sBioinformatics, 2023
Advance Access Publication Date: 18/12/2023
Original Research

4simEC: A Novel and User-Friendly Tool for Enzyme Prediction/Annotation Using Structure and Sequence Similarity

Yun-Ching Chien [1*], Panagiotis Delaroudis [1*], Anna Liuptak [1*], Yu-Jie Qiu [1*]

*To whom correspondence should be addressed.

[1]Department of Bioinformatics, KU Leuven

Abstract

Motivation: The Enzyme Commission (EC) number is a numerical classification system for enzymes based on the different chemical reactions. High-quality prediction of the EC numbers is critical for understanding the enzyme functions and the cellular metabolism. Over the last two decades, numerous tools have been developed to predict EC numbers. Traditionally, sequence similarity, particularly with BLAST, has been a mainstay for annotating new sequences with EC numbers. More recent methods like Foldseek enable EC number annotation through protein structural similarities. However, current databases and methods are not well suited for EC number annotation due to the lack of integrated sequence similarity and structural similarity.
Results: The new visualization tool allows users to query and get a knowledge graph about the inferred EC number. The tool also provides similar search results and predictive EC numbers for users to download. Comparative analyses against the available EC number prediction tools (DeepEC and ECPred) show that the new 4simEC tool has better performance of predicting EC numbers than DeepEC and ECPred.

Availability: Scripts provided through GitHub: https://gitlab.kuleuven.be/csb/enzymares/enzyme-knowledge-graph-dev

1 Introduction

Enzymes play a critical role in regulating various biological processes in the human body. Annotation of enzyme function has diverse applications, such as metagenomics, industrial biotechnology, and diagnosis of enzyme deficiency-caused diseases. Usually, the experience technique is using enzyme assay which provides a direct and accurate means of exploration[1]. However, due to the time, cost and expertise required, experiments may not deal with huge and newly discovered enzymes. As a result, when dealing with numerous and newly discovered enzymes, there is a need for an efficient annotation tool to facilitate scientific research and guide the formulation of experiments for validation. Moreover, the knowledge graph provides a valuable resource for understanding not only how sequences are linked but also identifying instances where no connection significantly exists[2].

1.1 Enzyme Commission (EC) Number System

Enzymes are classified into a hierarchical numerical system based on the types of reactions they catalyze that are established by the International Union of Biochemistry and Molecular Biology (IUBMB)[3]. Each EC number consists of four digits, with each digit representing a specific aspect of the enzyme's function[4]. The first digit broadly categorizes enzymes into one of six main classes: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases and (6) ligases, while subsequent digits provide increasing specificity about the enzymatic activity, substrate, and product. Take the fatty acid elongated enzyme, which is annotated as EC 2.3.1.199, as an example, the '2' denotes that it is a transferase; the '3' indicates that it acts upon acyl groups; the '1' shows that it transfers groups other than aminoacyl groups; and the '199' suggests that it is a very-long-chain 3-oxoacyl-CoA synthase.

1.2 Protein and EC number databases

The UniProt comprehensive database provides users with protein sequences functional annotation. The Swiss-Prot is part of the UniProt and specifically focuses on manually curated and annotated protein sequences[5]. The curation process involves experts who review and annotate experimental data, ensuring the accuracy and reliability of the information provided. According to Swiss-Prot released in November 2023, it contains 570,420 manually annotated proteins and 283,117 of which are enzymes. These enzymes are classified using the EC system. Some available databases for users to get information of EC numbers include: (1) Expasy Enzyme which is a repository of information relative to the nomenclature of enzymes[6]. (2) BRENDA database covers enzyme classes and metabolic pathway[7]. (3) KEGG Enzyme links enzymes to metabolic pathways and genome annotation[8].

1.3 Sequence similarity in EC number annotation

Traditionally, sequence similarity has been a cornerstone for annotating new sequences with EC numbers. Tools like BLAST (Basic Local Alignment Search Tool) finding regions of similarity between sequences[9]. The algorithm for scoring is using substitute matrix to assign a value to each aligned pair of amino acid letter. The more likely substitutions, the higher values, otherwise the unlikely substitutions have negative values. Finally, sum this value up and calculate the statistical significance of the matches. Through this strategy, BLAST can use sequence similarity to predict enzyme function by finding the correlation between unknown sequences and annotated sequences.

As instrumental progress and data get larger, such as BLAST are too slow to handle huge data sets. MMseqs2 (Many-against-Many sequence searching) is a tool suited to search and cluster huge protein sequence sets. MMseqs2 runs 10,000 times faster than BLAST, but it can achieve almost the same sensitivity as BLAST[10]. Clustering is a technique to assign protein sequences into different groups based on sequence similarity. Proteins exceeding sequence similarity threshold will be in the same group. Thus, clustering can discover the biological relationship of homologous sequences. Moreover, clustering can speed up downstream analysis by reducing highly similar sequences to a single representative cluster. The key strategy of MMseqs2 run fast involves precomputing similar k-mers above a certain similarity threshold for each target profile. These k-mers are stored in an index table. When searching for matches, the algorithm iterates over the overlapping, spaced k-mers in the query sequence, checking the index table only for exact matches (same k-mers) (Figure. 1)[11].

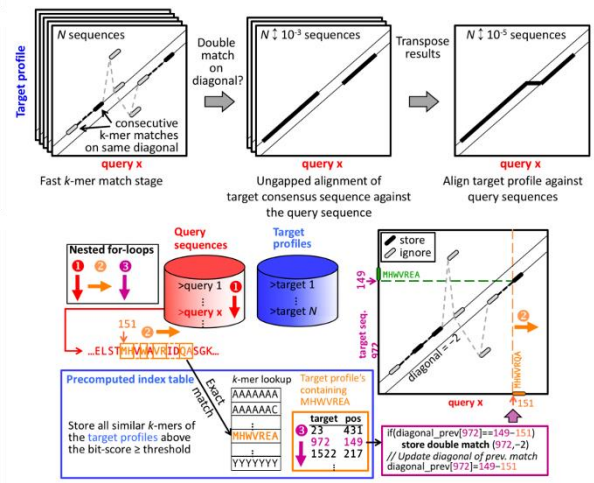

Figure 1. Algorithmic change to perform fast sequence profile searches using MMseqs2[11].

1.4 Importance of Structure Similarity in EC Number Prediction

However, the approach of predicting EC number based on sequence similarity has limitations, particularly when dealing with geometrically similar proteins that may share similar functions but exhibit low sequence similarity. Many proteins cannot be annotated because of detecting distant evolutionary relationships from sequences alone. Such limitations could result in the loss of valuable biological information. Therefore, searching for structural relationships independent of sequence relationship is critical and expected to significantly increase the number of deriving functional inferences.

More recently, advances in structural biology have underscored the significance of structural similarity in EC number prediction. Tools such as Foldseek can cluster protein structural information to infer enzymatic functions, providing valuable for predicting function based on structure similarity, even when sequence similarity is low[12]. The structural alignment tools are slow because structural similarity scores are non-local; Chang the alignment in one part would affects the similarity in all other parts. The state-of-the-art Foldseek tool copes with the speed by converting structure into a sequence and comparing structures using sequence alignments. Each residue is represented by a structural state letter (Figure. 2)[13,14]. This expresses the structural features that read off the secondary structure elements. The Foldseek tool enables us to utilize structural comparison methods to enhance homology inference, as well as to conduct analyses related to structure, function, and evolution. As a result, integrating both sequence and structural similarity enhances the accuracy and reliability of EC number prediction, bridging gaps by methods solely reliant on sequence data.

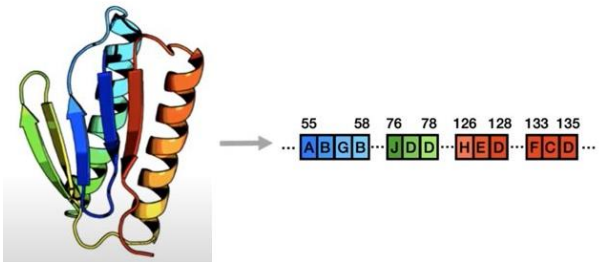

Figure 2. Structural alphabets reduce structure comparisons to much faster sequence alignments[13,14].

2 Materials and Methods

2.1 Data Resources

Protein sequences, integrated into the neo4j database and employed for the similarity search, were derived from Swiss-Prot, the reviewed segment of the UniProt database (https://www.uniprot.org/). Specifically, fasta files were obtained directly from the UniProt website while pdb files of the same proteins were taken from the precomputed Foldseek Swiss-Prot database. Metadata for the Swiss-Prot proteins, such as protein names, taxonomy, information about protein function, EC numbers, and links to external resources, were extracted from the UniProt website using customized columns.

2.2 Aggregating and Clustering Data

Protein sequences were grouped in clusters of sequentially and structurally similar proteins. Command line tools MMseqs2 (https://github.com/soedinglab/MMseqs2) and Foldseek (https://github.com/steineggerlab/foldseek) were used for the clustering based on sequence and structure similarity, respectively. Both clustering procedures were executed iteratively with diverse identity percentages, including 80%, 85%, 90%, 95%, and 99%. The resulting clusters and Swiss-Prot meta-data were further

manipulated and converted into CSV files for subsequent storage within the neo4j database.

2.3 Building neo4j graph database

The process of building the Neo4j graph database began with the initial transformation of Swiss-Prot metadata into separate CSV files. This involved the creation of five node labels: Sequence, ECNumber, Organism, Species and Genus. Additionally, two types of relationships, BELONGS_TO and HAS_A, were established to connect these nodes. Each sequence node was enriched with crucial metadata as node properties, including Entry ID, Entry Name, UniParc ID, Catalytic activity, Gene Names, Protein names, AlphaFoldDB ID, PDB ID and amino acid sequence.

Subsequently, CSV files for sequence and structure cluster nodes were generated based on cluster representatives, labeled as SequenceCluster and StructureCluster, respectively. Cluster members were identified by their sequence node ID and connected to representatives using the IS_IN relationship type, with the identity percentage thresholds included as properties of these relationships.

Finally, the neo4j-admin command line bulk importer was used for loading all the generated CSV files into the Neo4j database. To ensure data integrity, the "--skip-bad-relationships" parameter was utilized, allowing us to bypass importing relationships that referred to missing node IDs—specifically, those linked to obsolete or merged entries within Swiss-Prot.

2.4 Pipeline for similarity search

For sequence similarity search, the input query sequence provided is in FASTA format. The workflow for this process is illustrated in Figure 3.

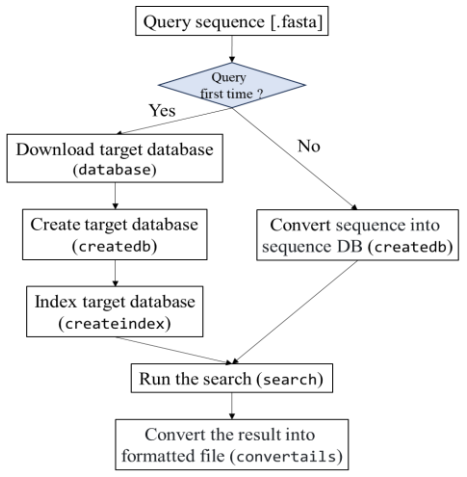

Figure 3. The pipeline of sequence similarity search
In each step of the pipeline, the corresponding command is denoted in parentheses. The Swiss-Prot database serves as the default target database. The search command employs the parameter "-a" to include alignment details. The resulting file is structured as a tab-separated list featuring 12 columns, which provide comprehensive alignment information, including target sequences, E-value and bit score.

In the case of structure similarity search, the query sequence is in PDB format. The workflow is depicted in Figure 4.

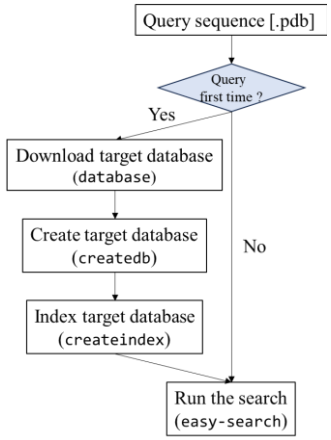

Figure 4. The pipeline of structure similarity search
The main module used in each step is indicated in parentheses. The default target database is the Alphafold/Swiss-Prot. The final file is in the same format as the one from structure similarity search.

2.5 Creating the Streamlit interface

4simEC has an interactive user interface built on Streamlit version 1.27.2 within python 3.9.18 . Streamlit is an open-source Python library that allows developers to quickly build user-friendly web interfaces. It provides flexible layouts and versatile widgets that could accommodate many customized requirements.

2.6 Benchmarking

Before the benchmarking there were some particular limitations to consider. Firstly, the Steamlit app allows only one input at a time making testing multiple queries a costly procedure timewise. Secondly, since the neo4j database includes proteins from the Swiss-Prot database, these would need to be excluded from any test dataset as the tool would just match them with themselves and simply display their already known EC number without actually predicting it.

After considering the limitations above, it was decided to use a small number of proteins to create two test datasets. The first one was made of unreviewed proteins not included in the Swiss-Prot database, whose EC numbers have been determined, in order to properly test the predictive ability of the tool. The second one did include proteins from Swiss-Prot, but these were removed from the neo4j database in order to test how the app would work on them as well.

DeepEC and ECPred are two tools designed for predicting EC numbers by employing different methodologies. DeepEC utilizes deep learning algorithms to predict EC numbers, while ECPred combines machine learning techniques with functional domain information. DeepEC uses three convolutional neural networks (CNNs) as a primary engine for predicting EC numbers and implements homology analysis for EC numbers that cannot be classified by the CNNs[15]. On the other hand, ECPred provides probabilistic predictions for enzymatic functions across all levels

of the EC hierarchy for uncharacterized protein sequences[16]. Among its predictors, SPMap relies on subsequences, which involve subsequence extraction, clustering, and the construction of probabilistic profiles. A complication here stems from the fact that many enzymes are involved in multiple different catalytic activities and thus possess many different EC numbers. In order to evaluate the success of any particular tool, we decided that if one of the correct EC numbers associated with an enzyme was predicted, it would be considered a hit.

3 Results and Discussion

3.1 The neo4j database

Upon the incorporation of seven node types and three relationship types, the neo4j database resulted in 570157 Sequence nodes, 5720 ECNumber nodes, 14509 Organism nodes, 2218 Species nodes, 5396 Genus nodes, 450480 SequenceCluster nodes and 439074 StructureCluster nodes. These nodes were interconnected by a total of 586727 BELONGS_TO, 298114 HAS_A and 5562180 IS_IN relationships. It is noteworthy that the IS_IN relationships have properties with identity percentage thresholds of 80%, 85%, 90%, 95% or 99%.

3.2 The 4simEC interface and EC number prediction

4simEC provides a Streamlit interface that enables users to run sequence and structure similarity search, and further infer the EC numbers of the query sequence based on the results.

Before conducting the similarity search, users are asked to connect to the neo4j database (Figure 5). After the search, the identified target sequences are presented in tab-separated list format along with other alignment information. Users can then specify the number of target sequences they prefer to visualize in a node-edge graph. Notably, in the graph, different node labels are distinguished by colors and edges connecting cluster nodes vary in width to represent different identity percentage thresholds (Figure 6). Additionally, external links to resources related to the selected target sequences, such as KEGG or BRENDA, would also be displayed if applicable.

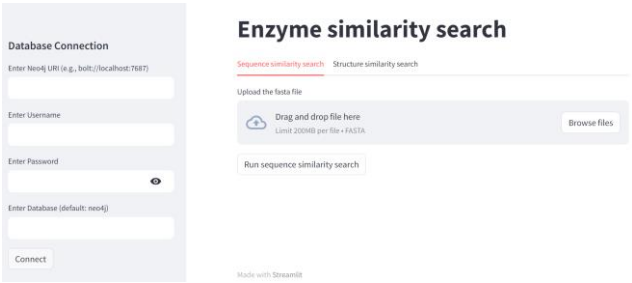

Figure 5. The initial page of 4simEC interface.
The snapshot illustrates the layout of the 4simEC. The sidebar on the left-hand side allows users to connect to the Neo4j database. Users can then select and conduct the sequence or structure similarity search by clicking on different tabs above. Afterward, users can upload their query sequence and start searching.

Furthermore, users can specify their preferences regarding EC number predictions based on either sequence or structure clusters, along with the desired similarity percentages (Table 1). The prediction would be displayed as a dataframe and allow users to download it for further analysis.

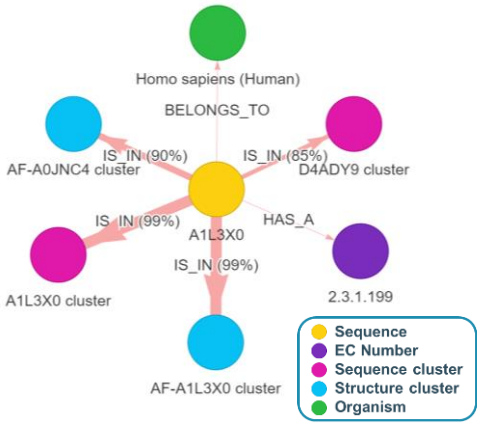

Figure 6. Example of a node-edge graph.
This graph consists of a single target sequence (in yellow), which belongs to human (in green) and has an EC number (in purple). Additionally, the width of the edges connecting cluster nodes varies according to identity percentages.

	Similar sequence	EC number	SequenceCluster (99%)	SequenceCluster (90%)	
0	A1L3X0	2.3.1.199	A1L3X0 (2.3.1.199)	A0JNC4 (2.3.1.199)	A1L3X0 (2.3.1.199)
1	A0JNC4	2.3.1.199	A0JNC4 (2.3.1.199)	A0JNC4 (2.3.1.199)	A1L3X0 (2.3.1.199)
2	Q9D2Y9	2.3.1.199	Q9D2Y9 (2.3.1.199)	D4ADY9 (2.3.1.199)	Q9D2Y9 (2.3.1.199)
3	D4ADY9	2.3.1.199	D4ADY9 (2.3.1.199)	D4ADY9 (2.3.1.199)	Q9D2Y9 (2.3.1.199)
4	Q9BW60	2.3.1.199	Q9BW60 (2.3.1.199)	Q9BW60 (2.3.1.199)	Q9JLJ5 (2.3.1.199)

Table 1. Example of EC number prediction.
This example table displays the top five target sequences along with their corresponding EC numbers. Additionally, it includes other members of the sequence cluster at 99% and 90% similarity. This dataframe provides comprehensive information about the inferred EC numbers. The columns can also be customized to show various combinations of clusters (sequence and/or structure) and identity percentages (80%, 85%, 90%, 95%, 99%). In this specific case, it is highly likely that the EC number of the query sequence is 2.3.1.199.

3.3 Benchmarking

Below are the results obtained from testing the 4simEC app against DeepEC and ECPred (Figure 7). As the graph clearly shows, the Streamlit app was more successful than both DeepEC and ECPred (spmap) for these two datasets. More specifically, it correctly identified the EC numbers for all ten Swiss-Prot proteins that were removed from the neo4j database as well as 88% of the unreviewed proteins from the TrEMBL dataset. In comparison, ECPred matched the app's performance in the Swiss-Prot dataset but did much worse in the unreviewed one while DeepEC performed well in both but was still below the app.

Additionally, in order to test the structural search method, we used the PDB files of the unreviewed dataset's proteins and ran those as inputs. Although the PDB files of three of those proteins could not be found, the EC numbers were correctly predicted for all of the rest.

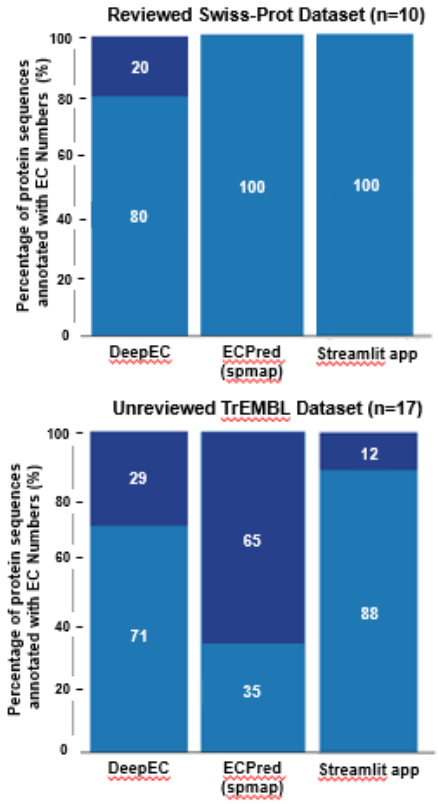

Figure 7. Comparison of DeepEC, ECPred, and 4simEC.
The light blue parts denote successful hit while the dark blue misses.

4 Conclusion

In summary, 4simEC is a novel and user-friendly tool designed for efficient similarity search and EC number annotation. By leveraging the speed and accuracy of state-of-the-art algorithms MMseqs2 and Foldseek, 4simEC enables fast EC number prediction based on either sequence or structure similarity. The tool utilizes a graph database that was built during the project. The database provides required data for EC number inference and offers additional information and graphical representation of the results. Remarkably, on a small dataset, 4simEC demonstrated a more accurate EC number prediction based on protein sequences in comparison with widely employed methods DeepEC and ECPred. Furthermore, 4simEC delivered accurate predictions based on protein structures for all tested proteins. It is worth noting that there is potential for improvement, particularly in enhancing functionality. Additionally, rigorous testing on larger datasets could be of interest.

Acknowledgements

We would like to thank Prof. van Noort for providing us feedback and suggestions on the project and our supervisor Jaldert François for the guidance through the project.

References

1. Jean-Philippe Goddard, Jean-Louis Reymond. Enzyme assays for high-throughput screening. Curr Opin Biotechnol. 2004;15(4):314-22. doi: 10.1016/j.copbio.2004.06.008.
2. Daniele Toti, Gabriele Macari, Enrico Barbierato, Fabio Polticelli. FGDB: a comprehensive graph database of ligand fragments from the Protein Data Bank. Database (Oxford). 2022;1-12. doi: 10.1093/database/baac044.
3. Ida Schomburg, Antje Chang, Dietmar Schomburg. Standardization in enzymology–Data integration in the world's enzyme iInformation system BRENDA. Perspec. Sci. 2014;15-23. doi: 10.1016/j.pisc.2014.02.002.
4. Athel Cornish-Bowden. Current iubmb recommendations on enzyme nomenclature and kinetics. Perspect. Sci. 2014;74–87. doi: 10.1016/j.pisc.2014.02.006.
5. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023; 51(D1): D523–D531. doi: 10.1093/nar/gkac1052.
6. Elisabeth Gasteiger, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D. Appel, and Amos Bairoch. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003; 31(13): 3784–3788. doi: 10.1093/nar/gkg563.
7. panelI. Schomburg, L. Jeske, M. Ulbrich, S. Placzek, A. Chang, D. Schomburg. The BRENDA enzyme information system–From a database to an expert system. 2017; 261:194-206. doi: 10.1016/j.jbiotec.2017.04.020.
8. Minoru Kanehisa. Enzyme Annotation and Metabolic Reconstruction Using KEGG. Methods Mol Biol. 2017; 1611:135-145. doi: 10.1007/978-1-4939-7015-5_11.
9. Scott McGinnis* and Thomas L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32: W20-W25. doi: 10.1093/nar/gkh435.
10. Maria Hauser, Martin Steinegger, Johannes Söding. MMseqs software suited for fast and deep clustering and searching of large protein sequence sets. 2016;32(9):1323-30. doi: 10.1093/bioinformatics/btw006.
11. Steinegger M and Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017; 35(11):1026-1028. doi: 10.1038/nbt.3988.
12. Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron L. M. Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. Nature. 2023; 622:637–645. doi: 10.1038/s41586-023-06510-w.
13. Matthew Hutson. Foldseek gives AlphaFold protein database a rapid search tool. Nature Technology Feature. 2023. doi: 10.1038/d41586-023-02205-4.
14. Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding & Martin Steinegger. Fast and accurate protein structure search with Foldseek. Nature Biotechnology. 2023. doi: 10.1038/s41587-023-01773-0.
15. Jae Yong Ryu, Hyun Uk Kim, Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996-14001. doi: 10.1073/pnas.1821905116.
16. Alperen Dalkiran, Ahmet Sureyya Rifaioglu, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, Tunca Doğan. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics. 2018;19(1):334. doi: 10.1186/s12859-018-2368-y.