validate these predictions were performed.

Sequence identity and 3D-structure similarity



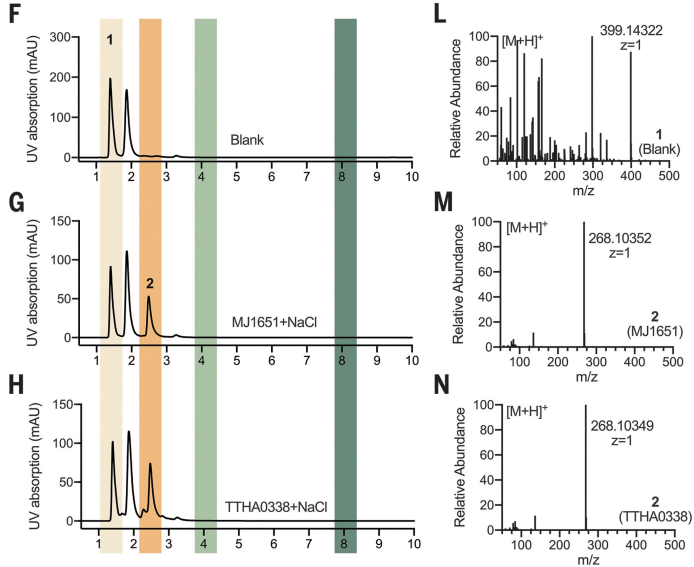
- The sequence identity for **SsFIA** and **ScFIA** is high.
- All have very similar structures with PC (positive control) enzymes.

SAM hydroxide adenosyltransferase MJ1651-TTHA0338 reaction

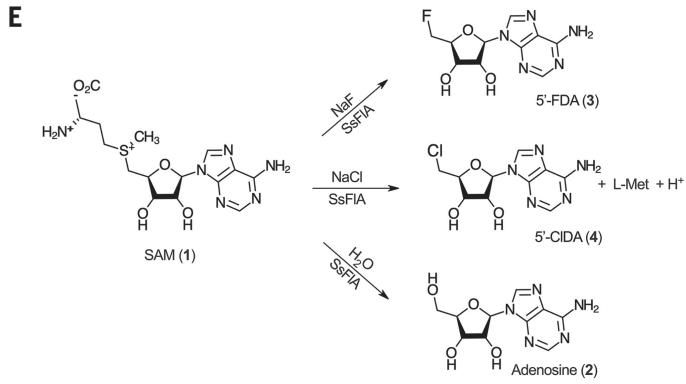


- In-vitro experiments confirmed that MJ1651 is SAM hydrolase (EC 3.13.1.8), as CLEAN predicted, rather than chlorinase (EC 2.5.1.94, mislabeled in UniProt) or fluorinase (EC 2.5.1.63, by computational tools).
- CLEAN also correctly annotated TTHA0338. All other six commonly used computational tools failed to predict MJ1641 and TTHA0338.

Blank and MJ1651-TTHA0338 reaction outcomes

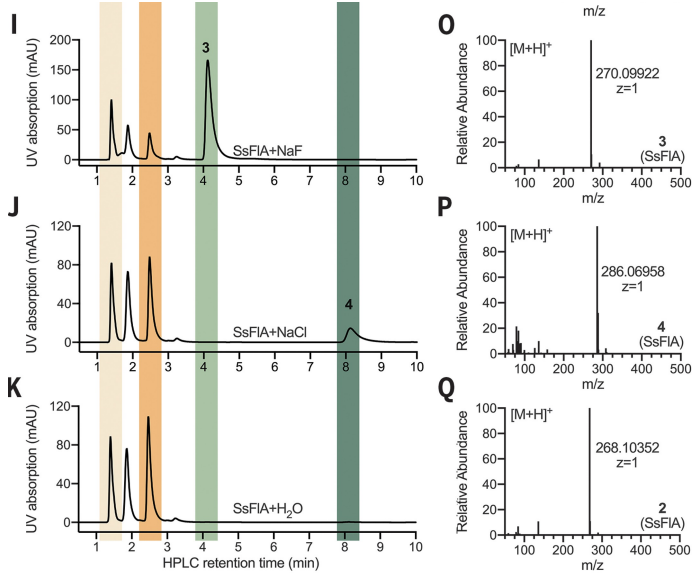


Nucleophilic substitution of SAM toward SsFlA



- SsFlA would be much harder
  - It has a very similar sequence and structure to PC enzymes.
  - It has 3 EC numbers (EC 2.5.1.63, EC 2.5.1.94, and EC 3.13.1.8).
- CLEAN can still correctly annotate SsFlA.

SsFlA reaction outcomes



Conclusion:

- CLEAN, a contrastive learning structure based on the ESM-1b backbone, to learn protein-function representations for enzymes.
- Very solid experiments: in silico and in vitro testing has shown that CLEAN outperforms six other leading tools in prediction performance.
- Comprehensive analysis of an uncharacterized halogenase dataset revealed CLEAN's ability to correctly identify hypothetical proteins and correct mislabeled ones.
- CLEAN has proven effective at identifying enzyme promiscuity, essential for enhancing the performance of existing enzymes.
- CLEAN's methodology, contrastive learning based on the representations from other pretrained LLM models, could be adapted to other predictive tasks beyond enzymatic activities.

References

[1] Altschul, Gish, et al., “Basic local alignment search tool”
[2] Altschul, Madden, et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”
[3] Blum et al., “The InterPro protein families and domains database: 20 years on”
[4] Krogh et al., “Hidden Markov models in computational biology: Applications to protein modeling”
[5] Radivojac et al., “A large-scale evaluation of computational protein function prediction”
[6] Hult and Berglund, “Enzyme promiscuity: mechanism and applications”
[7] Steinegger et al., “HH-suite3 for fast remote homology detection and deep protein annotation”
[8] “UniProt: the universal protein knowledgebase in 2021”
[9] Wheeler et al., “Database resources of the national center for biotechnology information”
[10] C. Zhang, Freddolino, and Y. Zhang, “COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information”
[11] Desai et al., “ModEnzA: accurate identification of metabolic enzymes using function specific profile HMMs with optimised discrimination threshold and modified emission