



APPROVED: 8 October 2024
doi:10.2903/sp.efsa.2024.EN-9063

Development of *in silico* methodologies to predict the toxicity of novel proteins in the context of food and feed risk assessment

L. Palazzolo¹, T. Laurenzi¹, O. Ben Mariem¹, A. Bassan², U. Guerrini¹, I. Eberini¹

¹. Dipartimento di Scienze Farmacologiche e Biomolecolari, Università degli Studi di Milano, Via Balzaretti 9, Milano, Italia

². INNOVATUNE, Via Giulio Zanon 130/D, Padova, Italia

Abstract

This report is the outcome of an EFSA procurement (OC/EFSA/GMO/2021/02 – LOT1) aiming at developing an *in silico* strategy to predict the toxicity of (novel) proteins. Up-to-date, commercially available tools predicting protein toxicity based on primary structures were evaluated for their accuracy and usability, using a curated dataset of annotated toxins and non-toxins from UniProt. ToxinPred2 and Toxify emerged as the top performers, showing both high accuracy and suitability for integration into an automated pipeline. Additional bioinformatics methods were explored, which provide sequence similarity-based information rather than direct predictions (BLAST, InterPro HMM profiles). By converting their outputs into features for machine learning models, a high prediction accuracy was achieved, though there is potential for improvement to reduce overfitting risks. An Artificial Intelligence (AI)-based consensus pipeline, integrating results from ToxinPred2, Toxify, and our machine learning models was developed. This consensus model reached a 95% accuracy rate in distinguishing toxins from non-toxins. Noteworthy, our BLAST-based machine learning model - although performance-wise comparable to BLAST - offers higher sensitivity and specificity across diverse queries than BLAST; it relies on database-based evolutionary relationships, which may significantly limit its applicability to novel or mutated toxins. Structure-based prediction methods are deemed impractical due to their resource intensity and reliance on accurate structural data; AI-driven structure prediction methods - like Rosetta and AlphaFold - are promising, however they are still under development and may not be suitable for the regulatory context yet. Recommendations are provided, including enhancement of the proposed consensus pipeline to create an independent open-source, user-friendly tool for evaluating the safety of (novel) proteins in food and feed; regular updates of the proposed databases and models; incorporation of 3D structures and in general validation of AI and machine learning models for regulatory uses.

© European Food Safety Authority, 2024

Keywords: *In silico*, protein toxicity, toxin, artificial intelligence, methodologies, tools.

In silico methodologies to predict the toxicity of novel proteins



Question number: EFSA-Q-2024-00605

Correspondence: NIF@efsa.europa.eu



Disclaimer: The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Suggested citation: Palazzolo L, Laurenzi T, Ben Mariem O, Bassan A, Guerrini U and Eberini I, 2024. Development of in silico methodologies to predict the toxicity of novel proteins in the context of food and feed risk assessment. EFSA supporting publication 2024:EN-9063. 99 pp. doi:10.2903/sp.efsa.2024.EN-9063

ISSN: 2397-8325

© European Food Safety Authority, 2024

Reproduction is authorized provided the source is acknowledged. EFSA may include images or other content for which it does not hold copyright. In such cases, EFSA indicates the copyright holder and users should seek permission to reproduce the content from the original source.

The designations employed and the presentation of material on any maps included in this scientific output do not imply the expression of any opinion whatsoever on the part of the European Food Safety Authority concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.



Table of contents

Abstract.....	1
Table of contents	4
1 Introduction	6
1.1 General background	6
1.2 A note on in silico tools and methodologies to predict protein toxicity	6
1.3 Background and terms of reference as provided by the requestor	7
1.4 Interpretation of the Terms of Reference	8
1.5 Additional information - Project plan	9
2 TASK 1: Protocols to gather information on tools and methodologies predicting protein toxicity and to set protein benchmark datasets	9
2.1 Protocol for literature search to identify tools and methodologies	9
2.2 Protocol to set protein datasets	11
3 TASK 2 and TASK 3: Gathering information on tools and methodologies predicting protein toxicity and setting protein benchmark datasets	12
3.1 Literature search: Tools and methodologies predicting protein toxicity (TASK 2)	12
3.1.1 Introduction	12
3.1.2 Methodology	12
3.1.3 Results	15
3.2 Selection of tools and methodologies for an in silico prediction strategy of protein toxicity	18
3.2.1 Results	19
3.3 Protein datasets setting (TASK 3)	20
3.3.1 Introduction and Aim	20
3.3.2 Evaluation of the application scope of the selected tools	20
3.3.3 Main toxic protein datasets creation	21
3.3.4 Lowering the redundancy within datasets	22
3.3.5 Generation of the True Negatives datasets	25
3.3.6 Generation of the Expected False Negatives (EFN) datasets	26
3.3.7 Generation of final datasets for each tool	27
4 TASK 4: assessment of tools and methodologies and pipeline definition	30
www.efsa.europa.eu/publications	

In silico methodologies to predict the toxicity of novel proteins



4.1	Testing tools against protein datasets	30
4.2	Testing methodologies	33
4.2.1	BLAST	33
4.2.2	Hidden Markov Model (HMM) profile alignment	35
4.3	Creation of a consensus model and pipeline	37
4.3.1	Generation of machine learning models based on bioinformatics methodologies .	37
4.3.2	Creation of a consensus pipeline	40
4.3.3	Comparison of protein toxicity predictive strategies	42
5	Conclusions and recommendations	45
6	Bibliography	49
	Appendix A Presentation of retrieved tools to predict protein toxicity	56
	Appendix B Methodologies	97

Annexes (the annexes can be retrieved at this link:
<http://doi.org/10.5281/zenodo.13904499>)

Literature search results	Annex 1
Tools and methodologies applications	Annex 2
Optimized non-redundant datasets	Annex 3
BLAST results	Annex 4
InterProScan results	Annex 5
Main dataset	Annex 6
Train and test datasets	Annex 7
SVC train and test datasets	Annex 8
Pipeline train and test datasets	Annex 9
Consensus model results	Annex 10



1 Introduction

1.1 General background

In the European Union (EU) proteins are evaluated for their safety in various areas of the food and feed risk assessment, with examples spanning from novel proteins in Genetically Modified Organisms (GMOs) and novel foods, to proteins produced by microbial pesticides in the plant protection products area. Since proteins can be associated with toxic effects for humans and animals, dedicated tools and methodologies are deployed to the assessment of their toxicity; these are inherited and adapted from chemical risk assessment and include in vivo toxicological studies, as well as in silico investigations, such as similarity searches for toxins. In recent years, high-quality information on proteins has been made publicly available and can form the basis to evolve, modernize, and strengthen protein safety assessment embracing the regulatory science trend into New Approach Methodologies (NAMs). In a previous EFSA procurement (NP/EFSA/GMO/2018/01), an integrated pipeline for literature and database search on toxic proteins was developed. Toxins and “toxin-antitoxin systems” retrieved from UniProtKB, RCSB Protein Data Bank, SwissModel and the InterPro Consortium were used as inputs to the pipeline and compilations of comprehensive dataset of proteins with toxic effects set up; relevant scientific information on in silico methods for protein toxicity prediction were also identified, a preliminary evaluation of some of these tools was performed, and results were then discussed in terms of accuracy, specificity, and sensitivity; “TOXAPEX”, a Python-based tool, was also created to manage all data and information coming from this preliminary investigation (Palazzolo et al., 2020).

1.2 A note on in silico tools and methodologies to predict protein toxicity

To the best of our knowledge, two different in silico approaches could be used to predict the potential toxicity of proteins. Both approaches use protein alignment profiles obtained through BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1997; Neumann et al., 2014) on which Machine Learning (ML) based software tools are then trained.

The first approach specifically predicts whether a protein is toxic. It is based on tools that consider parameters found to be critical to distinguish between toxic and non-toxic peptides: amino acid composition (AAC), dipeptide composition (DPC), or pseudo amino acid composition (PseAAC). For instance, cysteine frequency is much higher in toxic peptides. After training, ML tools can also be used to detect distant homologous sequences (classification of proteins and/or domains) and align phylogenetically-related sequences (multiple alignment) (Fan et al., 2011; Gelly et al., 2004; Gupta et al., 2013; Jain and Kihara, 2019; Saha and Raghava, 2007a; Sharma et al., 2022).

The second approach predicts the general protein function. Gene Ontology (GO) terms (Ashburner et al., 2000) could be used to predict protein function. Results pertaining toxicity are singled out. This second approach could either corroborate or benchmark the results obtained with the more specific tools mentioned earlier.



1.3 Background and terms of reference as provided by the requestor

- To explore methodology/ies to predict the toxicity of a novel protein in the context of risk assessment. These should search for structural and functional homology of the protein of interest to well-known toxic proteins (homologous proteins share common structural architecture and function; detection of homology to toxins can be used to infer toxic properties in a protein); and for the presence in the protein itself of “toxic molecular signatures” (i.e. structural/functional properties relevant in the molecular initiating events leading to toxicity). Methodological approaches to consider could span from “traditional” sequence similarity analysis (e.g. BLAST) to artificial intelligence such as machine learning or deep learning. This objective should be based on the previous work by Palazzolo et al, 2020 and expand further by a comprehensive literature search. The literature search will be based on principles described in EFSA guidance on the application of systematic review methodology to food/feed safety assessments to support decision-making (EFSA 2010-). It will provide an overview (description and grouping) of methodology/ies suitable to predict the toxicity of a novel protein subject of risk assessment (EFSA-GMO-2021-02).
- To identify candidate methodology/ies for subsequent implementation the area of proteins newly expressed in GMOs (insecticidal proteins, enzymes conferring resistance to herbicides, proteins intervening on plants metabolic pathways) should be considered.
- To propose specific protocols describing the identified methodology/ies; threshold criteria should be identified. Examples of application of the identified protocols will be provided.
- To identify strengths, limits and gaps of the proposed methodology/ies, and to report these.

The need, frequency and modality of tool updates should also be outlined.

The methodology/ies proposed could be a pipeline (i.e. a concatenation of existing tools), a new individual tool coded “de novo” or as an update of an existing one, or a combination of the two. The coding of the methodology/ies (pipeline or tool) is not part of the present call (EFSA-GMO-2021-02). The solution proposed should consider the specific needs of EFSA (risk assessment of food and feed), as for example the possible preference for sensitivity over specificity or the paramount importance of the thresholds, and it should be not just one of the viable options but the best viable option. The identified methodology/ies can constitute preparatory work for the future development of a pipeline, architecture and software. The development of a pipeline is not part of the current call.

This contract/grant was awarded by EFSA to:

Contractor: Università degli Studi di Milano

Contract title: Development of in silico methodologies to predict the toxicity of novel proteins in the context of food and feed risk assessment

Contract number: OC/EFSA/GMO/2021/02 – LOT1

1.4 Interpretation of the Terms of Reference

To fulfil all the goals, five different tasks were defined by EFSA (Table 1).

Table 1: Overview of Project Tasks

Task
Task 1 - Project plan and gathering protocols
Task 2 - Literature Search
Task 3 - Protein benchmark dataset
Task 4 - Test of selected tools/pipelines-
Task 5 - Final report

Tasks were discussed and agreed with EFSA, with further interpretation for task 2 and 4.

Task 2 (literature search) aims to gather exhaustive and comprehensive information about available tools and methodologies able to predict whether a protein could be classified as toxin; this is achieved via:

- methods based on primary sequence, supporting toxicity prediction and protein function prediction with GO "toxicity"
- methods based on 3D structure, supporting toxicity prediction.

The search includes peer-reviewed research publications, and publicly available databases, together with a detailed documentation of the search strategy, also according to the 'Guidance on Application of systematic review methodology to food and feed assessments to support decision making' (EFSA, 2010).

Task 4 (i.e. defining the optimal approach(es) for predicting protein toxicity, if applicable), identifying shortcomings and limitations, suggesting methodological advancements, and providing a risk assessment solution tailored to the purpose), is addressed via a multi-step strategy:

- in first instance an overall evaluation of the identified tools and methodologies to predict protein toxicity is conducted and their applicability strengths, limits, and gaps evaluated. Namely, The Accuracy, Specificity and Sensitivity will be assessed using specific datasets (benchmarks, i.e. True Positives, True Negatives and Expected False Negatives). Information on tools and methodologies and their relevance in the field of in silico toxicity prediction are also evaluated and discussed using thresholds criteria for protein toxicity prediction;
- subsequently a pipeline is proposed to predict if a newly identified protein could be associated to toxic activity. In this framework, all the relevant in silico methodologies identified will be used, to deliver to EFSA an integrated pipeline

and/or new individual tools able to predict protein toxicity basing on primary structure (i.e. sequence), secondary structure (i.e. folding) and tertiary structure (i.e. 3D structure).

1.5 Additional information - Project plan

The project plan was presented, discussed and agreed with EFSA at the kickoff meeting. In line with the ToRs and their interpretation, it includes:

- Definition of gathering protocols (literature, protein datasets) (Task 1)
- Database settings (Task 2 and Task 3)
 - literature search to identify tools, methodologies to predict protein toxicity and literature database setting
 - protein datasets setting
- Assessment of tools and methodologies and pipeline definition (Task 4)
 - Presentation of tools and methodologies to predict protein toxicity
 - Discussion and decision about tools and methodologies worth further investigation
 - Testing tools against protein datasets
 - Testing bioinformatics methodologies
 - Creation of a consensus model and pipeline

2 TASK 1: Protocols to gather information on tools and methodologies predicting protein toxicity and to set protein benchmark datasets

2.1 Protocol for literature search to identify tools and methodologies

The protocol for this literature search was developed following (EFSA, 2010) and organized in two sections:

- Review question and inclusion/exclusion criteria
- Methods for selecting the most relevant literature and to organize the results of the literature search

Review question and inclusion/exclusion criteria

This section includes a clear definition of the question and objective of the review and pre-definition of criteria for study inclusion or exclusion.

Review question: Which are the currently available tools and methodologies to predict the potential toxin activity of proteins and how are they applied to classify proteins as toxins or not, with particular attention in the context of food and feed.

Due to the descriptive nature of the review question, Population (P) and Outcome (O) key elements are defined as follows:

Population : Ensemble of identified literature about a specific tool/methodology

www.efsa.europa.eu/publications



Outcome: The tool/methodology is useful to predict if a protein can be classified as toxin or not (Yes/No outcome)

Inclusion/exclusion criteria:

Specific inclusion and exclusion criteria have been used to select the appropriate literature. In Table 2: the four identified criteria (time, language, publication type and publication content) are reported together with the inclusion/exclusion conditions.

Table 2: Inclusion and exclusion criteria. In italics: publication type definitions in PubMed.

	IN	OUT
Time	2012-2022	Before 2012
Language	English	Other than English
Publication Type	Primary research, review and application articles published on peer-reviewed journals; <i>systematic review; book chapters</i> **	Primary research, review and application articles only in pre-print format and not peer-reviewed (i.e., Bio-/Med-Rxiv); PhD thesis; master thesis; multicenter study; observational study; retracted publication; case reports; controlled clinical trial; clinical trial protocol; clinical trial; randomized controlled trial; equivalence trial; historical article; video-audio media; practice guideline; letter; webcast; congress; editorial; editorial research support, non-U.S. Gov't;
Publication content	Primary research, review and application articles about in silico tools for discriminating toxins from non-toxins; literature associated to predictive tools previously analysed (NTXpred, BTXpred, KNOTTIN database, ClanTox, ConoServer) (ref. Palazzolo 2020): literature associated to methodologies (BLAST; Protein modelling; Hidden Markov Model; Support Vector	Literature associated to deprecated tools; literature associated to predictive tools for small molecules.

Machine; Machine Learning; Artificial intelligence)

** Only if abstracts are available.

Methods for the literature search

The literature search addressing Task 2 has been planned in three subsequent steps:

- Step 1. Search Strings Definition combining specific keywords using AND/OR operators. Keywords include terms relevant to protein toxicity prediction tools and in silico methodologies.
- Step 2. Literature Retrieval using databases like PubMed, Web of Science, and Scopus. Google's search engine and the CAFA challenge were also utilized to identify predictive tools and gather their primary literature.
- Step 3. Database Setting to organize the selected publications into a database using Mendeley (<https://www.mendeley.com/reference-management/mendeley-cite>) with detailed bibliographic information, and reasons for inclusion or exclusion, alongside a table (Table 3) summarizing the tools or methodologies and their associated literature.

2.2 Protocol to set protein datasets

Subsets of proteins (datasets) suitable for evaluating the sensitivity and specificity of selected tools and methodologies were set according to the rationale and methodology below described.

Rationale

Well-defined toxic and non-toxic protein datasets are necessary to test the identified tools, methodologies, and proposed strategies. Each predictive tool may have its own application scope; for instance, some tools are designed to receive only short peptide sequences, while others are specific for certain classes of toxins or organisms. Thus, datasets for assessing a receiver operating characteristic (ROC) curve for each tool should be constructed considering the tool's specific application scope.

Methodology

For each tool and methodology to be evaluated, a dataset composed of an equal proportion of toxic and non-toxic proteins is populated according to the tool and methodologies' application scope. Tools with the same applicability scope have been tested using the same datasets. Toxic and non-toxic datasets are specifically defined as True positives (TPs) and True negatives (TNs) as below detailed.

True Positives (TPs): True Positives represent well-known annotated and expert-reviewed toxins selected from the UniProtKB database. The selection criteria are defined on a per-tool basis, according to the application scope of each tool and methodology. Each entry is an annotated toxin associated with information such as GO annotations, PFAM classification, and PDB protein structure. A programmatic approach is used to gather entries from the UniProtKB database according to the tool specifications and entry data. Care is taken in removing redundancy within the dataset with pairwise sequence alignment methodology described below (see §3.3.4).

True Negatives (TNs): A reference database of non-toxic proteins is constructed using the same infrastructure as for the TPs. These proteins are selected from the UniProtKB



database, considering only highly reviewed and annotated entries without any keywords associated with toxicity. TNs are filtered on a per-tool basis, and the final tool-specific TPs dataset is used to select a corresponding number of non-toxic entries. The selection aims to match the sequence length distribution and other relevant features of the TPs. This process is carried out to ensure the protein sets are similar in feature distribution but guaranteed to be non-toxic, allowing better estimation of tool efficiency.

Expected False Negatives (EFNs): For each tool, a third dataset is constructed, including just the toxic proteins that fell outside the tool's application scope, and thus are *expected* to be misclassified (*false negatives*). This procedure helps assess the tool's application scope beyond what was originally declared.

3 TASK 2 and TASK 3: Gathering information on tools and methodologies predicting protein toxicity and setting protein benchmark datasets

3.1 Literature search: Tools and methodologies predicting protein toxicity (TASK 2)

3.1.1 Introduction

Proteins may have toxin-like effects, with a variety of mechanisms and in a variety of settings. In our previous publication (Palazzolo et al., 2020), as the outcome of an EFSA procurement (NP/EFSA/GMO/2018/01), we reviewed relevant scientific information on *in silico* prediction methods for toxins for supporting the food and feed risk assessment. In the present EFSA procurement, the main goal was to develop methodology(ies) to predict the potential toxin properties of new proteins expressed in GMOs, of possible application on proteins in other food/feed assessment areas. To this aim, as a first step a thorough and comprehensive review of existing literature was carried out to gather and critically evaluate pertinent information concerning methodologies for: (1) determining whether a protein could be categorized as a toxin, or more broadly, (2) establishing the relationship between protein structure and function, particularly regarding toxic activity. Furthermore, a meticulous examination of the retrieved literature in this domain was conducted to assess the current state of the art, including the availability, applicability, and suitability of existing tools and the information organized.

3.1.2 Methodology

The planned literature search was carried out according to the Literature Search Protocol in three subsequent steps.

Step 1 - Search Strings Definition (Table 3).

To define the string(s) for the **literature search of in silico tools**, a preliminary search was performed in PubMed, Web of Science and Scopus with several keywords, alone and in combination, based on four semantic concepts:

- i. the relevance of the study in the context of proteins (protein/peptide),
- ii. the toxic activity of proteins (toxic),



- iii. the prediction of the toxic activity (prediction/*in silico* /computational),
- iv. the availability of a specific tool to perform the prediction (tool/software/application/program/server).

Only six search strings with all the four semantic concepts combined with the AND logical operator were considered of interest. It must be noted that some terms, such as “in silico” or “in-silico” and “predictive” or “prediction”, are automatically searched *via* a combination of Medical Subject Headings (MeSH) synonyms and they gave the same number of results (Annex 1), making us confident that the search carried out with the proposed strings also include all the MeSH synonyms of all the keywords relevant to this activity.

The keywords “function”, “mode of action”, “mechanism of action”, “MOA”, and “activity” were added upon a specific request made by EFSA during the kick-off meeting. As it can be seen in Annex 1, due to the large semantic areas covered by these two keywords, the inclusion of either “function” or “activity” or both leads to values that cannot be properly processed (67,346; 25,077; and 70,606 hits, respectively). Therefore, we included only “mechanism of action” OR “moa” OR “mode of action” to the final search string since these terms are considered more specific with respect to the toxin activity and so to the review question.

Accordingly, the final search string reported in Table 5 matches both the relevance for, and the focus on the review question and generates a manageable number of hits.

All strings (including those with single keywords) are reported in Annex 1.

To define the strings for the **literature search of *in silico* methodologies** a specific search was performed to evaluate the application of some *in silico* methodologies in the context of protein toxicity prediction. The search was carried out combining a single methodology, such as “BLAST”, “protein mode(l)ling”, “support vector machine”, “machine learning”, “hidden Markov model”, “artificial intelligence” with the string “protein AND toxin AND prediction” (Table 6). With this strategy, we skimmed all the articles not specifically related to the application to the protein toxicity field of the above-mentioned methodology (Annex 1). For both the searches, the retrieval of publications (both primary search and review) was performed using publicly available databases, such as ‘Web of Science’, ‘Scopus’ and ‘PubMed’, providing information on the review question. These databases are considered by the scientific community comprehensive sources of information. Predictive *in silico* tools were searched also by using public search engine tools and within the CAFA challenge.

Table 3: Search strings.

Activity	Search strings	Number of hits
Literature search of toxin prediction tools	english[Language] AND "last 10 years"[dp] AND ("mechanism of action" OR moa OR "mode of action" OR toxic) AND (protein OR peptide) AND (prediction OR "in silico" OR in-silico OR computational OR predictive) AND (tool OR software OR application OR program OR server)	4,245
	<ul style="list-style-type: none"> • english[Language] AND "last 10 years"[dp] AND ("protein 	8



Literature search of <i>in silico</i> methodologies	modeling" OR "protein modelling") AND protein AND ("mode of action" OR moa OR toxin) AND prediction	
	<ul style="list-style-type: none"> • english[Language] AND "last 10 years"[dp] AND ("support vector machine" OR svm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	25
	<ul style="list-style-type: none"> • english[Language] AND "last 10 years"[dp] AND ("machine learning" OR ml) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	65
	<ul style="list-style-type: none"> • english[Language] AND "last 10 years"[dp] AND ("hidden markov model" OR hmm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	6
	<ul style="list-style-type: none"> • english[Language] AND "last 10 years"[dp] AND "artificial intelligence" AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	571

Step 2 - Literature Retrieval. The second step was carried out in the same way for both the literature searches of tools and *in silico* methodologies. It consisted in the selection of relevant publications, according to the specific eligible inclusion and exclusion criteria reported in Table 3 and by reading the title and the abstract. In the case a study was considered not clearly convincing and relevant for the scope, the full paper was read and further analysed to avoid the exclusion of relevant information. Thus, these criteria helped us define a database of papers to be further inspected.

Step 3 - Database setting. All the selected references were stored in a dedicated library, using the reference management software Mendeley (<https://www.mendeley.com/reference-management/mendeley-cite>). The outcomes of this systematic search were documented by listing authors, article titles, journal names, dates, volumes, issues, and full text abstracts or using standard identifiers for the available proteins databases together with a statement explaining the reason to include/exclude the publication in/from the final database. Moreover, a table including the name of the tool or methodology, the associated primary literature with the corresponding digital object identifiers (DOI), the prediction method and the number of available application studies was created (Annex 2).



The literature retrieved was integrated with the information from curated databases for protein classification in family/domain/motif, carefully verifying their applicability in the literature associated to predictive ability of both the InterPro member databases (it is composed by 13 members: <https://www.ebi.ac.uk/interpro/>) and the Meme suite was carefully evaluated.

We also gathered relevant publications on the tools that applied to the CAFA challenge and ranked within the first positions, by selecting published researches available as references from the CAFA website. All the selected references were stored in a dedicated library, using the reference management software Mendeley (<https://www.mendeley.com/reference-management/mendeley-cite>).

3.1.3 Results

The outcome of the literature search for in silico prediction tools is summarized in Table 4.

Table 4: Literature search of toxin prediction tools: results.

Activity	Search strings	Number of hits
Literature search of in silico prediction tools	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND ("mechanism of action" OR moa OR "mode of action" OR toxic) AND (protein OR peptide) AND (prediction OR "in silico" OR in-silico OR computational OR predictive) AND (tool OR software OR application OR program OR server) 	4,245

The outcome of the literature search for in silico prediction methodologies is summarized in Table 5.

Table 5: Search strings.

Activity	Search strings	Number of hits
Literature search of in silico methodologies	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND "artificial intelligence" AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	571
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND ("support vector machine" OR svm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	65
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND ("machine 	



	learning" OR ml) AND protein AND ("mode of action" OR moa OR toxin) AND prediction	
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND ("support vector machine" OR svm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	25
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND ("machine learning" OR ml) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	8
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND "hidden markov model" OR hmm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	6
	<ul style="list-style-type: none"> english[Language] AND "last 10 years"[dp] AND "artificial intelligence" AND protein AND ("mode of action" OR moa OR toxin) AND prediction 	

The "Number of hits" column represents the quantity of search results retrieved for each specific search string. Rearranging the data in descending order of the number of hits highlights the significance of each search string, with the most productive search string ("english[Language] AND "last 10 years"[dp] AND "artificial intelligence" AND protein AND ("mode of action" OR moa OR toxin) AND prediction") appearing at the top.

Since some articles were found through string search in PubMed, Scopus and Web of Science, we filtered results to build a non-redundant table with 7831 entries (original articles, reviews and book chapters) (Annex 2). This table was carefully skimmed, associating to each entry a justification for admission/exclusion. We used mainly four criteria to accept or exclude an article:

- 1) The main focus of the article is about a tool or an *in silico* methodology applied to toxin prediction (include);
- 2) The article may report a tool or an *in silico* methodology applied to toxin prediction (include);
- 3) The article exclusively refers to small molecules' toxicity (exclude);
- 4) The article is completely off-topic (exclude).

According to these criteria, we included 106 research articles or book chapters. All of them were then carefully evaluated and classified according to the following definitions:



- a) Articles describing the tool (include as primary source);
- b) Articles describing the methodology (include as primary source);
- c) Articles in which Authors used a tool, a methodology and/or domain classification (include as application);
- d) Articles that didn't fit our goal (exclude).

The selected articles are reported with the primary reference, the abstract, the link to the website of the tool, the number of citations, the field of application, and some information regarding the training dataset and the method used for the prediction. Globally, the collected literature provides useful information on ten *in silico* prediction tools. All the *in silico* prediction predictive tools are based on *in silico* methodologies, such as SVM, NN, HMM and/or family/domains/motifs classifier, such as MEME, MAST and TOMTOM (Table 6).

Among the selected articles, some articles were flagged as "out of time" since they were published before 2012 but modified later, and for this reason included in the result of the search. Among these papers, those considered relevant for the scope of this activity, even if out of the pre-defined time range, were reported in Annex 2. Then the ranking criteria previously defined were applied to guide the evaluation of the methods/tools described in the primary sources. Accordingly, Table 6 summarizes all the results, while in the next pages each retrieved tool is presented (Primary citation, Abstract, Link, Citations, Field of Application, Training Dataset, Predicting Methods and other information, as relevant). A complete description of the tools is reported in Appendix A.

Table 6: Summary of selected tools, methodologies and domains/families/motifs. For each tool, the methodologies behind the architecture have been highlighted together with the release year. NTXPred, ToxinPred, ToxDL and ToxinPred2 are also based on sub-routines that take into account domains and motifs classification/prediction.

Tools	BTXPred	NTXPred	PredCSF	ToxinPred	ToxClassifier	NNTox	TOXIFY	ToxDL	ToxIBTL	ToxinPred2
Methods	PSI-BLAST	X	X	X		X				X
	PSIPRED	X		X						X
	HMMER	X				X				X
	CLUSTAL-W	X								X
	SVM	X	X	X	X	X				X
Domains	NN		X		X	X	X	X	X	X
	InterPro							X		
	MEME		X		X			X		
Motifs	LOGO			X						
	MAST		X							
	MERCI									X
Year	2007	2007	2011	2013	2016	2019	2019	2021	2022	2022

The search identified two studies reporting the direct application of two methodologies to the prediction of toxicity of proteins (SVM and HMM). Other methodologies are embedded in prediction tool/s, but not used individually to predict toxicity. See details in Table 6 and in Appendix B.

3.2 Selection of tools and methodologies for an in silico prediction strategy of protein toxicity

The identified tools and methodologies were ranked according to the criteria discussed and agreed with EFSA. This was needed to select the most appropriate tools/methodologies for setting a strategy for the *in silico* prediction of protein toxicity (Task 4). Tables 7 - 9 report ranking criteria to guide the tools/methodologies evaluation.

Table 7: Ranking criteria for tools

Ranking Criteria	Satisfying conditions	Scoring
Usability	At least one application OR at least one not-self citation of primary publication	10 points for each satisfying condition
Availability	Open source (web server or download) OR availability of documentation about tool (tutorial)	10 points for each satisfying condition
Robustness	Compatibility with updated operating systems (last versions of MacOS, Linux, Windows) OR with updated programming languages	8 points for each satisfying condition

Table 8: Ranking criteria for tools - modified and adopted during the evaluation

Ranking Criteria	Satisfying conditions	Scoring
Usability	At least one application OR at least one not-self citation of primary publication	10 points
Availability – Web Server (WS)	Open source (web server)	4 points
Availability – API (WS)	Application Program Interface	4 points
Availability – Open Source (OS)	Open source (downloadable source code)	8 points
Robustness	Compatibility with updated operating systems (last versions of MacOS, Linux, Windows) OR with updated programming languages	8 points
Out of time (OT)	Coded > 2012 OR updated >2012	10 points

Table 9: Ranking criteria for methodologies modified and adopted during the evaluation

Ranking Criteria	Satisfying conditions	Scoring
------------------	-----------------------	---------



Usability	At least one application alone OR in combination with other methodologies	10 points
Availability	Pre-compiled packages with the methodology of interest	Yes/No condition
Robustness	Methodology used for <i>in silico</i> prediction	Yes/No condition

3.2.1 Results

The outcome of the evaluation of tools and methodologies based on the presented criteria, is reported in Table 10 and Table 11. The following tools were admitted to the next activities: TOXIFY, ToxClassifier, NNTox, ToxDL, ToxIBTL, ToxinPred2. We discarded from selection BTXPred, NTXPred and PredCSF since all of them are out of time with respect to reference and their source code is not available. The original version of ToxinPred is considered superseded by Toxinpred2.

On the other hand, all the methodologies retrieved by literature (SVM and HMM) together with those embedded in prediction tools (Table 6) were admitted to the next activities.

Table 10: Tools evaluation

Tool	Usability	Avail WS	Avail API	Avail OS	Robustness	OT	Total
BTXPred	10	4	0	0	0	0	14
NTXPred	10	4	0	0	0	0	14
PredCSF	10	4	0	0	4	0	18
ToxinPred	10	4	0	0	0	10	24
ToxinPred2	10*	4	0	8	8	10	40
TOXIFY	10	0	0	8	8	10	36
ToxClassifier	10	0	0	8	8	10	36
NNTox	10	4	0	8	8	10	40
ToxDL	10	4	0	8	8	10	40
ToxIBTL	10	4	0	8	8	10	40

*Scoring based on previous version usability

Table 11: Methodologies evaluation

Methodology	Usability	Availability	Robustness
-------------	-----------	--------------	------------

SVM	10	Yes	Yes
BLAST	10	Yes	Yes
HMM	10	Yes	Yes
NN	10	Yes	Yes
PSIPRED	10	Yes	Yes
CLUSTAL-W	10	Yes	Yes

BLAST has been directly tested since it is at the basis of all the selected tools; HMM, PSIPRED and CLUSTAL have been tested within the pipeline. SVM/AI were used as an ensemble methodology, not directly applied. Therefore, it was not tested as such but in the workflow management to improve the accuracy of the pipeline.

3.3 Protein datasets setting (TASK 3)

3.3.1 Introduction and Aim

The objective of this task was to generate specific datasets (Annex 3), according to the outline proposed in §2.2, to be used to assess each tool performance. The test sets will include entries labelled as follows:

- **True positives (TP):** known toxic proteins, falling within the application scope of the tool.
- **True negatives (TN):** known non-toxic proteins, sharing high sequence similarity to the corresponding TPs.
- **Expected false negatives (EFN):** where applicable, known toxins that are expected not to be correctly classified by the tool, as they fall outside of the tool application scope.

Some tools are only able to operate on specific classes of toxins (e.g., animal venoms, conotoxins), while other tools are more general. Thus, we created multiple testing datasets specific for each tool application scope.

3.3.2 Evaluation of the application scope of the selected tools

To determine whether a tool can correctly predict toxicity of a given protein, we carefully reviewed the literature associated to the predictive tools selected in TASK 2 to identify their application scopes. Beyond the functional limitations (i.e the application scope as defined for the tool, see Table 6), technical limitations must be considered. Indeed, some tools can handle protein sequences in a specific range of residue numbers. Thus, for these tools, entries in the final datasets contains only sequences with an appropriate length suitable for the specific tools (Table 12).

Table 12: Application scopes of the selected tools and technical limitations.

Tool	Application Scope	Technical limitations
------	-------------------	-----------------------



ToxClassifier	Animal venoms	N.A.
NNTox	All toxins	N.A.
TOXIFY	Venom toxins	Sequences \leq 500 amino acids*
ToxDL	All toxins	N.A.
ToxIBTL	All toxins	N.A.
ToxinPred2	All toxins	N.A.
KNOTTIN	Toxins belonging to the knottins family	Sequences \leq 200 amino acids

*While not strictly a technical limitation, only proteins containing 500 amino acids were included in the final training dataset for this tool. Application scopes were inferred by the literature, while technical limitations have been derived after visiting the tools websites.

3.3.3 Main toxic protein datasets creation

All datasets contain entries retrieved from UniProtKB, using the keywords reported in Table 5. The search criteria for these datasets were defined similarly to how tools' authors retrieved the entries for the tool training and test; however, we decided to be more rigorous with the choice of toxins, making sure that all entries were manually reviewed by UniProt experts (reviewed:true keyword) and contained the GO "toxin activity", as extensively described in Palazzolo et al. (2020).

Based on the tool's application scope, three TP datasets were used to assess the tools performance (Table 13):

- **alltox**: All the reviewed UniProt entries matching the GO "toxin activity".
- **venom**: All the reviewed UniProt entries matching the GO "toxin activity" and containing the general keyword "venom" in every field.
- **knottin**: All the reviewed UniProt entries matching the GO "toxin activity" and the UniProt keyword "knottin".

Note that venom and knottin are subsets of alltox since both "venom" and "knottin" are UniProt second order keywords. All the venom- and knottin-flagged protein have "toxin activity" as main GO (Palazzolo et al, 2020).

The following True Negatives (TN) dataset was also defined:

- **allnontox**: All the reviewed UniProt entries that *do not* match the GO "toxin activity". This criterion should effectively select all reviewed (known and manually annotated) proteins that are not toxins nor have any toxic activity.

Table 13: Main datasets (True Positives) and associated query.

Dataset	query	#Entries
alltox	((go:0090729*) AND (reviewed+:true))	7411
venom	((go:0090729) AND (reviewed:true)) AND venom	6265

knottin	((go:0090729) AND (reviewed:true)) AND (keyword:KW-0960**)	1292
allnontox	(NOT (go:0090729)) AND (reviewed:true)	560591

*GO "Toxin activity"

**UniProt Keyword "knottin"

† Manual annotation and revision by experts

Each entry in the dataset contains the following fields: *UniProt identifier*, *GO – molecular function*, *Organism*, *Pfam*, *Motifs*, *Sequence*, *Sequence length*. These fields were chosen to provide data useful to classify proteins and toxins based on relevant information, such as molecular function, Pfam protein family and the presence of specific functional motifs. A list of all available fields that can be returned by UniProt REST API can be found here: https://www.uniprot.org/help/return_fields. We used an in-house Python program to download the datasets leveraging the UniProt REST API. Data were then manipulated using the Pandas Python library. The Python program is delivered to EFSA and can be available upon request.

3.3.4 Lowering the redundancy within datasets

3.3.4.1 Problem definition

The TP datasets described above suffer from a sampling bias due to several factors, including the research interest on specific organisms (e.g., animals) or specific classes of toxins, such as scorpion venom toxins. This results in the presence of relatively large groups of similar toxins, which belong to common families, that share very similar sequences. Instead, other toxins, which may be of new discovery, or did not gain enough interest, or simply are "unique" in the sense that they have no homologs in other organisms, have very little representativity. Hence, failure to balance the frequency between overrepresented and underrepresented protein sequences poses the potential to introduce bias into the assessment of tool accuracy. This bias may manifest by overemphasizing the tool's proficiency in recognizing oversampled toxins (such as scorpion venom toxins), which exhibit substantial similarity, and simultaneously underemphasizing its performance in relation to under-sampled toxins characterized by distinctive sequences. It is imperative to recognize that the prevalence of specific toxins within the UniProt database does not serve as an indicator of their prevalence in the natural environment. For instance: a dataset comprising animal venom toxins was acquired from UniProt using a suitable search query targeting all annotated animal venom toxins. A statistical examination of the dataset reveals that 90% of its entries pertain to scorpion toxins, while the remaining 10% encompasses diverse venom toxins originating from other species. Subsequently, a specialized tool for venom toxin recognition is assessed against this dataset, demonstrating a 95% sensitivity score, accurately classifying 95% of the tested toxins. However, a scrutiny of misclassified results elucidates that all discrepancies are confined to the 10% subset representing non-scorpion venom toxins. Consequently, the tool exhibited 100% accuracy in identifying scorpion toxins, characterized by their uniformity, but exhibited suboptimal performance (50%) in the identification of highly diverse sequences from other species. This outcome is deemed unsatisfactory, considering the tool's advertised utility in classifying all venom toxins.



3.3.4.2 Solution strategy

The identified solution was to use hierarchical clustering based on sequence similarity, and merge clusters based on a similarity threshold. Only a representative item per each cluster, i.e. the medoid, and the corresponding representative sequence has been used for each group. Further details are given in the following paragraph.

3.3.4.3 Implementation

An all-against-all pairwise sequence alignment of the toxin sequences contained in **alltox** has been performed using the Needleman-Wunsch algorithm (EMBOSS Needle) with BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) and -10/0.5 gap opening/extension penalties. These parameters are at their default values and are suitable for comparing protein sequences within a broad range of evolutionary distances. This resulted in a symmetric square matrix reporting the sequence similarity for all toxin pairs as computed by the alignment algorithm. The similarity score is defined as the number of positive pairings (identities and conservative mutations) between the aligned sequences, divided by the length of the alignment, hence it is represented by a number ranging between 0 (no pairings) and 1 (identical sequences). A *distance matrix* was then generated by taking the reciprocal of the similarity matrix ($1 - \text{similarity}$). Because **alltox** is a superset of **venom** and **knottin**, the generated "all-vs-all" distance matrix can be used to cluster all datasets. Each dataset was clustered hierarchically using a maximum-distance linkage method (Figure 1) thus ensuring that inter-cluster distance corresponds to the maximum distance between all members of each cluster. E.g.: $\text{distance}(A, B) = \max(\text{distance}(a, b) \text{ for } a \text{ in } A \text{ for } b \text{ in } B)$, where A and B are any two clusters and a and b are their elements.

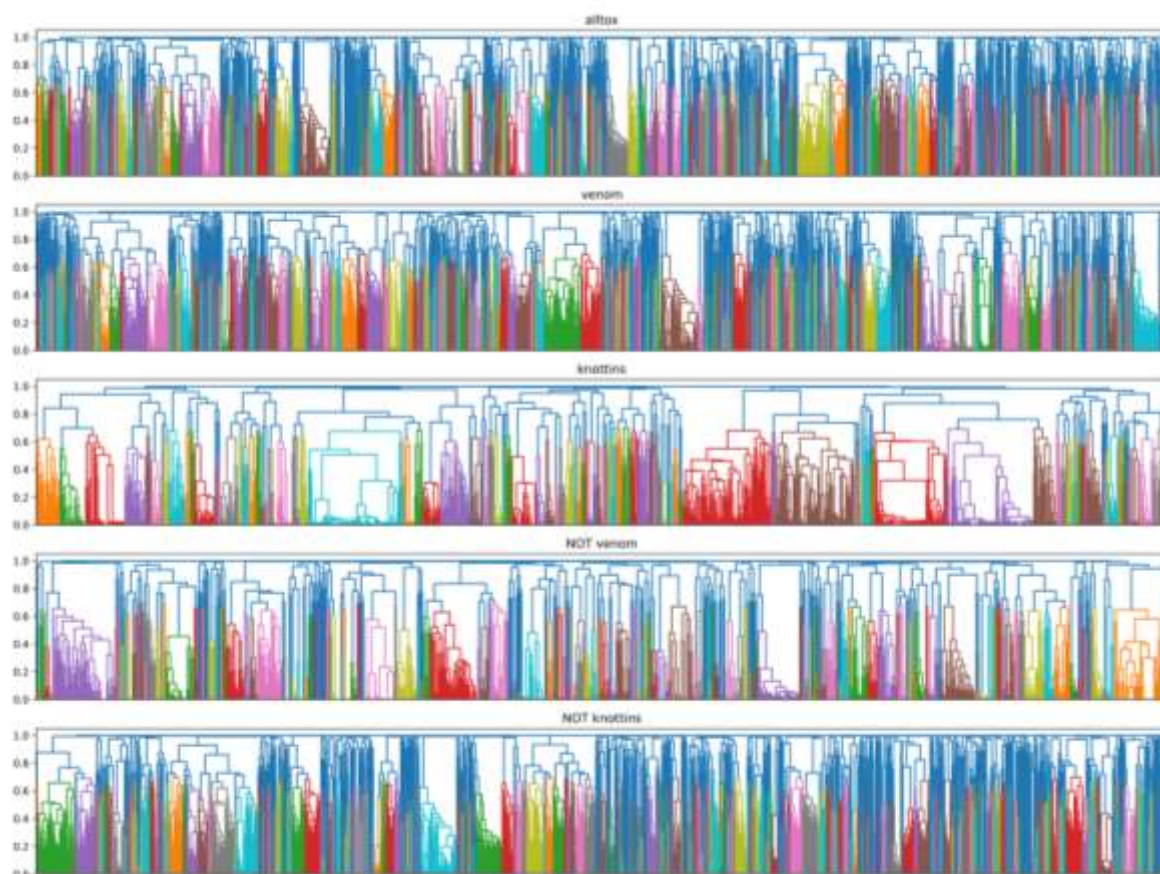


Figure 1: Dendrogram representation of datasets clustering. Colours indicate optimal clustering, only for visualization purposes. The datasets “NOT venom” and “NOT knottins” are obtained by subtracting entries contained in “venom” and “knottin” from “alltox”, respectively.

Clusters were then flattened (merged) so that the minimum distance between any two clusters is 0.2, i.e., the maximum similarity between elements picked from any two clusters is lower than 80% (Figure 2). This threshold was chosen as a rule of thumb. Then, for each dataset, a new non-redundant dataset was generated by retaining only the representative sequence (i.e. the medoid) per each of the identified clusters.

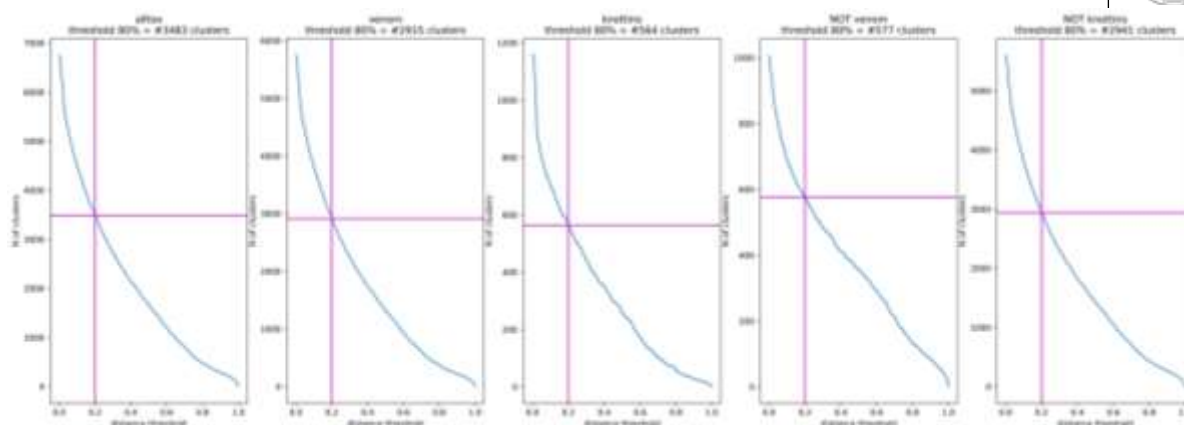


Figure 2: Clusters agglomeration based on distance threshold. The figure reports the final number of clusters after merging the clusters at the corresponding distance threshold.

3.3.4.4 Results

Three optimized non-redundant datasets were obtained, **alltox80**, **venom80** and **knottin80**, where all elements within each dataset share no more than 80% sequence similarity (Annex 3)

3.3.5 Generation of the True Negatives datasets

Given the above optimized datasets **alltox80**, **venom80** and **knottin80**, three corresponding datasets of non-toxic proteins were set; ideally, to provide a greater challenge for the tools, non-toxic proteins (TN) should have sequences as similar as possible to those from the corresponding toxin (TP) dataset. This aimed at testing the tool ability to properly classify sequences based on the *subtle* differences between toxins and non-toxins. Consider the example where a model is trained to recognize cats within pictures: to properly test the model one should include pictures of other animals, instead of pictures of cars, so to make sure that the model is focusing on the right features that distinguish a cat from other animals (causation), instead of focusing on incidental corollary features (correlation).

3.3.5.1 Implementation

To construct TN datasets that contain sequences similar to the corresponding **alltox80**, **venom80** and **knottin80** (TP) datasets, we implemented the following procedure. For each toxic protein within a main TP dataset, we identify the most similar non-toxin by aligning the toxin to all the non-toxic proteins of similar length contained in **allnontox**. Pairwise alignments were performed with the same methodology used to generate the similarity matrices (EMBOSS Needle with default settings: BLOSUM62 substitution matrix, -10 gap opening penalty and -0.5 gap extension penalty). Hence, for each toxin, we add the most similar non-toxin to the growing TN dataset; if the non-toxin was already contained in the TN dataset (this might happen when two or more toxins best align with the same non-toxin), the next top-scoring non-toxic protein is chosen, until a unique non-toxin is found. We chose to align each toxin to a subset of non-toxins with a similar length

www.efsa.europa.eu/publications

(± 5 amino acids) for performance reasons, following the assumptions that globally similar proteins also have similar sequence lengths. Otherwise, the total number of alignments to perform (all toxins vs. all non-toxins) would have been prohibitive (in the order of magnitude of 10).

3.3.5.2 Results

Three datasets were obtained, **alltox_TN**, **venom_TN** and **knottin_TN** containing known non-toxic proteins similar to toxic proteins in the sets **alltox80**, **venom80** and **knottin80**, respectively (Annex 3).

3.3.6 Generation of the Expected False Negatives (EFN) datasets

We designate Expected False Negatives (EFNs) for a particular tool those toxins anticipated to be inaccurately classified since outside the tool's scope. We consider that assessing the robustness of a tool's scope boundaries, as indicated by its developers, can provide valuable insights. For instance, a tool trained solely to identify venom toxins might struggle with accuracy when presented with non-venom toxins. However, if the tool demonstrates high accuracy with the EFN set, analyzing correctly classified sequences within this set can offer valuable insights on the tool real scope. The application scope of each tool and the consequent definition of out-of-scope toxins is hereby summarized in Table 14 Note that for the tools with the broader application scope (i.e., all toxins), the "out-of-scope" toxins cannot be defined.

Table 14: Tools application scope and their expected out-of-scope toxins.

Tool	Application Scope	Out-of-scope
ToxClassifier	Animal venoms	All toxins that are not animal venoms
NNTox	All toxins	N.A.
TOXIFY	Venom toxins	All toxins that are not animal venoms
ToxDL	All toxins	N.A.
ToxIBTL	All toxins	N.A.
ToxinPred2	All toxins	N.A.
KNOTTIN	Toxins belonging to the knottins family	All toxins that are not knottins

3.3.6.1 Implementation

EFNs datasets were generated by subtracting the entries included either in **venom** or **knottin** from **alltox**. Namely, **alltox \ venom** resulted in the dataset **alltox_not_venom** containing all toxins *minus* the toxins also belonging to the **venom** dataset; the same applies for **alltox \ knottin**, which resulted in the dataset **alltox_not_knottin**. Please refer to Figure 3 for additional information on datasets composition. The two newly generated sets were then clustered by similarity using the same procedure previously described.

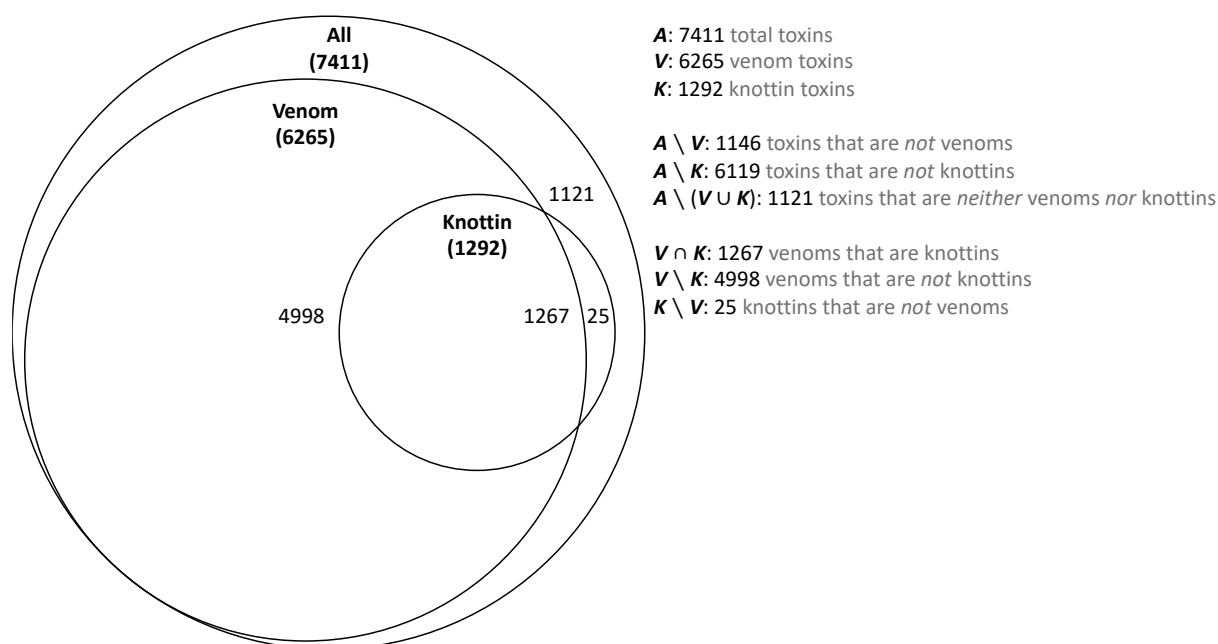


Figure 3: Venn diagram describing the alltox dataset composition (All). Numbers refer to “raw” datasets, before clustering. For clarity, circle areas are not to scale although they correlate with sets numerosity.

3.3.6.2 Results

Two datasets were obtained, **alltox_not_venom80** and **alltox_not_knottin80**. These low-redundancy datasets contain expected false negatives for the tools that are only able to recognize venoms or knottins respectively (Annex 3).

3.3.7 Generation of final datasets for each tool

3.3.7.1 Dataset filtering

Datasets were filtered to only include entries of appropriate lengths for tools with a limit to the maximum and/or minimum number of residues that can be contained in a query sequence. Since similar proteins have similar lengths, there’s no need for regenerating TNs based on sequence similarity. In this fashion the datasets **venom80_500**, **venom_TN_500**, **alltox_not_venom80_500**, **knottin80_200**, **knottin_TN_200**, and **alltox_not_knottin80_200** are easily obtained. Please refer to Table 15 for a summarized description of dataset names.

Table 15: Summary of dataset names and their description, organized by True Positives (TP), True Negatives (TN) and Expected False Negatives (EFN).

Dataset name	Description	Size
TP		



alltox80	All toxins, below 80% similarity	3,483 out of 7,411
venom80	Venom toxins, below80% similarity	2,979 out of 6,265
venom80_500	Venom toxins, below80% similarity, with maximum length of 500 amino acids.	2,939 out of 6,265
knottin80	Knottins, below 80% similarity	570 out of 1,292
knottin80_200	Knottins, below80% similarity, with maximum length of 200 amino acids.	570 out of 1,292
TN		
alltox_TN	Non-toxins similar to entries in alltox80	3,483 out of 560,591
venom_TN	Non-toxins similar to entries in venom80	2,979 out of 560,591
venom_TN_500	Non-toxins similar to entries in venom80, with maximum length of 500 amino acids	2,879 out of 560,591
knottin_TN	Non-toxins similar to entries in knottin80	570 out of 560,591
knottin_TN_200	Non-toxins similar to entries in knottin80, with maximum length of 200 amino acids	570 out of 560,591
EFN		
alltox_not_venom80	All toxins that are not venom toxins, below 80% similarity	578 out of 1,147
alltox_not_venom80_500	All toxins that are not venom toxins, below 80% similarity, with maximum length of 500 amino acids	441 out of 1,147
alltox_not_knottin80	All toxins that are not knottins, below 80% similarity	2,941 out of 6,120
alltox_not_knottin80_200	All toxins that are not knottins, below 80% similarity, with maximum length of 200 amino acids	2,447 out of 6,120

The selected tools were set to be tested with the generated **TP**, **TN** and **EFN** datasets as summarized in Table 16, while the strategy employed for the datasets generation and so far described is summarized in Figure 4.

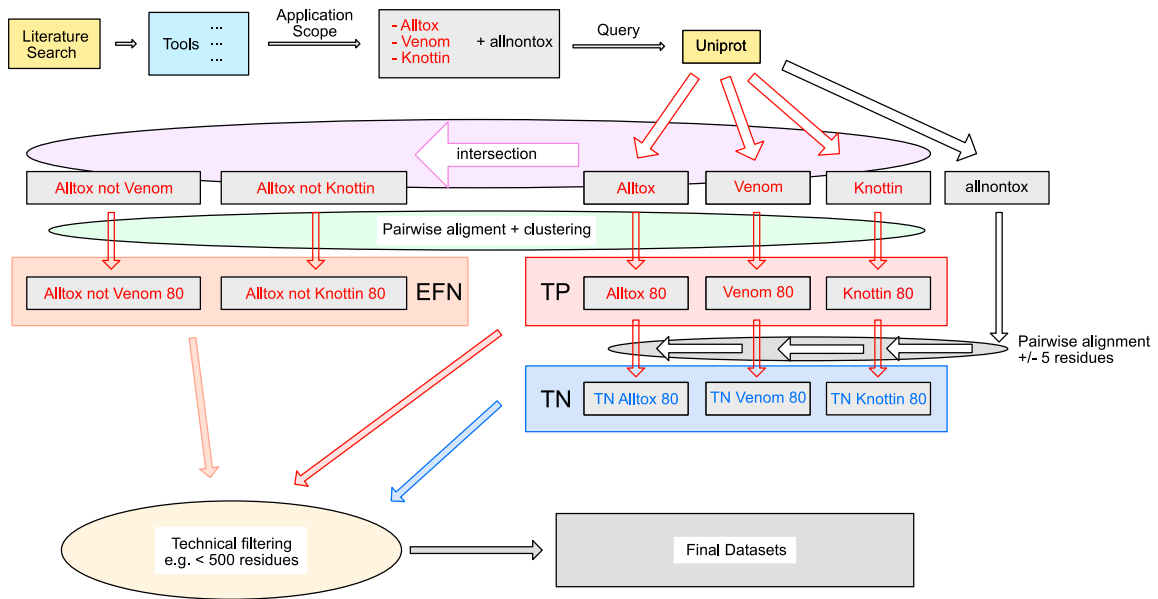


Figure 4: Flowchart describing the datasets generation strategy.

Table 16: Datasets to test individual tools.

Tool	Datasets		
	TP	TN	EFN
ToxCliasser	venom80_500	venom_TN_500	alltox_not_venom80_500
NNTox	alltox80	alltox_TN	N.A.
TOXIFY	venom80	venom_TN	alltox_not_venom80
ToxDL	alltox80	alltox_TN	N.A.
ToxIBTL	alltox80	alltox_TN	N.A.
ToxinPred2	alltox80	alltox_TN	N.A.
KNOTTIN	knottin80_200	knottin_TN_200	alltox_not_knottin80_200

4 TASK 4: assessment of tools and methodologies and pipeline definition

4.1 Testing tools against protein datasets

Tools selected during Task 2 were preliminary evaluated in Task 3.3 and then here thoroughly scrutinized again to understand their real usability and possibility to be included in a predictive pipeline. Table 17 summarizes the tools that were excluded, after a technical revision, alongside the update exclusion reasoning.

Table 17: Dropped tools

Tool	Motivation
ToxCClassifier	Unreachable
ToxDL	Unreachable / needs training
KNOTTIN	Not working / unreachable
DeepGraphGO *	Undocumented
FunfHMMER *	Not usable
MetaGO *	Mandatory structure, web server only with reply via e-mail
GOFGD *	Unreachable
ToxiTaxi	Not a predictive tool
ConoServer	Not a predictive tool
DBETH	Not a predictive tool
T1Tadb	Not a predictive tool
ClanTox	Unreachable
BTXPred	Outdated (website, one query at a time; not usable)
NTXPred	Outdated (website, one query at a time; not usable)
ToxinPred	Outdated (superseded)
PredCSF	Outdated (Website, results by e-mail, undocumented source code and missing installation instructions)

Unreachable: The tool website is offline, source code cannot be retrieved

Needs training: The model should be trained by the user

Not usable: The software cannot be utilized in a programmatic way

Outdated: The software was last updated more than 5 years ago

* tools from CAFA challenge not present in preliminary results (inserted later on EFSA request).

BTXPred, NTXPred, ToxinPred and PredCSF had already been excluded in Task 3, as for ToxiTaxi, ConoServer, T1Tadb and DBETH that are toxin database and not predicting tools.

In silico methodologies to predict the toxicity of novel proteins



All the remaining predictive tools were evaluated using TPs and TNs datasets from Task 3. Statistics of sensitivity and specificity of each method were calculated, and the method accuracy and applicability to toxin prediction was assessed. Table 18 reports the testing results.



Table 18: Tools evaluation statistics

Tool	dataset TP	dataset TN	dataset EFN	P	N	TP	FP	TN	FN	EFN (TPR)	TPR	TNR	PPV	NP V	FNR	FPR	FDR	FOR	ACC	BM	MCC
DeepGO	alltox_80	allnontox	N.A.	348	348	544	23	0	9	N.A.	0.16	0.99	0.96	0.54	0.84	0.01	0.04	0.46	0.57	0.15	0.23
DeepFri	alltox 80	allnontox	N.A.	348	348	126		282	221	N.A.	0.36	0.81	0.66	0.56	0.64	0.19	0.34	0.44	0.59	0.18	0.22
NNTox	alltox 80	allnontox	N.A.	348	311	329		311		N.A.	0.95	1.00	1.00	0.94	0.05	0.00	0.00	0.06	0.97	0.95	0.94
ToxinPred 2	alltox_80	allnontox	N.A.	348	348	330	179	168		N.A.	0.95	0.48	0.65	0.91	0.05	0.52	0.35	0.09	0.72	0.43	0.43
Toxify	venom80_50 0	tn_venom_50 0	alltox_not_ve nom80_500 (440)	287	287	259	120	167		0.33	0.90	0.58	0.68	0.86	0.10	0.42	0.32	0.14	0.74	0.48	0.52
TOXIBTL	alltox 80_50	allnontox_50	N.A.	104	107	914	294	783	132	N.A.	0.87	0.73	0.76	0.86	0.13	0.27	0.24	0.14	0.80	0.60	0.63

TPR True Positive Rate (Sensitivity, Recall, Hit Rate)
TNR True Negative Rate (Specificity, Selectivity)
PPV Positive Predictive Value (Precision)
NPV Negative Predictive Value
FNR False Negative Rate (Miss Rate)
FPR False Positive Rate (Fall-out)
FDR False Discovery Rate
FOR False Omission Rate
ACC Accuracy
BM Bookmaker Informedness
MCC Matthews Correlation Coefficient (Phi Coefficient)

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author(s).



From the careful observation of these results, CAFA tools DeepGO and DeepFRI have a very low accuracy (TPR < 40%), and, as such, they are not considered any further. NNTox is unapplicable to a real-case scenario, given that no GOs annotations exist for novel sequences. Notably, we tried to predict sequences GOs using the CAFA tool DeepGO and submit the predicted GOs to NNTox, unfortunately this resulted in a very poor accuracy. TOXIBTL is also excluded as usage of the source code is undocumented and it does not provide a programmatic access to the web server.

Hence, the tools ToxinPred2 and Toxify are singled out for the pipeline creation.

4.2 Testing methodologies

BLAST and HMM have been tested as such since the architecture of several tools is based on these methodologies. On the other hand, AI (i.e. SVM and NN) were used as workflow manager to improve the accuracy of the whole pipeline, building a consensus model.

4.2.1 BLAST

We assessed BLAST capability to discriminate between toxins and non-toxins by challenging the hypothesis that “toxins are more similar to other toxins than they are to non-toxins and, conversely, non-toxins are more similar to other non-toxins than they are to toxins.”

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families and is the gold standard in the field thanks to its accuracy. The core principle of BLAST involves breaking down the query sequence into smaller segments known as "words" that are used to search for matches within the database. BLAST uses a scoring system to assign significance scores to each match, indicating the level of similarity between the query and database sequences.

The BLAST algorithm works in five steps:

- 1. Word Generation:** BLAST starts by dividing the query sequence into overlapping words of a fixed length. These words act as the initial search seeds.
- 2. Seed Extension:** The algorithm extends the seeds in both directions to identify potential alignments, searching for matches that exceed a predefined threshold and considering factors such as sequence similarity and statistical significance;
- 3. Scoring:** BLAST employs a scoring system to assign values to matches based on the alignment quality, considering factors such as the presence of gaps, mismatches, and the overall similarity between the sequences;
- 4. Database Search:** BLAST performs a database search using the generated words and extended alignments, comparing the query sequence against the sequences stored in the database to identify similar regions;

5. **Ranking and Reporting:** The results are ranked based on the significance scores, with the most significant matches appearing at the top, providing various statistics and metrics to help researchers assess the reliability and significance of the matches.

We installed the standalone BLASTp program distributed by NIH-NCBI (<https://www.ncbi.nlm.nih.gov>). A database (herein referred to as target database) was built from the concatenation of the alltox (7,411) and allnontox (560,591) datasets, containing all reviewed toxins and all reviewed non-toxins, respectively.

Each sequence of this database was then used as a query and aligned against every other entry in the target database, recording the best 250 alignments for each query. The alignments were performed using BLAST standard settings (word-length = 5, gap opening penalty = 11, gap extension penalty = 11, conditional compositional score matrix adjustment, BLOSUM62 substitution matrix) which represent the best compromise for aligning a broad range of sequences with different identity levels. Results are reported in Annex 4.

Alignments were ranked based on their bit-score¹. We then analyzed the best matches for each protein in the dataset, measuring how likely it is that, if the query is a toxin, the closest match will also be a toxin, and how many toxins are present within the best 250 alignments for each query. The same procedure was applied also to non-toxins to measure the specificity of the methodology. Table 19 summarizes our results:

Table 19: Results of BLAST alignments

number of toxins which best match is a toxin:	6790 / 7240 (93.78%)
number of non-toxins which best match is a toxin:	461 / 559937 (0.08%)
mean fraction of toxins in the best 250 alignments for each toxin:	72.46%
mean fraction of toxins in the best 250 alignments for each non-toxin:	0.25%
number of toxins that align with at least one toxin within the best 250 alignments:	7061 / 7240 (97.53%)
number of non-toxins that align with at least one toxin within the best 250 alignments:	15954 / 559937 (2.85%)

¹ The BLASTP bit-score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The bit-score (S) is determined by the following formula: $S = (\lambda \times S - \ln K) / \ln 2$ where λ is the Gumble distribution constant, S is the raw alignment score, and K is a constant associated with the scoring matrix.

If the best match is a toxin, the methodology reaches a maximum sensitivity of 93.78% (true-positive rate) and a specificity (true-negative rate) of 99.91%, accuracy is 99.82%.

We focused on outliers misclassified by BLAST, i.e. toxins which best match is not a toxin, and toxins which do not entirely align with any other toxin and searching for common patterns among their GO and InterPro annotations.

We tested the method performance considering subsets of toxins from the dataset. Toxins were clustered at different distance thresholds based on their similarity, calculated using pairwise sequence alignment. For each distance threshold, BLAST alignment results were filtered by removing all the alignments where a query matched a target sequence that was not a representative medoid of the clustering procedure (<15 clusters), effectively increasing the heterogeneity of the toxins within the target database. Results show that even removing toxins from the main target database until there is at most 10% similarity between the remaining toxins, BLAST methodology still classifies correctly 60% of all toxins (Figure 5).

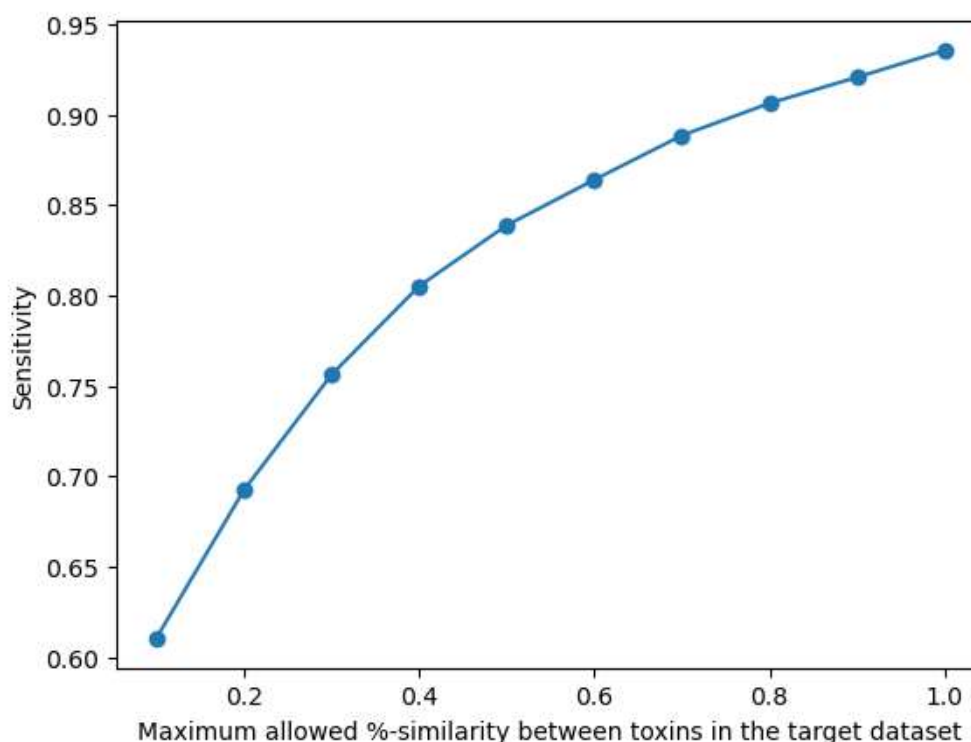


Figure 5: BLAST sensitivity at varying toxin dataset composition. The method sensibility was calculated at different clustering thresholds after keeping only the target toxins that were medoids of the clustering procedure.

4.2.2 Hidden Markov Model (HMM) profile alignment

InterPro (<https://www.ebi.ac.uk/interpro>) is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro uses predictive models, known
www.efsa.europa.eu/publications



as signatures, provided by several collaborating databases that collectively make up the InterPro consortium (CATH, CDD, HAMAP, MobiDB Lite, Panther, Pfam, PIRSF, PRINTS, Prosite, SFLD, SMART, SUPERFAMILY and NCBI-FAMs).

InterPro represents the state-of-the-art methodology for the task of classifying proteins into families and identify common domains (Jones et al., 2014). InterPro is able to classify protein sequences into families and predict their domains and functional sites using Hidden Markov Models. It integrates data from various sources, including protein sequence databases, protein family databases, and domain databases. By combining this information, InterPro offers a more comprehensive and accurate analysis of protein sequences.

InterPro uses several methods and tools to achieve its objectives:

1. **Sequence Analysis:** InterPro compares protein sequences against a collection of protein signatures, which are patterns or motifs associated with specific protein families or domains, identifying conserved regions and inferring functional information;
2. **Domain Prediction:** InterPro uses domain databases to predict the presence of specific protein domains within a given protein sequence that are structural and functional units that often play crucial roles in protein function;
3. **Protein Classification:** InterPro classifies proteins into families based on shared characteristics, such as sequence similarity, domain composition, and functional annotations;
4. **Functional Annotation:** InterPro provides functional annotations for proteins by integrating data from various sources, including Gene Ontology (GO) terms, protein::protein interaction databases, and literature information elucidating the biological roles and potential functions of proteins.

The software InterProScan is freely available (<https://github.com/ebi-pf-team/interproscan>) and can be used to annotate known and unknown protein sequences (Jones et al., 2014). However, it is worth noting that, since we are not dealing with unknown sequences, for our datasets, InterPro labels can be retrieved via UniProt cross-references. We nonetheless ensured that there was consistency between the InterPro labels retrieved via UniProt queries and the ones produced after running the InterProScan program on the whole toxins dataset. We found that there was almost full correspondence between InterPro queries, except for 8 toxins (ids: P15917, I2C090, Q91132, P40136, Q0ZZJ6, A0RZC6, J3S836, Q4MV79) for which the UniProt annotation did not fully match the one from InterPro, possibly due to automatic annotation issues. However, InterPro was able to produce annotations only for 5861 of 7,351 (80%) unique toxin sequences, regardless of the source of the annotation (Table 20).

InterProScan is a widely used tool in bioinformatics that utilizes HMMs to identify protein domains and functional motifs in protein sequences. It plays a crucial role in characterizing the functional properties of proteins based on their sequence information.

InterProScan integrates various protein signature databases, such as Pfam, PRINTS, ProSite, and others, which contain pre-built HMM profiles representing protein domains and motifs.



These HMM profiles are generated based on known protein families and their characteristic sequence patterns.

When a protein sequence is input into InterProScan, it searches against these HMM profiles using an algorithm called HMMER. HMMER compares the protein sequence to the HMM profiles and calculates a statistical score that represents the likelihood of the sequence belonging to a particular domain or motif.

The output of InterProScan provides valuable information about the presence of specific protein domains, functional sites, and other important features in the protein sequence. This information helps researchers understand the potential functions, interactions, and evolutionary relationships of proteins.

By leveraging the power of HMMs, InterProScan enables researchers to annotate and classify protein sequences at a functional level, allowing for a deeper understanding of the biological roles and properties of proteins. This information is crucial for studying protein evolution, protein-protein interactions, and the overall functioning of biological systems.

Table 20: Number of toxins with associated InterPro tags.

Number of toxins:	7,596
Number of unique sequences:	7,351
Number of toxins with InterPro xrefs:	5,861
Number of of toxins with computed InterPro tags:	5,861
Number of toxins with full correspondence between InterPro tags:	5,853

Given the very large number of InterPro annotation tags, results obtained from InterProScan annotation of sequences (Annex 5) cannot be readily utilized to discriminate toxins from non-toxins, hence we decided to train an SVC machine learning model to read the generated tags and predict protein toxicity. The model generation will be described in section 4.3.

4.3 Creation of a consensus model and pipeline

4.3.1 Generation of machine learning models based on bioinformatics methodologies

We developed two machine learning predictive models by leveraging the results obtained from applying BLAST and InterPro (HMM) bioinformatics methodologies. The primary objective was to condense their extensive outputs into a single scalar probability value, which would then serve as a foundation for constructing a consensus model.

To train and validate these models, we curated a dataset, referred to herein as “the dataset” (Annex 6). This dataset comprised all toxins from the previously mentioned alltox80 dataset, which encompasses all UniProt reviewed toxins with an 80% similarity threshold. Additionally, we included an equal number of their most similar non-toxin counterparts (the alltox_TN



dataset), as well as an equal number of randomly selected non-toxins from the allnontox dataset. The rationale for such dataset composition is to provide the models with examples of non-redundant toxins (alltox80), examples of a broad variety of non-toxins (randomly selected non-toxins), and examples of non-toxins with similar features to toxins (alltox_TN), that will force the models to carefully weight only the relevant features that are able to discriminate toxins from non-toxins.

The resulting database has the following composition:

- 3,594 toxins with no more than 80% similarity between themselves
- 3,594 non-toxins similar to the selected toxins
- 3,594 non-toxins randomly selected.

4.3.1.1 MLP Classifier based on BLAST results

We trained a multi-layer perceptron classifier on the following features extracted from BLAST all-vs-all alignment results:

- 1 or 0 whether the best alignment is a toxin or not
- Fraction of toxins within the first 250 alignments
- Percent identity of the first toxin within the first 250 alignments
- Percent identity of the first non-toxin within the first 250 alignments
- Bit-score of the first toxin within the first 250 alignments
- Bit-score of the first non-toxin within the first 250 alignment

The dataset was split into a training and test set (Annex 7) as reported in Table 21. The model hyperparameters were searched using a grid search approach, evaluating the number of hidden layers and their size. Models were ranked using a 5-fold cross-validation scheme. The resulting top scoring model has 19 layers with 350 perceptrons each.

Table 21: Main set and its division in both training and test sets.

MAIN SET	
10404 entries: 3448 toxins, 6956 non-toxins.	
TRAINING SET (66%)	TEST SET (33%)
6,970 entries: 2,276 toxins, 4,694 non-toxins.	3,434 entries: 1,172 toxins, 2,262 non-toxins.

The best model was then evaluated on the test set with the following results:

Sensitivity	Specificity	Accuracy
91%	97%	95%

For comparison, the performance of BLAST (see paragraph 4.2.1) alone on the same main dataset had a sensitivity of 89.41% and a specificity of 97.45%. These values were calculated using the first feature of the blast model, that is, whether the first match is a toxin or not. As the dataset is composed of toxins and non-toxins (similar respectively 80% maximum to toxins of interest), we are effectively reducing the chances to find a toxin best match for a toxin query, hindering the accuracy of BLAST alone. Nevertheless, our model was still able to correctly classify protein sequences with remarkable accuracy.

4.3.1.2 Support Vector Classifier based on InterPro annotations

To classify toxins and non-toxins based on InterPro, the dataset was filtered to contain only the proteins which had InterPro annotations (generated via InterProScan) and split into a training and test set (Annex 8) as reported in Table 22.

For each entry in the dataset, InterPro labels were encoded into a binary vector of 3,269 columns, which correspond to the number of unique InterPro tags in the training set. This is a common procedure used to represent categorical features for machine learning algorithms, named one-hot encoding: each InterPro label is represented as a column and has the value of 1 for every row corresponding to any proteins with an InterPro label, otherwise it has the value of 0.

Table 22: Main set and its division in both training and test sets.

MAIN SET	
10782 entries: 3594 toxins, 7188 non-toxins.	
TRAINING SET (66%) 7,223 entries: 2,407 toxins, 4,816 non-toxins.	TEST SET (33%) 3,559 entries: 1,187 toxins, 2,372 non-toxins.

Then a Support Vector Machine Classifier (SVC) machine learning model was trained on the 7223x3269 training set matrix. SVMs are well suited for this application, since they are effective in high dimensional spaces, also in cases where the number of dimensions is greater than the number of samples.

We used the SVC implementation provided by the scikit-learn Python library (Pedregosa et al., 2011), using a radial basis function kernel of 3rd degree polynomial and regularization parameters $C = 10$ and $\gamma = 0.09$, scaled on the number of features and their variance. The model hyperparameters were optimized using a grid search algorithm and models were



validated using a 5-fold cross-validation approach.² The model performance was then evaluated on the test set with the following results:

Sensitivity	Specificity	Accuracy
58%	96%	83%

The poor sensitivity of the model on its test set may be due to a bias in the hyperparameter search, which resulted in a high value for the regularization parameter C and a low value of gamma. This combination of parameters likely caused overfitting on the model training set, and, despite the 5-fold cross validation, the model was deemed as the best model by the optimization algorithm.

4.3.2 Creation of a consensus pipeline

The aim of this task was to generate a consensus model, considering the evaluations of the predictive tools and the models generated using Blast and InterProScan (IPS) and producing a single probability output of whether the sequence represents a toxin.

To train and test this model, we used the same dataset previously described for the generation of the BLAST and IPS models. For each entry in the dataset, we collected the scores from ToxinPred2 and Toxify, and the predictions from the models generated using BLAST and IPS.

On these features, we trained a multi-layer perceptron classifier after splitting the dataset into training and test sets (Annex 9) as reported in Table 23. Figure 6 summarizes the devised workflow.

² A 5-fold cross validation is a process when all data is randomly split into k folds, in our case k = 5, and then the model is trained on the k – 1 folds, while one-fold is left to test a model.

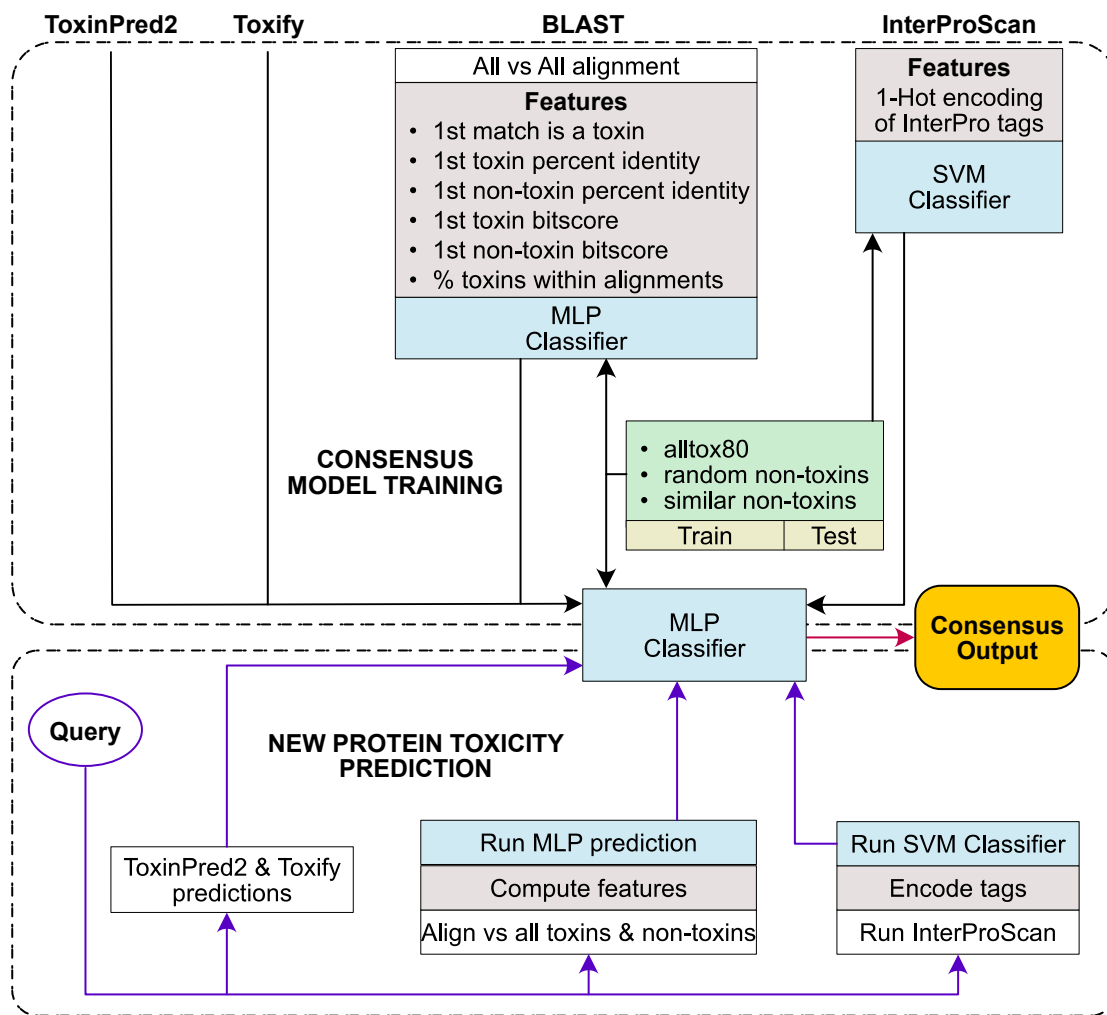


Figure 6: Schematic diagram of models training workflow (top rectangle): predictions on the dataset (green rectangle) are obtained using the predictive tools ToxinPred2 and Toxify. An MLP classifier is built on the features extracted from BLAST all-vs-all alignment; InterPro annotations are generated via IPS and used to train an SVM classifier. Predictions generated via ToxinPred2, Toxify and the two machine learning models are then used to train a consensus MLP classifier. A hypothetical workflow for an unknown protein query sequence is reported (bottom rectangle): i) ToxinPred2 and Toxify are used to infer protein toxicity, ii) the query sequence is aligned against all toxins and all non-toxins, the BLAST model is run upon calculation of the relevant features; iii) InterPro tags are generated via IPS, one-hot-encoded and fed to the SVM classifier. All predictions are then used to predict protein toxicity using the consensus model.

The consensus model hyperparameters were searched using a grid search approach, evaluating the number of hidden layers and their size. Models were ranked using a 5-fold cross-validation scheme. The resulting top scoring model has 25 layers with 400 perceptrons each.

Table 23: Main set and its division in both training and test sets.

MAIN SET	
10782 entries: 3594 toxins, 7188 non-toxins.	
TRAINING SET (66%)	TEST SET (33%)
7223 entries: 2407 toxins, 4816 non-toxins.	3559 entries: 1187 toxins, 2372 non-toxins.

The best model was then evaluated on the test set with the following results:

Sensitivity	Specificity	Accuracy
91%	96%	94%

4.3.3 Comparison of protein toxicity predictive strategies

To compare all the prediction strategies, encompassing the selected tools Toxify and ToxinPred2, and our machine learning models based on InterPro and BLAST, we calculated statistics of their performance against the whole dataset.³ Such dataset was the same used for our ML models. Machine learning models performance (MLP) was calculated on their respective test sets.

Table 24: Sensitivity, specificity and accuracy of the tested pipelines

Model	Sensitivity	Specificity	Accuracy
BLAST MLP	91%	97%	95%
InterPro SVC	58%	96%	83%
Consensus MLP	91%	96%	94%

The score metrics for each predictive tool and methodology on the same dataset (Annex 6) are reported in Table 24.

ToxinPred2 is the prediction tool with the highest sensitivity score, however, it should be considered that its specificity and accuracy (dependent on the false positives rate) are

³ This dataset contains all toxins from the alltox80 dataset (UniProt reviewed toxins not more similar than 80%), an equal number of their most similar non-toxin counterparts (the alltox_TN dataset), and an equal number of randomly sampled non-toxins from the allnontox dataset.



suboptimal, especially when considering false positives generated from hard-to-classify non-toxins similar to toxins (i.e. EFN).

Our models, in particular the BLAST and consensus MLPs, outperform every other predictive tool, having both sensitivity, specificity and, most importantly, accuracy over 95%.

Table 25: Summary of classification metrics among the different prediction approaches and combined consensus model. The best result for each column is highlighted. The maximum and mean prediction metrics are calculated by taking the highest value, or the mean value, respectively, among the Toxify, ToxinPred2, BLAST model and InterProScan model prediction scores.

prediction source	sensitivity	specificity			accuracy		
		whole	ntox	rndntox	whole	ntox	rndntox
Toxify	82%	79%	63%	98%	80%	72%	89%
ToxinPred2	97%	67%	40%	95%	77%	68%	96%
BLAST MLP	91%	97%	95%	100%	95%	93%	96%
InterPro SVC	92%	91%	91%	99%	92%	91%	92%
Consensus MLP	92%	96%	93%	100%	95%	92%	96%
Maximum	100%	64%	33%	94%	76%	66%	97%
Mean	95%	89%	73%	99%	89%	84%	97%

whole: the entire dataset; ntox: non-toxins similar to toxins; rndntox: randomly sampled non-toxins

There is a gain in sensitivity for the InterPro and consensus models performance on the whole main dataset compared to their performance on their respective test sets. This is because the whole dataset contains all the data that were used to generate the models train and test sets, to which each model might be slightly overfitted. Remarkably, the BLAST MLP model did not exhibit this behavior, indicating that the model is solid and does not suffer from overfitting. However, a large amount of these same data was also likely used to train ToxinPred2 and Toxify by their authors, thus evening out the overfitting bias in their comparison with our models. The common and unique source available to retrieve toxin sequences that is UniProt, and the impossibility to fabricate new or unknown toxin sequences data, and given the limited availability of toxin sequences, led us to consider this as the best possible strategy to maximize the usage of the available data on toxin sequences.

Table 25 reports the mutual relationships of misclassification between the different predictive strategies (Figure 7). ToxinPred2 emerges as the tool most able to correctly classify toxins where other methods fail, however, our BLAST and consensus MLP models are the ones with better specificity.



Figure 7: Orthogonality of predictive strategies. Heatmaps show the percentage of how many entries misclassified by method *i* (rows) are correctly classified by method *j* (columns).

As shown in Figure 8, scores are reported with respect to the presence of hard-to-classify non-toxins like toxins and to randomly sampled non-toxins. All prediction strategies classify toxins with high sensitivity, however predictive tools (ToxinPred2 and Toxify) struggle to correctly classify non-toxins like toxins, with the notable exception of the models generated on InterPro tags, BLAST alignments and the consensus MLP model. While the consensus and BLAST models work for every protein, the domain of applicability of the InterPro model is limited to the InterProScan ability to generate tags for a given sequence.

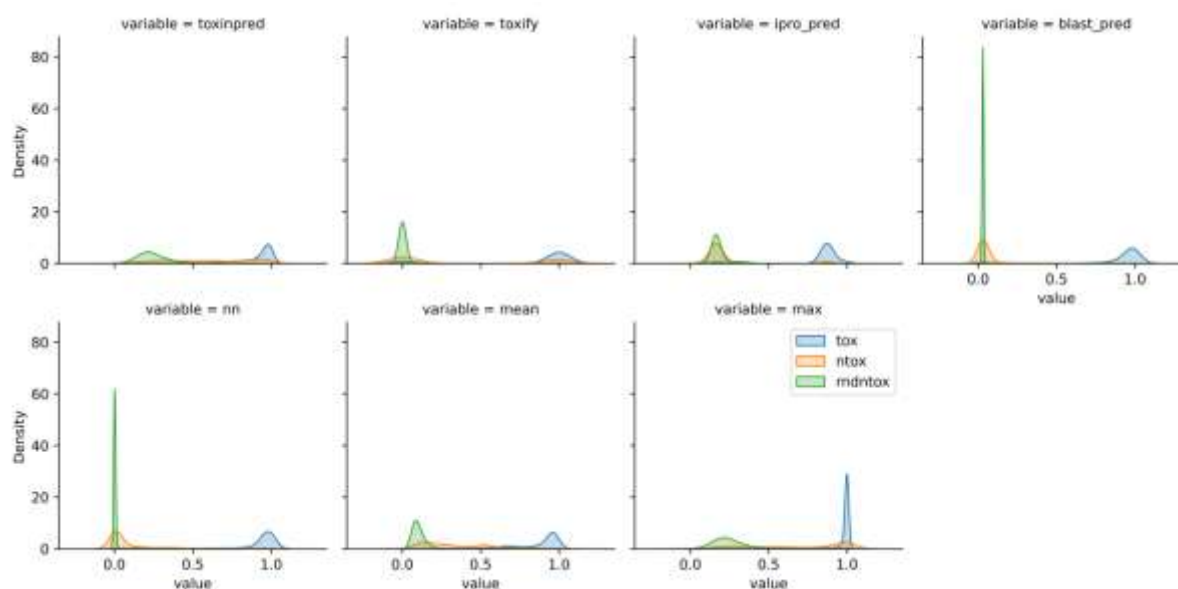


Figure 8: Distribution of prediction scores among different classification approaches with respect to sequence labels. Tox: toxins; ntox: non-toxins similar to toxins; rndntox: random non-toxins.

The consensus model was also trained using the leave-one-out methodology, i.e. removing a feature at a time. Results are reported in Table 26 and in Annex 10. From this analysis emerges that, removing one or more features leads to a drop in sensitivity compared to the full consensus model. The features that contribute the most to the model accuracy are predictions based on BLAST and InterProScan methodologies, whereas specifically removing the tool Toxify has no impact on the model accuracy.

Table 26: Consensus model generated using the leave-one-out strategy.

Removed features	sensitivity	specificity	accuracy
None	91%	94%	93%
Toxify	86%	96%	93%
ToxinPred2	88%	94%	92%
BLAST MLP	88%	92%	90%
InterPro	87%	94%	92%
Tools	89%	95%	93%
Methods	80%	87%	85%

Tools: Toxify and ToxinPred2 predictions; Methods: BLAST and InterPro models predictions

These results are not surprising, as both BLAST and InterProScan directly rely on alignments against annotated databases, *via* UniProtKB or InterPro HMM profiles, and thus leverage their knowledgebase, based on the central assumption of biochemistry that proteins sharing a similar sequence also share similar architecture and function.

5 Conclusions and recommendations

In silico prediction of protein toxicity is currently carried out under various regulatory frameworks in the EU. Generally, it is carried out by blasting strategies, based on the primary protein sequence and against toxins databases (often local and proprietary). Strengthening the *in silico* prediction of protein toxicity is highly needed to properly inform their risk assessment, for instance identifying tailored *in vitro* or *in vivo* studies. Against this background, we developed a preliminary integrated *in silico* pipeline to predict the toxicity of (novel) proteins, based on commercially available toxin prediction tools and other bioinformatic methodologies.

As a first step, we carried out a horizon scanning of commercially available tools for the *in silico* prediction of protein toxicity and thoroughly tested the most up-to-date for their predictive accuracy using a curated benchmark dataset. This was specifically designed to



contain a balanced set of annotated toxins and non-toxins sourced from the UniProt reference database; careful consideration on redundancy ensured the removal of closely related or identical sequences from the benchmark dataset. Of note, this dataset is a valuable resource not only for protein toxicity prediction but also for other applications requiring well-defined, annotated protein sequences: its rigorous curation process makes it an ideal material for evaluating various bioinformatics tools and models, as it serves as a reliable reference point for protein function analysis, classification or further studies on sequence-function relationships. The specific capabilities and limitations of the selected *in silico* toxicity prediction tools were scrutinized. Their intended application scope was considered (e.g. designed for general toxicity prediction, focused on specific types of proteins or optimized for certain environments or organisms) and special attention was given to the sequence length limitations imposed by each tool (different tools may have varying capacities for processing short versus long protein sequences). Noteworthy, all the tools identified in this study rely on the primary structure of proteins as the foundational input for their predictions, without considering higher-order structural elements such as secondary, tertiary, or quaternary structures. By operating exclusively on the primary structure, these tools typically use algorithms designed to analyze specific sequence motifs, patterns, or physicochemical properties associated with toxicity. The selected tools were also evaluated as regards their usability for the construction of an automated pipeline for prediction of protein toxicity, i.e. the capability to be programmatically interacted with, either through a stand-alone application or via a web-based automatic programmatic interface (API). ToxinPred2 and Toxify were the top-scoring tools in terms of protein toxicity prediction accuracy and usability for the development of an automated pipeline.

In response to EFSA's request, we extended our search to include tools capable of predicting general protein functions, beyond just toxin activity. The rationale behind this broader exploration was to assess whether function prediction tools, which are designed to capture a wider range of biological activities, could offer insights into protein toxicity. To identify the most promising candidates, we focused on tools that had performed exceptionally well in the CAFA (Critical Assessment of Functional Annotation) challenge, an international competition that ranks methods based on their ability to predict protein function accurately. However, despite their strong performance in predicting general protein function, these tools proved inadequate when applied specifically to the task of protein toxicity prediction. As a result, they were excluded from further testing in this context.

We also explored the applicability of other bioinformatic methodologies to protein toxicity prediction. Namely, we tested BLAST and InterPro Hidden Markov Model (HMM) profile alignments. BLAST is a heuristic local sequence alignment algorithm and program useful to infer evolutionary relationships among protein sequences. InterPro is a database of HMM profiles of protein families and domains, and its program InterProScan can generate annotations for protein sequences. It is important to note that these methodologies do not predict anything *per se*, but instead provide information on protein sequences based on their similarity, and consequently homology, with entries of pre-existing and annotated databases (i.e., UniProt and InterPro). Thus, we organized the output of these methodologies into cleverly-devised features to be used in the training of machine learning models, with the aim to discriminate toxins from non-toxins. Our models provide astounding accuracy on protein



toxicity prediction, and there is room for improvement with respect to the risk of model overfitting.

Finally, we developed an AI-based consensus model was developed to integrate the outputs of the selected predictive tools, including ToxinPred2, Toxify, and machine learning models constructed using BLAST and InterPro data. This consensus model is capable of distinguishing toxins from non-toxins with a 95% predictive accuracy and, to the best of our knowledge, it represents the state-of-the-art in protein toxicity prediction based solely on sequence analysis. When comparing predictions, our BLAST MLP (Multi-Layer Perceptron) model demonstrates a similar level of accuracy to that of the consensus model. However, there is a pivotal distinction between the two. The BLAST MLP model accuracy heavily depends on the presence of phylogenetic relationships between the query sequence and known proteins in the database. This reliance can be associated with limitations, as the current toxic protein datasets tend to exhibit sampling bias, with an overrepresentation of toxins from specific organisms (e.g., animals) and toxin classes (e.g., scorpion venom toxins). As a result, BLAST-based predictions are often skewed toward well-represented toxin families, while novel or rare toxins — those lacking close homologs or stemming from under-investigated sources — are less reliably detected. In contrast, the consensus model combines outputs from multiple tools, some of which, like ToxinPred2 and Toxify, are not reliant on phylogenetic relationships but instead focus on sequence-based features. This allows them to perform better in cases where toxins may not have clear homologs in databases, such as UniProtKB, or where the toxins originate from artificial mutations or less studied organisms. InterPro also contributes by identifying functional domains and complementing the analysis.

In conclusion, we developed an integrated *in silico* pipeline that combines proprietary available toxin prediction tools, such as ToxinPred2 and Toxify, with bioinformatics methodologies like BLAST and InterPro. These tools were rigorously evaluated for their predictive accuracy and ease of integration into an automated system. A key outcome of our work is the development of an AI-based consensus model, which combines the outputs of multiple methodologies, including ToxinPred2, Toxify, BLAST, and InterPro. This model achieves a 95% predictive accuracy in distinguishing toxins from non-toxins, representing the state-of-the-art in protein toxicity prediction based on sequence analysis. Compared to traditional homology-based methods, like the BLAST MLP model, which performs well when phylogenetic relationships are present, the consensus model offers enhanced performance, particularly for novel or rare toxins where homologs may be absent or poorly represented in databases like UniProtKB. Structure-based predictive methods, while theoretically promising, were not considered practical for this task due to the resource-intensive nature of determining protein structures. Although advances in AI-based tools, like AlphaFold and Rosetta, are impressive, they remain under rapid development and are not yet suitable for a regulatory use.

Moreover, we would like to recommend some points:

1) Development of an Open-Source, User-Friendly Tool

Future efforts should prioritize the creation of an open-source, stand-alone tool for protein toxicity prediction. This tool should be designed with user-friendliness in mind, allowing non-



bioinformatician risk assessors to perform accurate, preliminary toxicity evaluations with ease. This tool would improve accessibility and transparency in regulatory processes.

2) Regular Updates to Databases and Models

To ensure the continued reliability of predictions, it is essential that the models, particularly those relying on sequence similarity (e.g., BLAST, ToxinPred2, and Toxify), be regularly updated with the latest protein sequence data from databases such as UniProt and InterPro. Keeping models current will enhance their ability to identify novel and emerging toxins accurately.

3) Incorporation of 3D Structural Information

Although sequence-based models are highly effective, future developments should explore the potential of integrating 3D structural information into prediction models. The generation of structural feature descriptors, alongside primary sequence data, could further improve the accuracy and sensitivity of toxicity predictions, especially for proteins with unknown homologs.

4) Addressing Sampling Bias in Databases

To improve the reliability of BLAST-based predictions, efforts should be made to reduce the sampling bias inherent in current toxic protein datasets. Expanding the range of organisms and toxin classes represented in databases will provide a more balanced dataset, leading to more generalizable and accurate toxicity predictions.

5) Collaboration with Regulatory Bodies

Continued collaboration with regulatory bodies, such as EFSA, is essential to ensure that the developed tools align with the requirements of regulatory frameworks. This collaboration will facilitate the integration of these tools into official risk assessment workflows, and ensure their applicability in real-world decision-making processes.

6) Evaluation of AI and Machine Learning Models for Regulatory Use

While AI-based methods have shown tremendous potential, further validation of these models is necessary to ensure their accuracy and robustness, particularly in a regulatory context. Additional benchmarks, stress tests, and external validations should be conducted to ensure that machine learning-based predictions are trustworthy, stable and suitable for regulatory adoption.



6 Bibliography

- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Z.X., 2016. Tensorflow: a system for large-scale machine learning. OSDI 265–283.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* <https://doi.org/10.1038/75556>
- Bailey, T.L., Johnson, J., Grant, C.E., Noble, W.S., 2015. The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. <https://doi.org/10.1093/nar/gkv416>
- Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. <https://doi.org/10.1093/nar/28.1.45>
- Baranek, J., Pogodziński, B., Szipluk, N., Zielezinski, A., 2020. TOXiTAXi: a web resource for toxicity of *Bacillus thuringiensis* protein compositions towards species of various taxonomic groups. *Sci. Rep.* 10, 1–12. <https://doi.org/10.1038/s41598-020-75932-7>
- Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L.G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nospikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roehert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y.,



Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., 2017. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. <https://doi.org/10.1093/nar/gkw1099>

- Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T.G., Fan, J., Castro, L.G., Garmiri, P., Georgiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K.C., Coudert, E., Cucho, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T.B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., Ruch, P., Teodoro, D., 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2007. GenBank. *Nucleic Acids Res.* 35, D21–D25. <https://doi.org/10.1093/nar/gkl986>
- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C., 2002. The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 58, 899–907. <https://doi.org/10.1107/S0907444902003451>
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370. <https://doi.org/10.1093/nar/gkg095>
- Brendel, V., 1992. PROSET—a fast procedure to create non-redundant sets of protein sequences. *Math. Comput. Model.* 16, 37–43. [https://doi.org/10.1016/0895-7177\(92\)90150-J](https://doi.org/10.1016/0895-7177(92)90150-J)
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10. <https://doi.org/10.1186/1471-2105-10-421>



- Carnate, M., Ed, M., 2008. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci* 95, 5857–5864. <https://doi.org/10.1073/pnas.95.11.5857>
- Chakraborty, A., Ghosh, S., Chowdhary, G., Maulik, U., Chakrabarti, S., 2012. DBETH: A database of bacterial exotoxins for human. *Nucleic Acids Res.* 40, 615–620. <https://doi.org/10.1093/nar/gkr942>
- Chan, A.P., Pertea, G., Cheung, F., Lee, D., Zheng, L., Whitelaw, C., Pontaroli, A.C., SanMiguel, P., Yuan, Y., Bennetzen, J., Barbazuk, W.B., Quackenbush, J., Rabinowicz, P.D., 2006. The TIGR Maize Database. *Nucleic Acids Res.* 34, D771–776. <https://doi.org/10.1093/nar/gkj072>
- Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A Library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2. <https://doi.org/10.1145/1961189.1961199>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.* 1724–1734. <https://doi.org/10.3115/v1/d14-1179>
- Cole, T.J., Brewer, M.S., 2019. TOXIFY: A deep learning approach to classify animal venom proteins. *PeerJ* 2019. <https://doi.org/10.7717/peerj.7200>
- Darty, K., Denise, A., Ponty, Y., 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 1974–1975. <https://doi.org/10.1093/bioinformatics/btp250>
- Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>
- EFSA, 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA J.* 8. <https://doi.org/10.2903/j.efsa.2010.1637>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D., 2019. The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. <https://doi.org/10.1093/nar/gky995>
- Fan, Y.-X., Song, J., Kong, X., Shen, H.-B., 2011. PredCSF: An integrated feature-based approach for predicting conotoxin superfamily. *Protein Pept. Lett.* 18, 261–267. <https://doi.org/10.2174/092986611794578341>
- Fan, Yong-Xian, Song, J., Kong, X., Shen, H.-B., 2011. PredCSF: An Integrated Feature-Based Approach for Predicting Conotoxin Superfamily. *Protein Pept. Lett.* 18, 261–267. <https://doi.org/10.2174/092986611794578341>
- Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. <https://doi.org/10.1093/nar/gkv1344>



- Fozo, E.M., Makarova, K.S., Shabalina, S.A., Yutin, N., Koonin, E. V, Storz, G., 2010. Abundance of type I toxin-antitoxin systems in bacteria: Searches for new candidates and discovery of novel families. *Nucleic Acids Res.* 38, 3743–3759. <https://doi.org/10.1093/nar/gkq054>
- Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gacesa, R., Barlow, D.J., Long, P.F., 2016. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Comput. Sci.* 2016. <https://doi.org/10.7717/peerj-cs.90>
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein identification and analysis tools on the ExpASY server. *Proteomics Protoc. Handb.* 571–607.
- Gelly, J.C., Gracy, J., Kaas, Q., Le-Nguyen, D., Heitz, A., Chiche, L., 2004. The KNOTTIN website and database: A new information system dedicated to the knottin scaffold. *Nucleic Acids Res.* 32, 156–159. <https://doi.org/10.1093/nar/gkh015>
- Gracy, J., Le-Nguyen, D., Gelly, J.C., Kaas, Q., Heitz, A., Chiche, L., 2008. KNOTTIN: The knottin or inhibitor cystine knot scaffold in 2007. *Nucleic Acids Res.* 36, 314–319. <https://doi.org/10.1093/nar/gkm939>
- Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Raghava, G.P.S., 2013. In Silico Approach for Predicting Toxicity of Peptides and Proteins. *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0073957>
- Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S., 2007. Quantifying similarity between motifs. *Genome Biol.* 8. <https://doi.org/10.1186/gb-2007-8-2-r24>
- Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>
- Jain, A., Kihara, D., 2019. NNTox: Gene Ontology-Based Protein Toxicity Prediction Using Neural Network. *Sci. Rep.* 9, 17923. <https://doi.org/10.1038/s41598-019-54405-6>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S., 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Jungo, F., Bairoch, A., 2005. Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon* 45, 293–301. <https://doi.org/10.1016/j.toxicon.2004.10.018>
- Jungo, F., Bougueleret, L., Xenarios, I., Poux, S., 2012. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon* 60, 551–557. <https://doi.org/10.1016/j.toxicon.2012.03.010>
- Kaas, Q., Westermann, J.C., Halai, R., Wang, C.K.L., Craik, D.J., 2008. ConoServer, a



- database for conopeptide sequences and structures. *Bioinformatics* 24, 445–446. <https://doi.org/10.1093/bioinformatics/btm596>
- Kaas, Q., Yu, R., Jin, A.H., Dutertre, S., Craik, D.J., 2012. ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkr886>
- Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences, in: 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. pp. 655–665. <https://doi.org/10.3115/v1/p14-1062>
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. <https://doi.org/10.1093/molbev/mst010>
- Katoh, K., Toh, H., 2008. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 9. <https://doi.org/10.1186/1471-2105-9-212>
- Kawashima, S., Kanehisa, M., 2000. AAindex: Amino acid index database. *Nucleic Acids Res.* 28, 374.
- Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M., 2006. MUSTANG: A multiple structural alignment algorithm. *Proteins Struct. Funct. Genet.* 64, 559–574. <https://doi.org/10.1002/prot.20921>
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. <https://doi.org/10.1093/bioinformatics/btm404>
- Li, S., Chen, J., Liu, B., 2017. Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics* 18. <https://doi.org/10.1186/s12859-017-1842-2>
- Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. <https://doi.org/10.1093/nar/gkw1129>
- McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405. <https://doi.org/10.1093/bioinformatics/16.4.404>
- Neumann, R.S., Kumar, S., Shalchian-Tabrizi, K., 2014. BLAST output visualization in the new sequencing era. *Brief. Bioinform.* 15, 484–503. <https://doi.org/10.1093/bib/bbt009>
- Palazzolo, L., Gianazza, E., Eberini, I., 2020. Literature search – Exploring in silico protein toxicity prediction methods to support the food and feed risk assessment. *EFSA Support. Publ.* 17. <https://doi.org/10.2903/sp.efsa.2020.en-1875>



- Pan, X., Zuallaert, J., Wang, X., Shen, H. Bin, Campos, E.P., Marushchak, D.O., De Neve, W., 2020. ToxDL: Deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* 36, 5159–5168. <https://doi.org/10.1093/bioinformatics/btaa656>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Porollo, A.A., Adamczak, R., Meller, J., 2004. POLYVIEW: A flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics* 20, 2460–2462. <https://doi.org/10.1093/bioinformatics/bth248>
- Postic, G., Gracy, J., Périn, C., Chiche, L., Gelly, J.C., 2018. KNOTTIN: The database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkx1084>
- Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., Finn, R.D., 2018. HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. <https://doi.org/10.1093/nar/gky448>
- Saha, S., Raghava, G.P.S., 2007a. Prediction of neurotoxins based on their function and source. *In Silico Biol.* 7, 369–387.
- Saha, S., Raghava, G.P.S., 2007b. BTXpred: Prediction of bacterial toxins. *In Silico Biol.* 7, 405–412.
- Schneider, T.D., Stephens, R.M., 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100. <https://doi.org/10.1093/nar/18.20.6097>
- Sharma, N., Naorem, L.D., Jain, S., Raghava, G.P.S., 2022. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform.* 1–12. <https://doi.org/10.1093/bib/bbac174>
- Starcevic, A., Moura-Da-Silva, A.M., Cullum, J., Hranueli, D., Long, P.F., 2015. Combinations of long peptide sequence blocks can be used to describe toxin diversification in venomous animals. *Toxicon* 95, 84–92. <https://doi.org/10.1016/j.toxicon.2015.01.005>
- Stothard, P., Wishart, D.S., 2005. Circular genome visualization and exploration using CGView. *Bioinformatics* 21, 537–539. <https://doi.org/10.1093/bioinformatics/bti054>
- Tatusov, R.L., Natale, D.A., Garkavtsev, I. V, Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E. V, 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28. <https://doi.org/10.1093/nar/29.1.22>
- Tourasse, N.J., Darfeuille, F., 2021. T1TAdb: The database of type I toxin-antitoxin systems. *Rna* 27, 1471–1481. <https://doi.org/10.1261/rna.078802.121>
- Wei, Lesong, Ye, X., Sakurai, T., Mu, Z., Wei, Leyi, 2022. ToxIBTL: Prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 38,



1514–1524. <https://doi.org/10.1093/bioinformatics/btac006>

Wei, Lesong, Ye, X., Xue, Y., Sakurai, T., Wei, Leyi, 2021. Atse: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief. Bioinform.* 22, 1–13. <https://doi.org/10.1093/bib/bbab041>

Wong, E.S.W., Hardy, M.C., Wood, D., Bailey, T., King, G.F., 2013. SVM-Based Prediction of Propeptide Cleavage Sites in Spider Toxins Identifies Toxin Innovation in an Australian Tarantula. *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0066279>

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415. <https://doi.org/10.1093/nar/gkg595>

Zuker, M., 1989. On finding all suboptimal foldings of an RNA molecule. *Science* (80-.). 244, 48–52. <https://doi.org/10.1126/science.2468181>

Appendix A Presentation of retrieved tools to predict protein toxicity

A.1. Presentation of retrieved tools to predict protein toxicity

A.1.1. Neural network/Artificial intelligence application to protein toxicity prediction

Primary citations: Vishnoi S, Matre H, Garg P, Pandey SK. *Artificial intelligence and machine learning for protein toxicity prediction using proteomics data*. Chem Biol Drug Des. 2020 Sep;96(3):902-920.

Abstract: *Instead of only focusing on the targeted drug delivery system, researchers have a great interest in developing peptide-based therapies for the procurement of numerous class of diseases. The main idea behind this is to anchor the properties of the receptor to design peptide-based therapeutics. As these macromolecules have distinct physicochemical properties over small molecules, it becomes an obligatory field for the treatment of diseases. For this, various in silico models have been developed to speculate the proteins by virtue of the application of machine learning and artificial intelligence. By analysing the properties and structural alert of toxic proteins, researchers aim to dissert some of the mechanisms of protein toxicity from which therapeutic insights may be drawn. Numerous models already exist worldwide emphasizing themselves as leading paramount for toxicity prediction in protein macromolecules. Few of them comparatively compete with the other predictive protein toxicity models and convincingly give a high-performance result in terms of accuracy. But their foundation is quite ambiguous, and varying approaches are found at the level of toxicoproteomic data utilization while building a machine learning model. In this review work, we present the contribution of artificial intelligence and machine learning approaches in prediction of protein toxicity using proteomics data.*

Citations: 6 citations.

A.1.2. BTXpred (Saha and Raghava, 2007b)

Primary reference: Saha S, Raghava GP. *BTXpred: prediction of bacterial toxins*. In Silico Biol. 2007;7(4-5):405-12. PMID: 18391233.

Abstract: *This paper describes a method developed for predicting bacterial toxins from their amino acid sequences. All the modules, developed in this study, were trained and tested on a non-redundant dataset of 150 bacterial toxins that included 77 exotoxins and 73 endotoxins. Firstly, support vector machines (SVM) based modules were developed for predicting the*

bacterial toxins using amino acids and dipeptides composition and achieved an accuracy of 96.07% and 92.50%, respectively. Secondly, SVM based modules were developed for discriminating enterotoxins and exotoxins, using amino acids and dipeptides composition and achieved an accuracy of 95.71% and 92.86%, respectively. In addition, modules have been developed for classifying the exotoxins (e.g. activate adenylate cyclase, activate guanylate cyclase, neurotoxins) using hidden Markov models (HMM), PSI-BLAST and a combination of the two and achieved overall accuracy of 95.75%, 97.87% and 100%, respectively. Based on the above study, a web server called 'BTXpred' has been developed, which is available at <http://www.imtech.res.in/raghava/btxpred/>. Supplementary information is available at <http://www.imtech.res.in/raghava/btxpred/supplementary.html>.

Link: The link reported in the primary article (<http://www.imtech.res.in/raghava/btxpred/>) is deprecated. Active link is the following: <https://webs.iitd.edu.in/raghava/btxpred/>

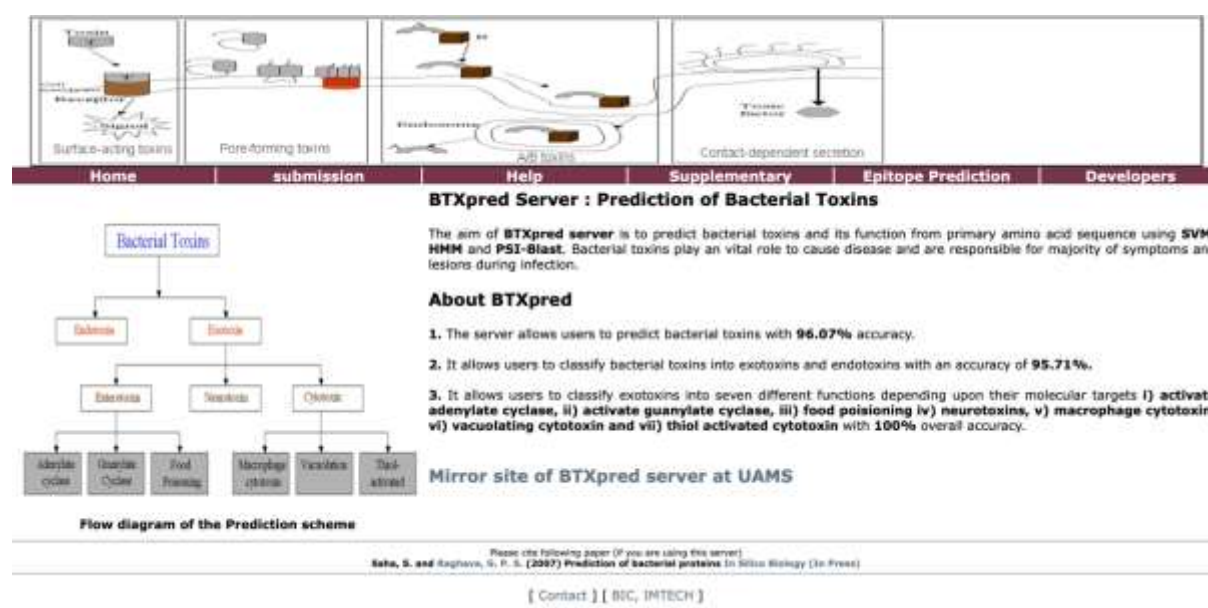


Figure 9: BTXPred homepage.

Citations: 46 citations (40 after 2012)

Field of application: bacterial toxins

Training dataset: A non-redundant dataset of 150 bacterial toxins with 77 exotoxins and 73 endotoxins was obtained pruning sequences that have more than 90% sequence identity with PROSET software (Brendel, 1992). Authors state that "the data set is available at <http://www.bioinfo.de/isb/2007/07/0028/>" but today the link is unavailable; the active link is the following: <https://webs.iitd.edu.in/raghava/btxpred/supplementary.html>

Exotoxins were further classified according to their molecular targets, i) activate adenylate cyclase, ii) activate guanylate cyclase, iii) food poisoning, iv) neurotoxins, v) macrophage



cytotoxin, vi) vacuolating cytotoxin and vii) thiol activated cytotoxin; viii) hemolysin. The True Negative dataset is composed by 500 non-toxin proteins from prokaryotic and eukaryotic origin. These proteins were retrieved by searching for the term "function" in the "Comment" field in Swiss-Prot, but excluding entries with the term "toxin" in the same field by the 'BUTNOT' option. Moreover, Authors manually checked the obtained proteins to verify that none of them was classified as toxin.

Predicting methods: BTXPred uses three different methodologies (SVM, HMM, and PSI-BLAST) to predict if a protein is a bacterial toxin. BTXPred also predicts the bacterial toxin function from primary amino acid sequence using SVM, HMM and PSI-BLAST.

SVM was implemented using the freely downloadable software package SVM_light with radial basis function (RBF) kernel. The input vectors used are amino acid composition (20 vectors) and dipeptide composition (400 vectors) of each protein sequence.

In addition, HMM profiles were generated for seven sub-classes of exotoxins using HMMER (Eddy, 1998). The multiple alignment of protein sequences was obtained using CLUSTAL-W. The program hmmbuild of HMMER has been used to build profile HMM and then calibrated using hmmcalibrate. The program Hmmpfam was used for searching a query sequence against the created profile HMM database.

Finally, PSI-BLAST module is used to align the query sequence against test dataset (Altschul et al., 1997). Performance for all the modules was evaluated using a 5-fold cross-validation technique: the dataset was randomly divided into five sub-sets: four of them were used for training and the remaining for testing. This process is repeated five times so that each set is used for testing once. Leave-one-out cross-validation (LOOCV) technique was also used for the evaluation of the modules developed for the prediction of sub classes of exotoxins.

Figure 10: Submission form of BTXPred

Results: Table 27 and table 28 report the statistics declared in the primary research article.

Table 27: Performance SVM- and of PSI-BLAST-based methods in prediction of bacterial toxins declared by Authors. PPV is the positive predictive value while MCC is the Matthew's correlation coefficient.

Approach	Sensitivity	Specificity	PPV	Accuracy	MCC
Amino Acids	0.92	1	1	0.96	0.93
Dipeptides	0.86	0.99	0.98	0.93	0.86
PSI-BLAST (Toxin)	0.67	NA	NA	NA	NA

Table 28: Performance SVM- and of PSI-BLAST-based methods discriminating between exotoxins and endotoxins declared by Authors. PPV is the positive predictive value while MCC is the Matthew's correlation coefficient.

Approach	Sensitivity	Specificity	PPV	Accuracy	MCC
Amino Acids	1	0.91	0.93	0.96	0.92
Dipeptides	0.94	0.91	0.92	0.93	0.86
PSI-BLAST (exotoxin)	0.46	NA	NA	NA	NA
PSI-BLAST (endotoxin)	0.90	NA	NA	NA	NA

A.1.3. NTXpred (Saha and Raghava, 2007a)

Primary reference: Saha S and Raghava GP. *Prediction of neurotoxins based on their function and source*. In *Silico Biol.* 2007;7(4-5):369-87. PMID: 18391230.

Abstract: *We have developed a method NTXpred for predicting neurotoxins and classifying them based on their function and origin. The dataset used in this study consists of 582 non-redundant, experimentally annotated neurotoxins obtained from Swiss-Prot. A number of modules have been developed for predicting neurotoxins using residue composition based on feed-forwarded neural network (FNN), recurrent neural network (RNN), support vector machine (SVM) and achieved maximum accuracy of 84.19%, 92.75%, 97.72% respectively. In addition, SVM modules have been developed for classifying neurotoxins based on their source (e.g., eubacteria, cnidarians, molluscs, arthropods have been and chordate) using*

www.efsa.europa.eu/publications

amino acid composition and dipeptide composition and achieved maximum overall accuracy of 78.94% and 88.07% respectively. The overall accuracy increased to 92.10%, when the evolutionary information obtained from PSI-BLAST was combined with SVM module of source classification. We have also developed SVM modules for classifying neurotoxins based on functions using amino acid, dipeptide composition and achieved overall accuracy of 83.11%, 91.10% respectively. The overall accuracy of function classification improved to 95.11%, when PSI-BLAST output was combined with SVM module. All the modules developed in this study were evaluated using five-fold cross-validation technique. The NTXpred is available at www.imtech.res.in/raghava/ntxpred/ and mirror site at <http://bioinformatics.uams.edu/mirror/ntxpred>.

Link: The link reported in the primary article (www.imtech.res.in/raghava/ntxpred/) is deprecated. Active link is the following:

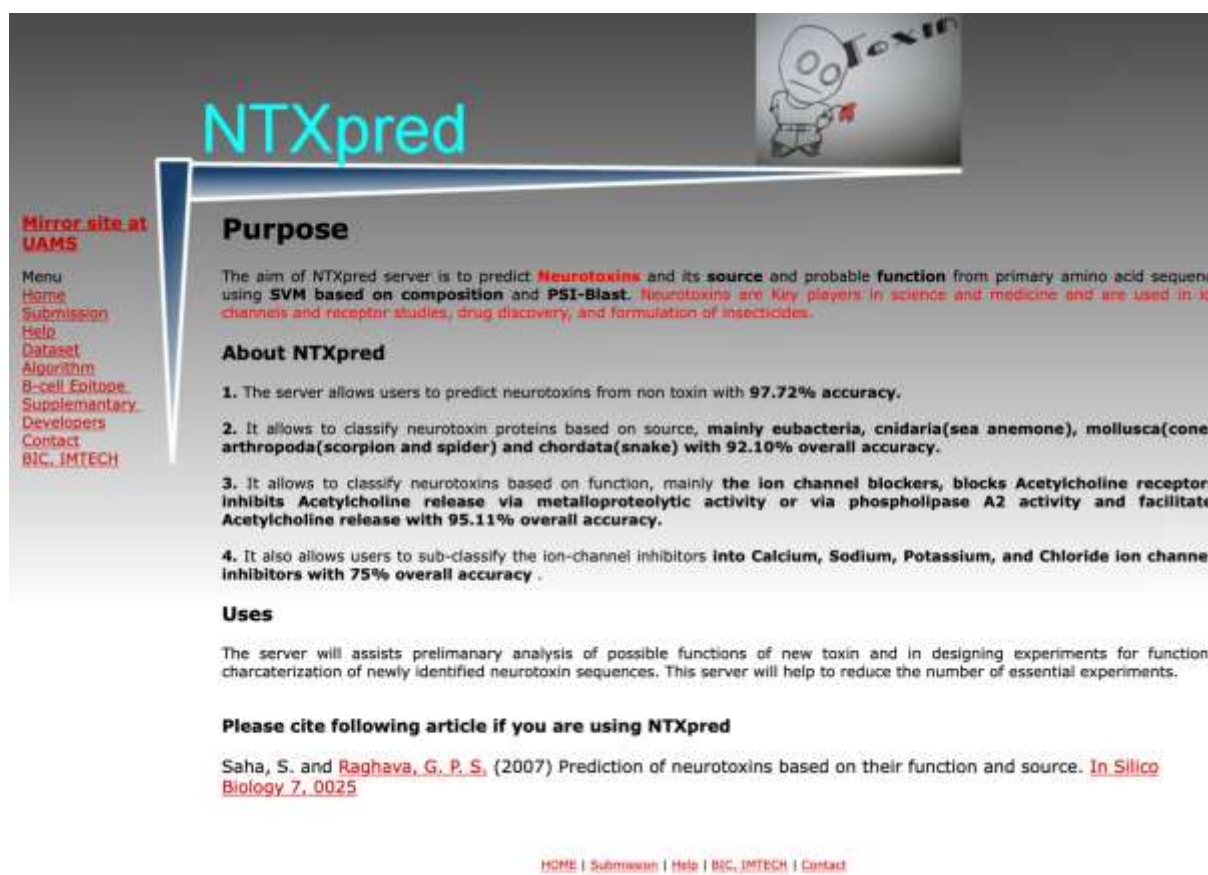


Figure 11: NTXPred homepage.

Citations: 39 citations (31 after 2012)



Field of application: neurotoxins

Training dataset: A non-redundant dataset of 582 neurotoxins was used for training as True Positive dataset. These neurotoxins were classified according to their source into (i) eubacteria (13); (ii) cnidaria (23); (iii) mollusca (95); (iv) arthropoda (313); and (v) chordata (138). Then, these non-redundant neurotoxin sequences were further classified into five sub-classes based on their target of action as (i) ion channels blockers (332); (ii) blockers of acetylcholine receptors (89); (iii) inhibitors of neurotransmitter release via metalloproteolytic activity (8); (iv) inhibitors of acetylcholine release with phospholipase A2 activity (21); and (v) facilitators of acetylcholine release (10). Finally, ion channel blockers were sub-classified into (i) calcium (81); (ii) chloride (8); (iii) potassium (91); (iv) sodium (150) ion channel blockers. The True Negative dataset is composed by 582 non-toxin proteins. These proteins were retrieved by searching for the term "function" in the "Comment" field of UniProtKB/Swiss-Prot but excluding entries with the term "toxin" in the same field by the 'BUTNOT' option. Moreover, Authors manually checked the obtained proteins to verify that none of them was classified as toxin. True Negative dataset of NTXPred share some proteins with the BTPred one.

PROSET was used software to prune the proteins with more than 90% sequence identity.

Predicting methods: NTXPred uses three different methodologies (SVM, HMM, and PSI-BLAST) to predict if a protein is a neurotoxin. SVM was implemented using the freely downloadable software package SVM_light with the RBF kernel. Neurotoxin function is predicted using N SVMs binary classifiers able to handle the multi-classification problem using 1 vs r (one against rest) strategy. The i^{th} SVM was trained with all samples in the i^{th} class with positive labels and the rest of the samples with negative labels. Five SVM modules were trained for the classification of neurotoxin according to their source into (i) eubacteria, (ii) cnidaria, (iii) mollusca, (iv) arthropoda and (v) chordata. Five further SVM modules were trained for the classification of neurotoxin according to their function into (i) ion channel blockers, (ii) acetylcholine receptor blockers, (iii) metalloproteolytic activity acetylcholine release inhibitors, (iv) phospholipase A₂ acetylcholine release inhibitors and, (v) acetylcholine release facilitators.

PSI-BLAST is used to align the query sequence against test dataset (Altschul et al., 1997). MEME (Multiple Em for Motif Elicitation) and MAST (Motif Alignment and Search Tool) were also used to compute structural features (motifs).

NTXPred is also based on a feed-forwarded neural networks (FNN) and a partial recurrent neural network (RNN) with a single hidden layer, implemented with free simulation packages SNNS, version 4.2, from Stuttgart University.

Figure 12: Submission form of NTXPred

Results: Table 29 reports the Cooper’s statistics declared in the primary research article.

Table 29: Performance of NTXPred methods in prediction of neurotoxins, as declared by Authors. PPV is the positive predictive value while MCC is the Matthew’s correlation coefficient.

Approach	Sensitivity	Specificity	PPV	Accuracy	MCC
FNN	0.90	0.79	0.88	0.84	0.69
RNN	0.89	0.96	0.96	0.93	0.86
SVM AA Composition (C)	0.96	0.97	0.98	0.98	0.94
SVM Dipeptide (D)	0.94	0.98	0.98	0.96	0.92

C + Length	0.98	0.97	0.97	0.97	0.95
D + Length	0.97	0.95	0.95	0.96	0.92
PSI-BLAST	0.98	0.95	NA	NA	NA
MEME/MAST	0.36	0.99	NA	NA	NA
C + MEME/MAST	0.97	0.97	0.97	0.97	0.94

A.1.4. PredCSF (Yong-Xian Fan et al., 2011)

Primary reference: Fan YX, Song J, Shen HB, Kong X. *PredCSF: an integrated feature-based approach for predicting conotoxin superfamily*. Protein Pept Lett. 2011 Mar;18(3):261-7.

Abstract: Conotoxins are small disulfide-rich peptides that are invaluable channel-targeted peptides and target neuronal receptors. They show prospects for being potent pharmaceuticals in the treatment of Alzheimer's disease, Parkinson's disease, and epilepsy. Accurate and fast prediction of conotoxin superfamily is very helpful towards the understanding of its biological and pharmacological functions especially in the post-genomic era. In the present study, we have developed a novel approach called PredCSF for predicting the conotoxin superfamily from the amino acid sequence directly based on fusing different kinds of sequential features by using modified one-versus-rest SVMs. The input features to the PredCSF classifiers are composed of physicochemical properties, evolutionary information, predicted secondary structure and amino acid composition, where the most important features are further screened by random forest feature selection to improve the prediction performance. The results show that PredCSF can obtain an overall accuracy of 90.65% based on a benchmark dataset constructed from the most recent database, which consists of 4 main conotoxin superfamilies and 1 class of non-conotoxin class. Systematic experiments also show that combining different features is helpful for enhancing the prediction power when dealing with complex biological problems. PredCSF is expected to be a powerful tool for in silico identification of novel conotoxins and is freely available for academic use at <http://www.csbio.sjtu.edu.cn/bioinf/PredCSF>.

Link: The web server is available at: <http://www.csbio.sjtu.edu.cn/bioinf/PredCSF>

PredCSF: An integrated feature-based approach for predicting conotoxin superfamily

[Read Me](#) | [Data](#) | [Citation](#) | [software](#) |

Step 1: Input the mature peptide of **one query protein (Example):**

Step 2: Input the full sequence of **one query protein (Example):**

Step 3: Input your email address:

Contact @ [Hong-Bin](#)

Figure 13: PredCSF homepage.

Citations: 18 citations

Field of application: Conotoxins

Training dataset: The mature peptides and the corresponding full sequences of conotoxins were extracted from the Swiss-Prot release 57.8 (released on 22-Sep-09). Because the number of entries in some superfamilies like P, S, J, L, D, V and C were less than 10 entries, too few to have statistical significance, these superfamilies were excluded. The I-conotoxin superfamily was not included because there are still some debates about the classification scheme of the I superfamily. [...] The remaining data set included 403 conotoxin sequences from A, M, O, and T superfamilies. To reduce the bias of sequence homology, the redundant sequences with pairwise sequence identity greater than 80% were excluded. The final data set was composed by 261 entries from four superfamilies: A (63 entries), M (48 entries), O (95 entries) and T (55 entries). Authors also added 60 short cysteine rich mature sequences of non-conotoxin sequences as a negative control data set. In the online Supporting information whole information about datasets used were provided.

Predicting methods: PredCSF uses different methodologies to predict if a protein is a conotoxin. PredCSF consider physicochemical information, discrete wavelet transforms, www.efsa.europa.eu/publications

EFSA Supporting publication 2024:EN-9063



position-Specific Scoring Matrix (PSSM) Information, Secondary Structure (SS) Information, Amino Acid Composition and Modified One-Versus-Rest SVMs.

The discrete wavelet transform is used to design feature vector elements that incorporates physicochemical properties of amino acids. The wavelet transform decomposes a signal into several groups (vectors) of coefficients in which different coefficient vectors contain information about the characteristics of the sequence at different scales. Coefficients at coarse scales capture global features of the proteins, whereas coefficients at fine scales contain local details.

PSSM was generated as matrix of Lx20 using PSI-BLAST (Altschul et al., 1997) to search the non-redundant protein sequence database through three iterations with 0.001 as the E-value cut-off for multiple sequence alignment against the sequence of the conotoxin.

SS information was retrieved by PSIPRED program (<http://bioinf.cs.ucl.ac.uk/psipred/>) and then protein was represented by a matrix. With this representation the secondary structure content ratios were computed for the whole protein chain.

The global information of mature peptides, represented by the amino acid composition (AAC), was also considered. AAC can be represented by a 20-dimensional vector, where each element denotes each amino acids occurrence in the whole sequence.

SVM was implemented using the LIBSVM package (Version 2.89) (<https://pypi.org/project/libsvm/>). The one-versus-rest (o-v-r) or one-versus-one (o-v-o) approaches were also implemented to decompose multiclass into a series of binary SVMs. This method includes the construction of each binary SVM classifier and five SVM classifiers, i.e. SVM-A specifically for the A type superfamily, SVM-M for the M superfamily, SVM-O for the O superfamily, SVM-T for the T superfamily and SVM-N for the negative control dataset were constructed.

Results: Table 30 and table 31 report the statistics declared in the primary research article.

Table 30: Predictive performance of the PredCF algorithm by the Jackknife Test for 321 mature peptides, as declared by Authors. A, M, O, T are the four conotoxin superfamilies, while N is the true negative.

	A (%)	M (%)	O (%)	T (%)	N (%)
Sn	84	94	94	94	87
Sp	91	96	90	93	85
MCC	85	94	88	92	83

Table 31: Predictive performance of PredCF using different feature subset. SS is the prediction of secondary structure, AAC the amino acid composition, PCP the wavelet features from physicochemical properties, PSSM the position specific scoring matrix. A, M, O, T are the four conotoxin superfamilies, while N is the true negative.

Features	Sn (%)					Sp (%)					Overall Acc
	A	M	O	T	N	A	M	O	T	N	
SS	52	8	92	95	68	66	44	71	60	76	68
AAC	63	65	82	56	68	67	58	71	67	77	69
PCP	78	92	83	91	68	84	90	75	93	76	82
PSSM	78	94	92	96	80	87	90	87	93	83	88
PSSM+SS	78	94	94	96	80	87	90	89	91	84	88
PSSM+SS+AAC	81	96	95	93	82	89	90	89	94	84	89
PSSM+SS+AAC+PCP	84	94	94	95	87	91	96	90	93	85	91

A.1.5. ToxinPred (Gupta et al., 2013)

Primary reference: Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, et al. (2013) *In Silico Approach for Predicting Toxicity of Peptides and Proteins*. PLoS ONE 8(9): e73957.

Abstract:

Background: Over the past few decades, scientific research has been focused on developing peptide/protein-based therapies to treat various diseases. With the several advantages over small molecules, including high specificity, high penetration, ease of manufacturing, peptides have emerged as promising therapeutic molecules against many diseases. However, one of the bottlenecks in peptide/protein-based therapy is their toxicity. Therefore, in the present study, we developed in silico models for predicting toxicity of peptides and proteins.

Description: We obtained toxic peptides having 35 or fewer residues from various databases for developing prediction models. Non-toxic or random peptides were obtained from SwissProt and TrEMBL. It was observed that certain residues like Cys, His, Asn, and Pro are abundant as well as preferred at various positions in toxic peptides. We developed models based on machine learning technique and quantitative matrix using various properties of peptides for predicting toxicity of peptides. The performance of dipeptide-based model in terms of accuracy was 94.50% with MCC 0.88. In addition, various motifs were extracted from the

toxic peptides and this information was combined with dipeptide-based model for developing a hybrid model. In order to evaluate the over-optimization of the best model based on dipeptide composition, we evaluated its performance on independent datasets and achieved accuracy around 90%. Based on above study, a web server, ToxinPred has been developed, which would be helpful in predicting (i) toxicity or non-toxicity of peptides, (ii) minimum mutations in peptides for increasing or decreasing their toxicity, and (iii) toxic regions in proteins.

Conclusion: ToxinPred is a unique in silico method of its kind, which will be useful in predicting toxicity of peptides/ proteins. In addition, it will be useful in designing least toxic peptides and discovering toxic regions in proteins. We hope that the development of ToxinPred will provide momentum to peptide/protein-based drug discovery (<http://crdd.osdd.net/raghava/toxinpred/>).

Link: The link reported in the primary reference is deprecated. Active link is the following: <https://webs.iiitd.edu.in/raghava/toxinpred/algo.php>.

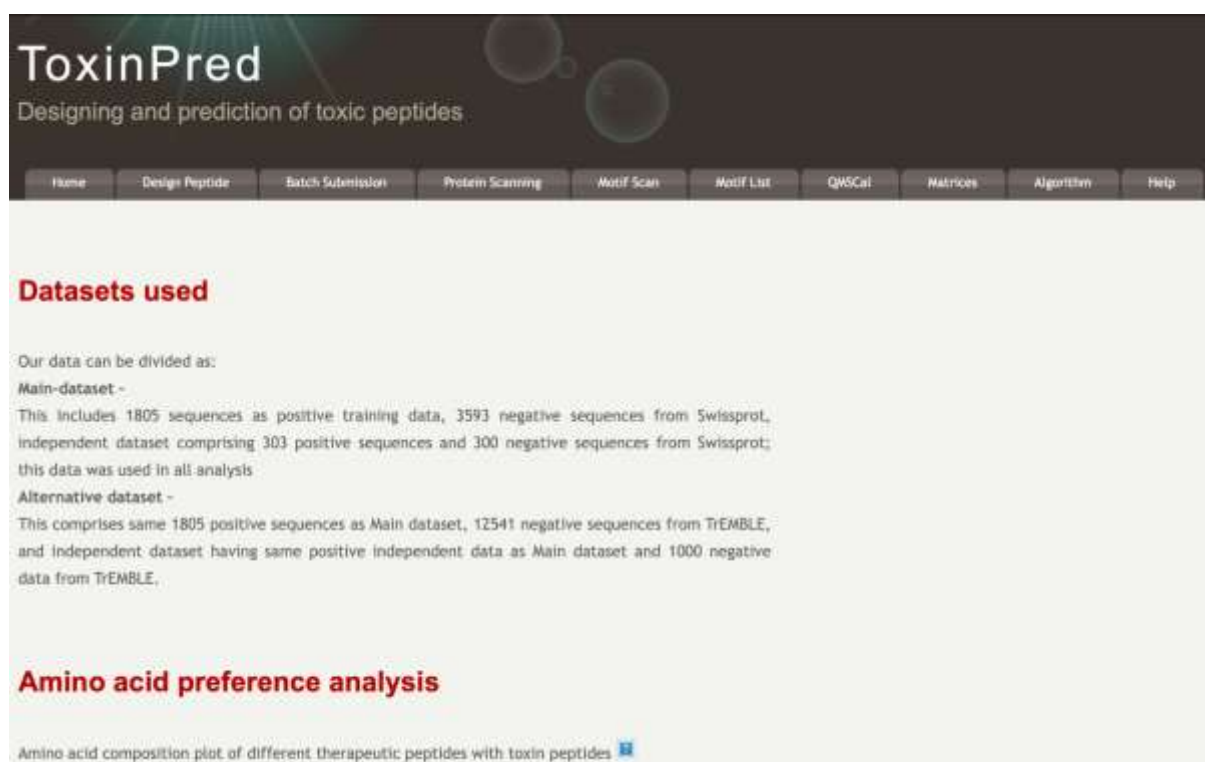


Figure 14: ToxinPred homepage.

Citations: 549 citations

Field of application: toxins

www.efsa.europa.eu/publications

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author(s).



Training dataset: The main dataset used for training and testing of ToxinPred was formed using experimentally validated toxic peptides (obtained from various databases) and well-annotated non-toxin peptides/proteins from UniProtKB/Swiss-Prot. Such dataset includes 1805 toxic peptides as positive examples and 3593 non-toxic peptides as negative examples. An alternate dataset formed with the same 1805 toxic-peptides/proteins and 12541 non-toxin peptides/proteins obtained from UniProtKB/TrEMBL. To evaluate biases in the performance of developed models, different independent datasets were created. The first includes 303 toxic proteins/peptides (called positive examples) and 300 non-toxic peptides/proteins or negative examples extracted from UniProtKB/SwissProt. None of the negative or positive examples was included in the main dataset. This dataset was referred as main independent dataset and used for evaluating models developed on the main dataset. Similarly, a second independent dataset was created in order to evaluate the performance of models developed on alternate dataset, this dataset consists of 303 positive examples extracted from UniProtKB/SwissProt and 1000 negative examples extracted from UniProtKB/TrEMBL, which were not included in the alternate dataset.

Predicting methods: ToxinPred implements SVM to predict if a peptide with primary structure length lower than 35 amino acids is a toxin using the freely downloadable software package SVM_light. To increase reliability, SVM classification based on amino acid composition and dipeptide composition were combined in a hybrid approach with the motif information. In this approach, first, various motifs are searched in the query peptides (based on MEME and two-sample logo software), and if any of the toxic motifs of toxic peptide is found, the SVM score is increased by the value of 5. This final score was used for the prediction.

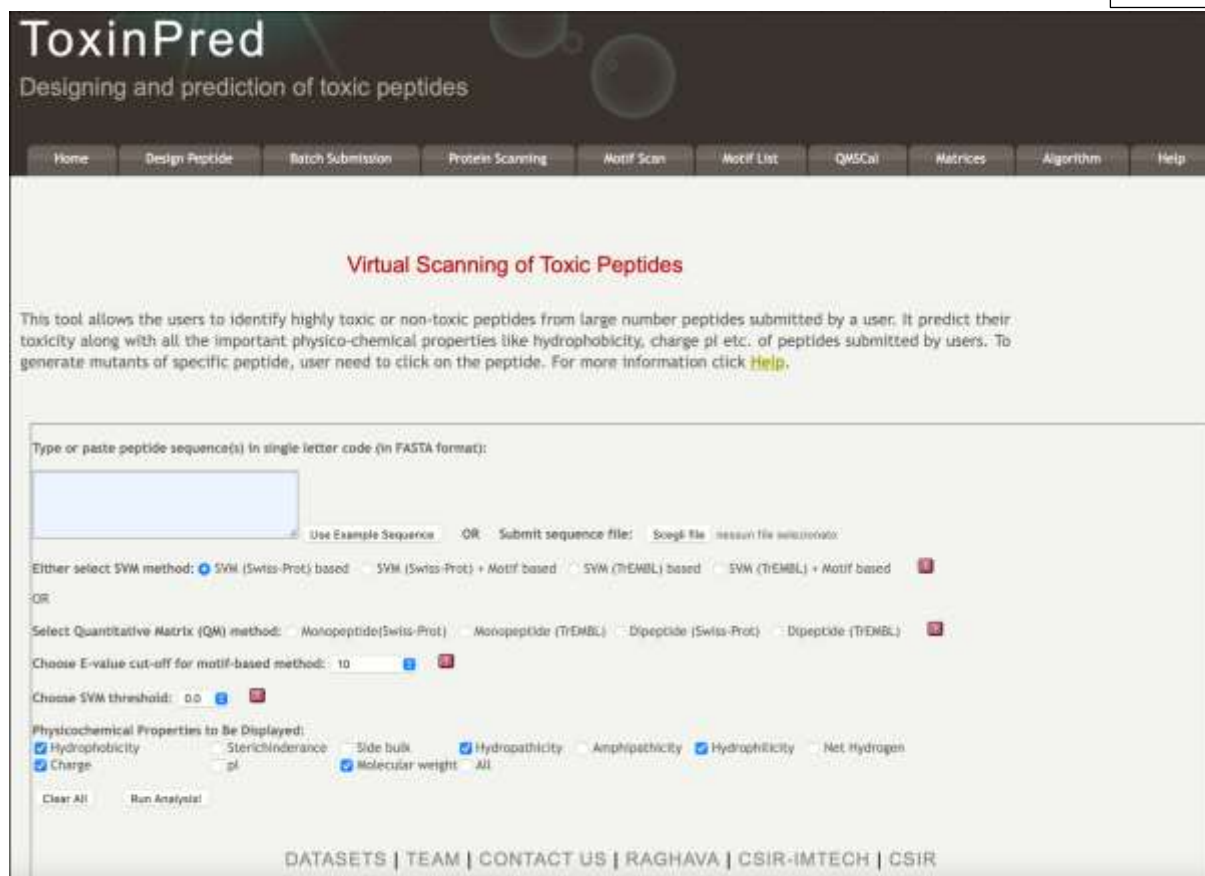


Figure 15: ToxinPred protein scanning tool.

Results: In Table 32 are reported the Cooper’s statistics declared in the primary research article.

Table 32: Performance of VMD-based models developed on main dataset using various type of composition like residue, dipeptide and terminal residues composition, as declared by Authors.

Features	Sensitivity	Specificity	Accuracy	MCC	AUC
AAC	0.93	0.94	0.94	0.87	0.97
C5AAC	0.84	0.84	0.84	0.65	0.88
C10AAC	0.89	0.92	0.91	0.80	0.94
N5AAC	0.82	0.81	0.81	0.59	0.88
N10AAC	0.90	0.87	0.88	0.75	0.94



DPC	0.94	0.95	0.94	0.88	0.98
-----	------	------	------	------	------

A.1.6. ToxClassifier (Gacesa et al., 2016)

Primary reference: Gacesa R, Barlow DJ, Long PF, Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. 2016, PeerJ Comput. Sci. 2:e90

Abstract: *Ascribing function to sequence in the absence of biological data is an ongoing challenge in bioinformatics. Differentiating the toxins of venomous animals from homologues having other physiological functions is particularly problematic as there are no universally accepted methods by which to attribute toxin function using sequence data alone. Bioinformatics tools that do exist are difficult to implement for researchers with little bioinformatics training. Here we announce a machine learning tool called 'ToxClassifier' that enables simple and consistent discrimination of toxins from non-toxin sequences with >99% accuracy and compare it to commonly used toxin annotation methods. 'ToxClassifier' also reports the best-hit annotation allowing placement of a toxin into the most appropriate toxin protein family, or relates it to a non-toxic protein having the closest homology, giving enhanced curation of existing biological databases and new venomics projects. 'ToxClassifier' is available for free, either to download (<https://github.com/rgacesa/ToxClassifier>) or to use on a web-based server (<http://bioserv7.bioinfo.pbf.hr/ToxClassifier/>).*

Link: The link to the web server temporary unavailable (<http://bioserv7.bioinfo.pbf.hr/ToxClassifier/>), while the link to the github source code is freely accessible: (<https://github.com/rgacesa/ToxClassifier>).

Citations: 26 citations

Field of application: toxins contained into animal venom

Training dataset: Four different databases were used selecting protein from UniProtKB/SwissProt as follows:

- 1) "Positive" dataset was extracted from UniProtKB/Swiss-Prot_ToxProt (Jungo et al., 2012) using the search query: taxonomy: "Metazoa [33208]" (keyword:toxin OR annotation:(type: "tissue specificity" venom)). All duplicate entries with identical sequence or sequence identifier were removed, resulting in 8,093 sequences.
- 2) "Easy negative" dataset was obtained by random sampling of 50,000 sequences in UniProtKB/SwissProt database (Bateman et al., 2017), washing out duplicates. Final dataset included 47,144 protein sequences.
- 3) "Moderate difficulty" negative dataset was designed to match highly curated toxin-like proteins with physiological function; it was created by BLASTp searching UniProtKB/SwissProt database with Positive dataset, with e-value cut-off of 1.0e-10, resulting in 8,034 proteins.
- 4) "Hard negative" dataset was constructed from UniProtKB/TrEMBL database (Bairoch and Apweiler, 2000) instead of Swiss-Prot. As with Moderate dataset, it was created from www.efsa.europa.eu/publications



results of BLASTp using Positive dataset as query and UniProtKB/TrEMBL as target database. Duplicates and sequences also occurring in Positive, Easy or Moderate datasets were removed for total of 7,403 sequences.

Predicting methods: Models describing protein sequences were constructed as follows:

- (1) Single Amino acid frequency model (OF): model uses length of sequence and frequency of each amino acid as input features.
- (2) Amino acid dimer frequency model (BIF): model uses length of sequence, frequency of each amino acid and of each amino acid 2-mer.
- (3) Naivetox-bits model (NTB): input features for this model are the number of "tox-bits" for each 'tox-bits' HMM listed in the "tox-bits" database (Starcevic et al., 2015).
- (4) Scored "tox-bits" model (STB): STB is a modification of the NTB model, with HMM bit-scores replacing the number of 'tox-bits' in each 'tox-bit' HMM model.
- (5) Tri-Blast Simple (TBS) model: TBS uses BLASTp searches against positive (UniProtKB/SwissProt-ToxProt) and two negative control databases (close non-toxins from UniProtKB/SwissProt and non-toxins from UniProtKB/TrEMBL); features include bit-score, query length, subject length, query/subject length ratio, query coverage, percentage of identity, percentage of positive matches; features also include amino-acid frequencies. Scores are computed from the 'best-hit' in each database, with a BLAST e-value of $1.0e-10$.
- (6) Tri-Blast Enhanced A (TBEa) model: TBEa model is an expanded variant of TBS, with amino dimer frequencies included in the model.
- (7) Tri-Blast Enhanced B (TBEb): model is a variation of TBEa, trained on 80% of the input dataset and with a BLAST e-value cut-off value of $1.0e+3$ for the detection of similar toxin or non-toxic sequences.

Support Vector Machine (SVM), Gradient Boosted Machine (GBM) and Generalised Linear Model (GLM) classifiers were trained for each of the models. Annotation models simulating manual annotation were constructed based on BLAST and HMMER

Results: In Table 33 are reported the Cooper's statistics declared in the primary research article.

Table 33: Prediction accuracy on positive and negative datasets, as well as range of measurements calculated for all test data, as declared by Authors. Annotation models used as classifier inputs either: the frequency of amino acids (TBSim) or combinations of two amino-acids (BIF); the presence of absence or 'Tox-Bits' (SToxA); HMM scores for 'ToxBits' (SToxB); a selection of BLAST output co-variants (TBEa); a variation on TBSim and TBEa (TBEb). Classifier Learning Machines used were: Gradient Boosted (GBM), Support Vector (SVM) and Generalised Linear Model (GLM). The datasets were a 'Positive' control containing only validated animal toxins, an 'Easy' dataset composed of non-toxin sequences, a 'Moderate' dataset comprising curated non-toxin sequences but with homology to 'Positive' sequences, and a 'Hard' dataset that included all sequences from the 'Moderate' dataset, together with un-curated sequences also with homology to 'Positive' sequences.

Annotation model	Classifier	Accuracy (Positive toxin dataset)	Accuracy (Easy non-toxin dataset)	Accuracy (Moderate non-toxin dataset)	Accuracy (Hard non-toxin dataset)
TBSim	GBM	0.80	0.99	0.98	0.92
	SVM	0.80	1	0.98	0.94
	GLM	0.55	0.99	0.96	0.84
BIF	GBM	0.83	1	0.98	0.94
	SVM	0.89	1	0.98	0.96
	GLM	0.71	0.99	0.99	0.91
SToxA	GVM	0.64	1	0.98	0.94
	SVM	0.84	1	0.96	0.91
SToxB	GBM	0.75	1	0.99	0.93
	SVM	0.85	1	0.99	0.92
	GLM	0.03	1	1	0.99
TBEa	GBM	0.88	1	1	0.99
	SVM	0.93	1	1	0.97
	GLM	0.96	1	1	0.94
TBEb	GBM	0.82	1	1	1
	SVM	0.96	1	1	0.97
	GLM	0.93	1	1	0.99

A.1.7. NNTox (Jain and Kihara, 2019)

Primary reference: Jain A, Kihara D. NNTox: Gene Ontology-Based Protein Toxicity Prediction Using Neural Network. *Sci Rep* 9, 17923 (2019).

Abstract: *With advancements in synthetic biology, the cost and the time needed for designing and synthesizing customized gene products have been steadily decreasing. Many research laboratories in academia as well as industry routinely create genetically engineered*

proteins as a part of their research activities. However, manipulation of protein sequences could result in unintentional production of toxic proteins. Therefore, being able to identify the toxicity of a protein before the synthesis would reduce the risk of potential hazards. Existing methods are too specific, which limits their application. Here, we extended general function prediction methods for predicting the toxicity of proteins. Protein function prediction methods have been actively studied in the bioinformatics community and have shown significant improvement over the last decade. We have previously developed successful function prediction methods, which were shown to be among top-performing methods in the community-wide functional annotation experiment, CAFA. Based on our function prediction method, we developed a neural network model, named NNTox, which uses predicted GO terms for a target protein to further predict the possibility of the protein being toxic. We have also developed a multi-label model, which can predict the specific toxicity type of the query sequence. Together, this work analyses the relationship between GO terms and protein toxicity and builds predictor models of protein toxicity.

Link: <https://github.com/kiharalab/NNTox>

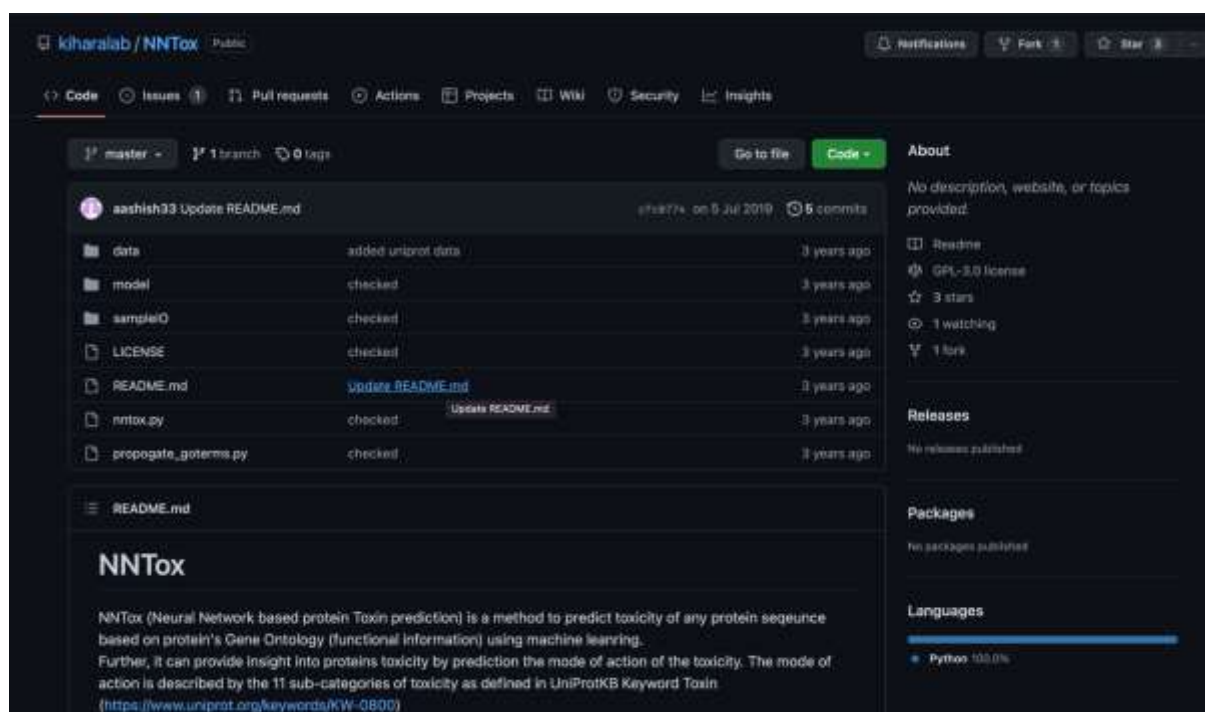


Figure 16: GitHub web page for NNTox

Citations: 6 citations

Field of application: toxins

Training dataset: A non-redundant dataset containing 488 toxin proteins was assembled with the following procedure. The UniProtKB/Swiss-Prot database was queried for the keyword "Toxin" (UniProtKB KW-0800) resulting in 6,497 entries which were then filtered according to sequence similarity to remove redundancy. Finally, GO Annotations were collected in a set to www.efsa.europa.eu/publications



characterise toxicity at GO level. Another non-redundant dataset of 6,594 non-toxin proteins was also collected from UniProtKB/Swiss-Prot selecting entries not tagged with the keyword “Toxin”, and with 95% of GO terms belonging to the toxin GO term set. Again, sequence similarity was used to remove redundancy. A third dataset of the mode of action was assembled with 270 non-redundant toxins divided in 11 sub-classes: cardiotoxin, enterotoxin, neurotoxin, ion channel impairing toxin, myotoxin, dermonecrotic toxin, hemostasis impairing toxin, G-protein coupled receptor impairing toxin, complement system impairing toxin, cell adhesion impairing toxin, and viral exotoxin.

Predicting methods: A five-layer fully connected feedforward neural network is used for the toxin/non-toxin prediction, with an input layer of 2,596 neurons representing the GO term feature vector. Then three further layers with 200 neurons each were used to feed either a SoftMax binary classifier or a cross entropy multi-label classifier.

Neural network was trained with backpropagation using the ADAM optimizer, implemented in TensorFlow. A five-fold nested cross validation was performed to tune four hyper-parameters: the number of neurons in hidden layer [10, 50, 100, 200, 500], the regularization strength [10, 1, 0.1, 0.01, 0.001], the learning rate [10, 1, 0.1, 0.01, 0.001] and the number of epochs [100, 500, 1000, 2000, 5000].

PFP⁴ was also used to predict the protein function for a toxin. Briefly, PFP uses PSI-BLAST (Altschul et al., 1997) to retrieve similar sequences from a database to a query sequence and obtains GO-term annotations from the sequences with an E-value of up to 125. Then, each GO term will be assigned with a score that reflects the E-value of sequences that have the GO term in their annotation as well as the conditional probability that the GO term occurs given other GO terms are observed. PFP-predicted GO terms were also used to NNTox neural network.

Results: In table 34 are reported the statistics declared in the primary research article.

Table 34: Summary of the toxin prediction, as declared by Authors.

Method	Precision	Recall	F1 score
With GO annotation			
Baseline exact	0.03	0.63	0.05
Baseline 1 mismatch	0.02	0.71	0.04
Baseline 1 mismatch	0.02	0.77	0.04
NNTox (GO Annotation)	0.90	0.90	0.90
With PFP prediction			

⁴ PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. s

Baseline exact	0.11	0.16	0.13
Baseline 1 mismatch	0.10	0.18	0.13
Baseline 1 mismatch	0.12	0.26	0.16
PFP	0.87	0.55	0.66
NNTox (PFP)	0.80	0.75	0.78
PFP + NNTox (PFP)	0.81	0.78	0.79

A.1.8. TOXIFY (Cole and Brewer, 2019)

Primary reference: Cole TJ, Brewer MS, TOXIFY: a deep learning approach to classify animal venom proteins. 2019, PeerJ 7:e7200

Abstract: *In the era of Next-Generation Sequencing and shotgun proteomics, the sequences of animal toxigenic proteins are being generated at rates exceeding the pace of traditional means for empirical toxicity verification. To facilitate the automation of toxin identification from protein sequences, we trained Recurrent Neural Networks with Gated Recurrent Units on publicly available datasets. The resulting models are available via the novel software package TOXIFY, allowing users to infer the probability of a given protein sequence being a venom protein. TOXIFY is more than 20X faster and uses over an order of magnitude less memory than previously published methods. Additionally, TOXIFY is more accurate, precise, and sensitive at classifying venom proteins.*

Link: <https://github.com/tijeco/toxify>

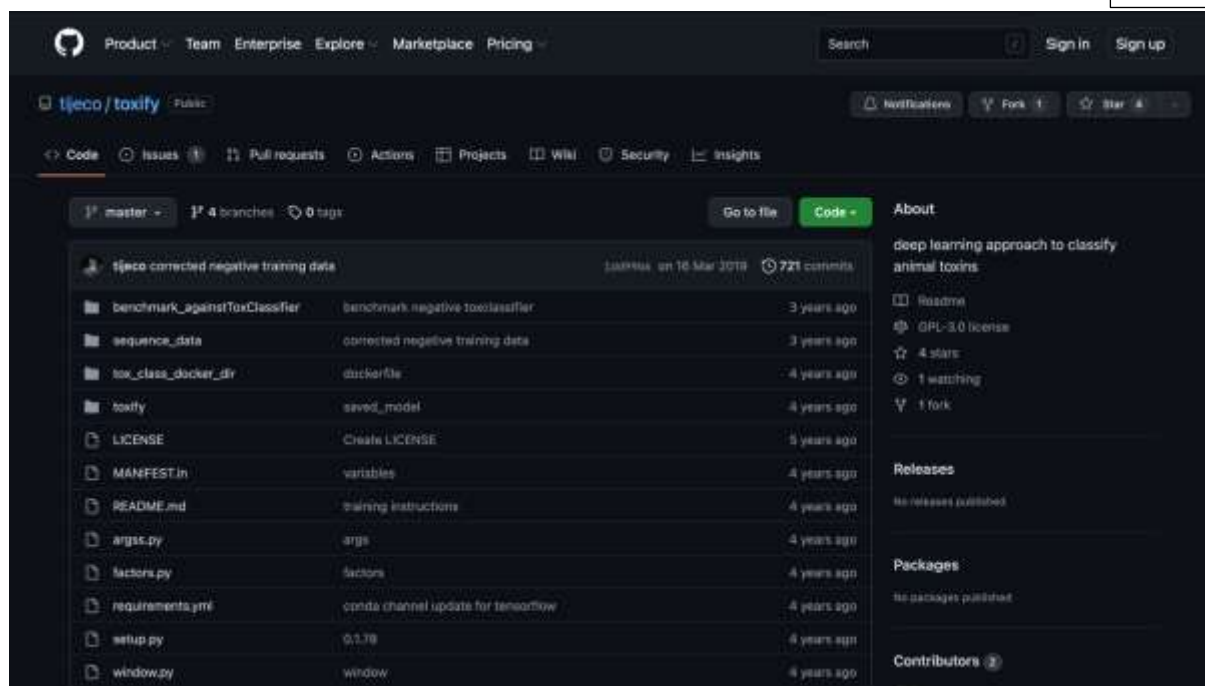


Figure 17: GitHub web page for TOXIFY

Citations: 6 citations

Field of application: venom toxins

Training dataset: To allow for a proper comparison between ToxClassifier and TOXIFY, the training datasets for TOXIFY comprised only protein sequences from UniProtKB that were uploaded/available prior to June 2016, when ToxClassifier was published. Protein sequences uploaded to UniProtKB between June 2016 and October 2018 were not included as training data for either ToxClassifier or TOXIFY and were used as benchmark comparisons between the two methods. Datasets were obtained using the following two procedures.

To train models to classify venom proteins, the training sets were constrained to only include verified venom proteins from UniProtKB/Swiss-Prot. This dataset, referred henceforth as “positive”, was constructed using the following search terms (annotation:(type:“tissue specificity” venom)). This resulted in a total of 6,133 venom protein sequences.

“Negative” data sets comprised 50,000 random, non-venom proteins from UniProtKB/Swiss-Prot using the following search term (NOT annotation:(type:“tissue specificity” venom) AND reviewed:yes).

Due to venom proteins generally being low mass and relatively short (e.g., <30 amino acids), only proteins containing ≤ 500 amino acids were included in the final training dataset. This brought the size of the positive dataset down to a total of 4,808 proteins and the negative dataset to 32,391 proteins. Training data consisted of a random 80% subset of the positive and negative sequences, and the remaining 20% was set aside for model validation. Less

than 5% of the dataset contained sequence redundancy, which is an artifact of the databasing procedure in UniProtKB/Swiss-Prot.

Predicting methods: Recurrent Neural Networks (RNN) are an ideal tool for classifying ordered sets of items, such as amino acid sequences, because they specify hidden states that depend on the input as well as the prior hidden state. Gated Recurrent Units (GRUs) are a high-performing RNN that have gained popularity since being introduced by (Cho et al., 2014), due to faster performance over traditional Long Short-Term Memory approaches. Using TensorFlow v1.8.0 libraries as the back-end (Abadi et al., 2016), Authors constructed a venom protein classifier using GRU with 270 hidden units with a learning rate of 0.01. Training occurred for 50 epochs, and the training accuracy and training loss in accuracy (from a logit cost function) were recorded at every 2nd epoch. The trained model was then used to calculate the probability that a given protein should be classified as an animal venom.

Results: In table 35 are reported the statistics declared in the primary research article.

Table 35: Performance of TOXIFY as declared by Authors.

Method	Acc	Spec	Sens	Bacc	NPV	PPV	F1	MCC
TOXIFY	0.86	0.96	0.76	0.86	0.80	0.95	0.85	0.74

A.1.9. ToxDL (Pan et al., 2020)

Primary reference: Pan X, Zuallaert J, Wang X, Shen HB, Campos EP, Marushchak DO, De Neve W. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*. 2021 Jan 29;36(21):5159-5168. doi: 10.1093/bioinformatics/btaa656. PMID: 32692832.

Abstract:

Motivation: Genetically engineering food crops involves introducing proteins from other species into crop plant species or modifying already existing proteins with gene editing techniques. In addition, newly synthesized proteins can be used as therapeutic protein drugs against diseases. For both research and safety regulation purposes, being able to assess the potential toxicity of newly introduced/synthesized proteins is of high importance.

Results: In this study, we present ToxDL, a deep learning-based approach for in silico prediction of protein toxicity from sequence alone. ToxDL consists of (i) a module encompassing a convolutional neural network that has been designed to handle variable-length input sequences, (ii) a domain2vec module for generating protein domain embeddings and (iii) an output module that classifies proteins as toxic or non-toxic, using the outputs of the two aforementioned modules. Independent test results obtained for animal proteins and cross-species transferability results obtained for bacteria proteins indicate that ToxDL

outperforms traditional homology-based approaches and state-of-the-art machine-learning techniques. Furthermore, through visualizations based on saliency maps, we are able to verify that the proposed network learns known toxic motifs. Moreover, the saliency maps allow for directed in silico modification of a sequence, thus making it possible to alter its predicted protein toxicity.

Availability and implementation: ToxDL is freely available at <http://www.csbio.sjtu.edu.cn/bioinf/ToxDL/>. The source code can be found at <https://github.com/xypan1232/ToxDL>.

Supplementary information: Supplementary data are available at Bioinformatics online.

Link: ToxDL is freely available at <http://www.csbio.sjtu.edu.cn/bioinf/ToxDL/>. The source code can be found at <https://github.com/xypan1232/ToxDL>.

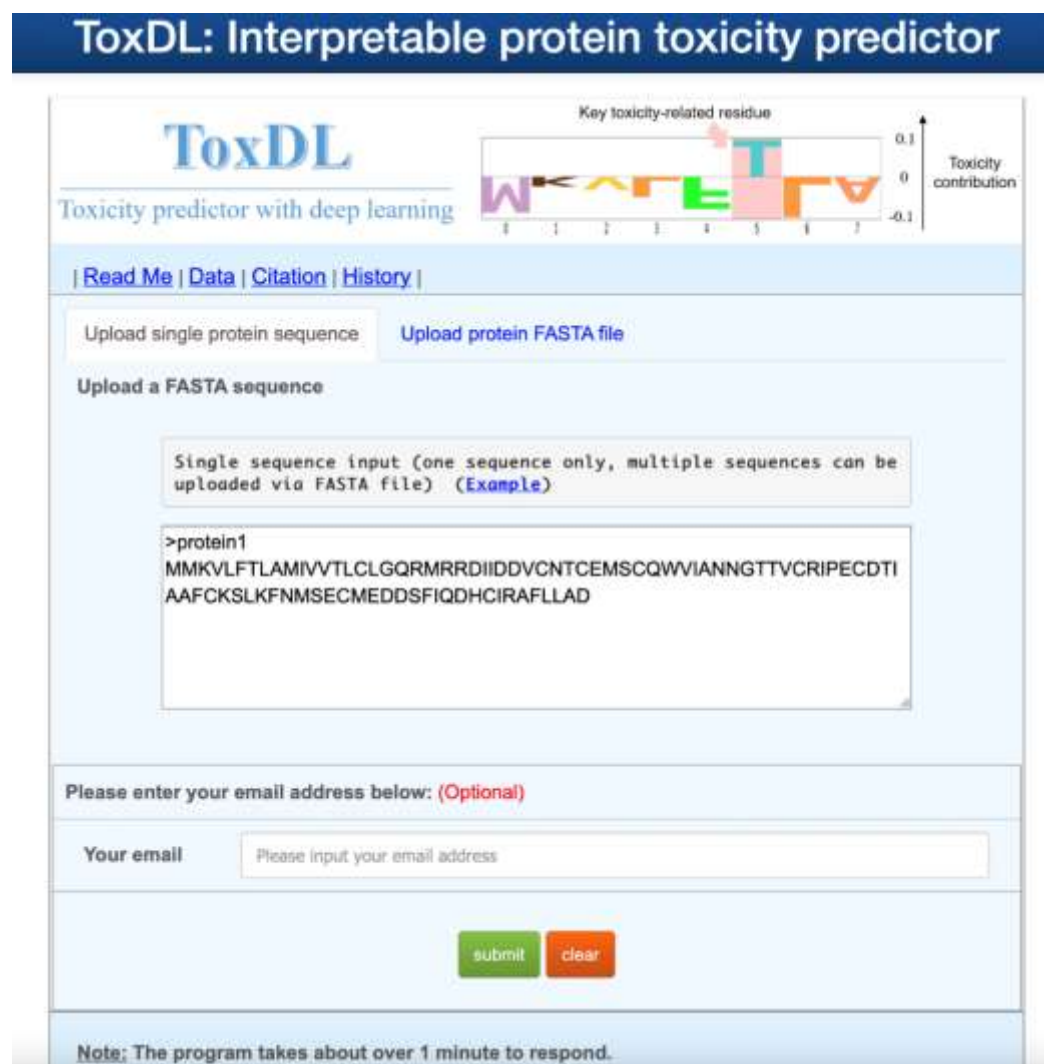


Figure 18: TOXDL homepage.



Citations: 9 citations

Field of application: toxins

Training dataset: All the 6164 proteins annotated as toxin in the Animal Toxin Annotation Project (Jungo and Bairoch, 2005) found in UniProtKB in May 2019 were used as positive samples. As negative samples, 6164 proteins and 903 venom proteins were randomly extracted from the same species of positive samples that are not flagged as toxic in UniProtKB/Swiss-Prot. Training set was created randomly selecting the 80% of the positive (toxin) and negative (non-toxin) samples, using the remaining 20% for creating the validation set and an independent test set. d-hit-2d (Fu et al., 2012) was used to remove redundant sequences from the test set with similarity threshold of at least 40% (the minimum value for cd-hit-2d). In the same fashion homologous sequences were held out from the validation set. This process resulted in an initial validation set of 309 non-toxic proteins and 25 toxic proteins, and an initial test set of 754 non-toxic proteins and 59 toxic proteins. Pfam clans (El-Gebali et al., 2019) were used to ensure the absence of proteins with domains from the same Pfam clans between the test and validation sets. resulting in a test set of 59 toxic proteins and 670 non-toxic proteins, and a validation set of 25 toxic proteins and 277 non-toxic proteins. Since the positive set is extracted from the Animal Toxin Annotation Project, a further bacteria test set, using the test set from BTXPred (Saha and Raghava, 2007b), with 183 toxic proteins and 500 non-toxic proteins was considered. Again cd-hit-2d was applied to remove similarity with a cutoff of 40%.

Predicting methods: ToxDL tool uses a multimodal deep learning-based approach for predicting protein toxicity. The output from a CNN module with the average embeddings of all domains found in a protein, is fed into an output component that generates a toxicity probability. In detail, the CNN module of ToxDL takes a one-hot encoded protein sequence as input, and subsequently performs convolutional, dropout and max pooling operations the output of which is then fed to a specialized layer to deal with the variable length of the input sequences, exploring five different approaches. After concatenating the output of the CNN module with the averaged domain embedding vector, ToxDL transfers the resulting vector to the output component, which consists of a fully connected layer, a dropout layer and a softmax output unit. The five approaches to deal with variable input lengths are: zero-padding, global max pooling, Gated Recurrent Unit (GRU), dynamic max pooling and finally, a dynamic k-max pooling layer (Kalchbrenner et al., 2014) was added after the last max pooling layer, resulting in a fixed output size. Instead of keeping one value after max pooling, dynamic k-max pooling collects the k highest activations in each channel in the same order of occurrence. Information about protein domains is integrated into ToxDL, via InterProScan (Jones et al., 2014) training a Skip-gram model to automatically learn protein domain embeddings, for a total of 269 resulting domains. The HMM models from Pfam v32.0 (El-Gebali et al., 2019) and were also used in different baseline methods.

Finally, MEME (Bailey et al., 2015) and TOMTOM (Gupta et al., 2007) were used to generate toxic motifs. Four variants of ToxDL can be built:

ToxDL-ODE: This variant only uses the 256-D protein domain embeddings as its input, which is then directly connected to the output component.

ToxDL-CNN: This variant only uses the CNN module.

www.efsa.europa.eu/publications



EFSA Supporting publication 2024:EN-9063

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author(s).



ToxDL-One: Instead of using learned embeddings for representing protein domains, we use a one-hot encoding for the 269 toxic protein domains. Specifically, each protein is represented using a 269-D binary vector, with a one indicating the presence of a particular domain. This vector is directly fed to the output component.

ToxDL-OD: For this variant, the one-hot encoded vectors for the 269 toxic protein domains, as described for the ToxDL-One variant, are concatenated with the output of the CNN module. This combination is then fed to the output component.

Results were compared with various baseline methods: BLAST, BLAST-score, InterProScan, hmmsearch, ToxinPred, ClanTox and TOXIFY

Results: In table 36 are reported the Cooper's statistics declared in the primary research article.

Table 36: Performance of ToxDL on the animal protein test set as declared by Authors.

Methods	F1 score	MCC	auROC	auPRC
ToxDL-One	0.34	0.44	0.61	0.57
ToxDL-OD	0.77	0.75	0.98	0.85
ToxDL-ODE	0.60	0.50	0.95	0.65
ToxDL-CNN	0.76	0.74	0.98	0.85
ToxDL	0.81	0.79	0.99	0.91

A.1.10. ToxIBTL (Wei et al., 2022)

Primary reference: Wei L, Ye X, Sakurai T, Mu Z, Wei L, *ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning*, Bioinformatics, Volume 38, Issue 6, 15 March 2022, Pages 1514–1524.

Abstract:

Motivation: Recently, peptides have emerged as a promising class of pharmaceuticals for various diseases treatment poised between traditional small molecule drugs and therapeutic proteins. However, one of the key bottlenecks preventing them from therapeutic peptides is their toxicity toward human cells, and few available algorithms for predicting toxicity are specially designed for short-length peptides.

Results: We present ToxIBTL, a novel deep learning framework by utilizing the information bottleneck principle and transfer learning to predict the toxicity of peptides as well as proteins. Specifically, we use evolutionary information and physicochemical properties of peptide sequences and integrate the information bottleneck principle into a feature representation learning scheme, by which relevant information is retained and the redundant information is minimized in the obtained features. Moreover, transfer learning is introduced to transfer the common knowledge contained in proteins to peptides, which aims to improve the feature



representation capability. Extensive experimental results demonstrate that ToxIBTL not only achieves a higher prediction performance than state-of-the-art methods on the peptide dataset, but also has a competitive performance on the protein dataset. Furthermore, a user-friendly online web server is established as the implementation of the proposed ToxIBTL.

Availability and implementation: The proposed ToxIBTL and data can be freely accessible at <http://server.wei-group.net/ToxIBTL>. Our source code is available at <https://github.com/WLYLab/ToxIBTL>.

Supplementary information: Supplementary data are available at Bioinformatics online.

Link: ToxIBTL and data can be freely accessible at <http://server.wei-group.net/ToxIBTL>. The source code is available at <https://github.com/WLYLab/ToxIBTL>.

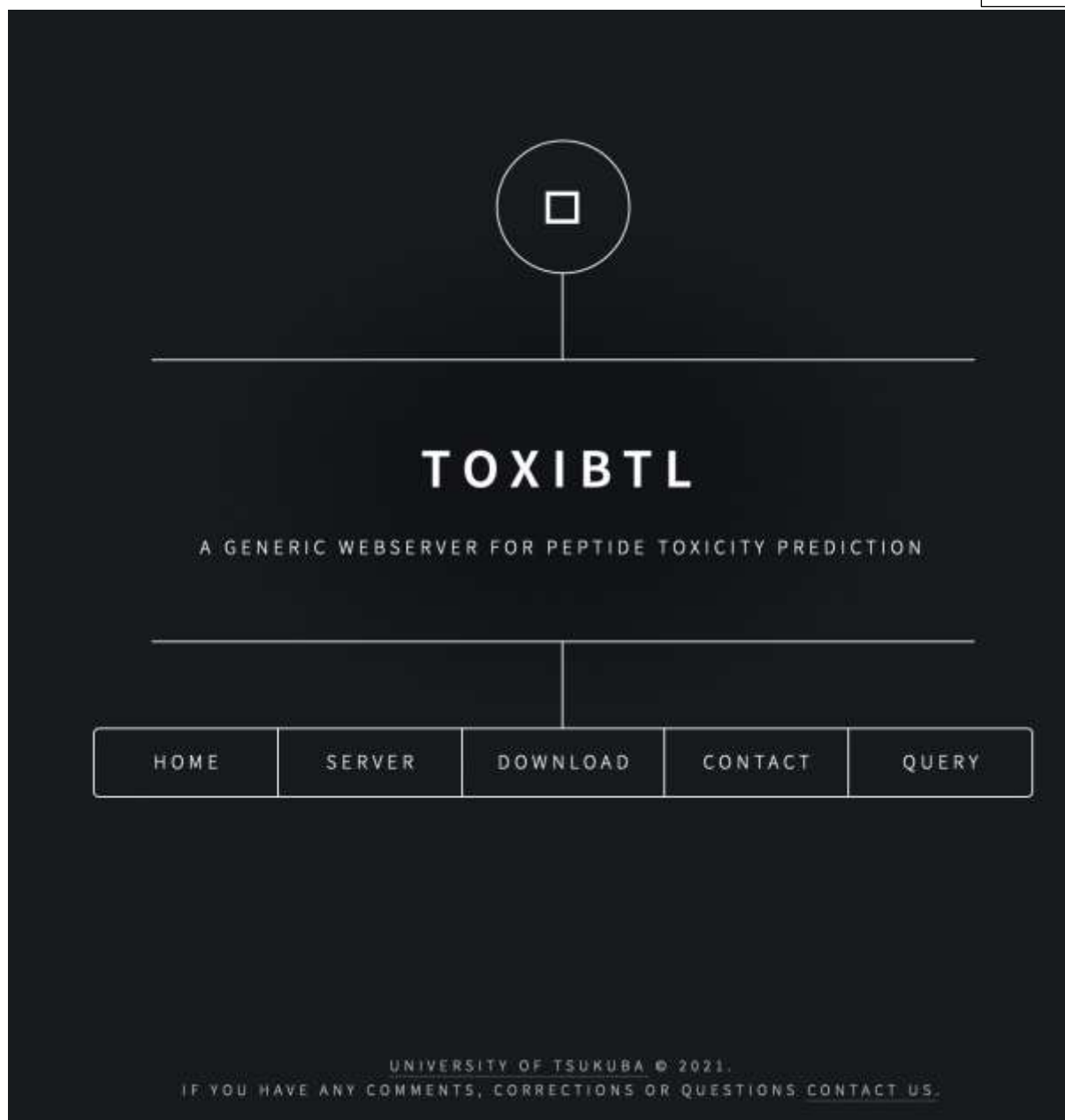


Figure 19: ToxIBTL homepage.

Citations: 1 citation

Field of application: toxins

Training dataset: Two datasets were used to train ToxIBTL. The first, established by (Pan et al., 2020) (ToxDL), was employed to build models for predicting protein toxicity. It contains 4,472 toxic animal proteins used as positive samples and 6,341 non-toxic animal proteins used as negative samples. Each sequence in the testing set has a similarity < 40% to that in the training set, meanwhile, there are no protein sequences with the same domain from the www.efsa.europa.eu/publications



Pfam clans (El-Gebali et al., 2019) between these two sets. The second benchmark dataset created in Author's previous work (Wei et al., 2021) (ATSE) was used to build models for peptide toxicity prediction, and consists of 3,864 samples with a range of 10–50 residues. The positive samples in this dataset are toxic peptide sequences, which are experimentally validated. Similarly, the negative samples are non-toxic peptide sequences, which have the same number as the positive ones. The sequence similarity between any two peptide sequences is less than 90%, which can avoid the evaluation bias introduced by sequence similarity. For training, about 85% of toxic and non-toxic peptides are randomly selected to fine-tune our model for predicting the toxicity of the peptide, and the remaining peptides are adopted as testing set to evaluate the performance of the fine-tuned model.

Predicting methods: The workflow of ToxIBTL mainly contains three steps, namely sequence encoding, optimization and classification. In the first step, to encode the evolutionary information, raw sequences are converted to evolutionary profiles and fed into the hybrid network CNN_BiGRU to automatically capture latent local and global information; simultaneously, to capture physicochemical information, raw sequences are sent into the FECS model to obtain graphical features and statistical features. In the second step, Authors directly concatenated the evolutionary and physicochemical features are directly concatenated and optimized by means of and used the information bottleneck principle to optimize the concatenated features. In the third step, the optimized features are used to determine the sequence as toxic or non-toxic one. A model was initially trained on the protein dataset and then used to fine-tune a new model on the peptide set.

The standard BLOSUM62 scoring matrix was used to encode peptide (or protein) sequences and the information derived from this process used to design a hybrid network. This network is called CNN_BiGRU and consists of CNN and BiGRU and it can effectively capture the contextual and semantic information of peptide (or protein) sequences. Specifically, the BLOSUM62 matrix of a peptide (or protein) sequence is fed into a 2D convolutional layer with a non-linear activation function (e.g. relu) to extract the local correlation between amino acids through the local perceptual domain. Afterward, the output of the convolutional layer is taken as the input of the BiGRU layer to obtain the long and short dependency information amongst extracted local correlation and capture sequence-order effects (Li et al., 2017).

To better represent each peptide (or protein) sequence to encompass the perspective of its biophysical and biochemical properties, FECS, a feature extraction model of protein sequence using the physicochemical properties of amino acids and statistical information of protein sequences was introduced to extract the graphical and statistical features of peptide (or protein) sequences.

For the graphical feature encoding, 158 physicochemical properties of amino acids are effectively used to transform a peptide (or protein) sequence into a 158-dimensional numerical vector, which are selected from the AAindex database (Kawashima and Kanehisa, 2000). First, the 20 amino acids are ranked in ascending order according to their physicochemical indices. Second, the ranked 20 amino acids are sequentially positioned on the circumference of the bottom of a right circular cone of height. Subsequently, 400 amino acid pairs are arranged on the underside of the right circular cone. Then, the 3D graphical curve S of the sequence P is obtained.

To represent a peptide (or protein) sequence, Authors linearly combine the evolutionary features extracted from CNN-BiGRU network with the physicochemical features from FECS model.

Results: In table 37 are reported the Cooper's statistics declared in the primary research article.

Table 37: Performance of ToxIBDL and on the toxin test set as declared by Authors.

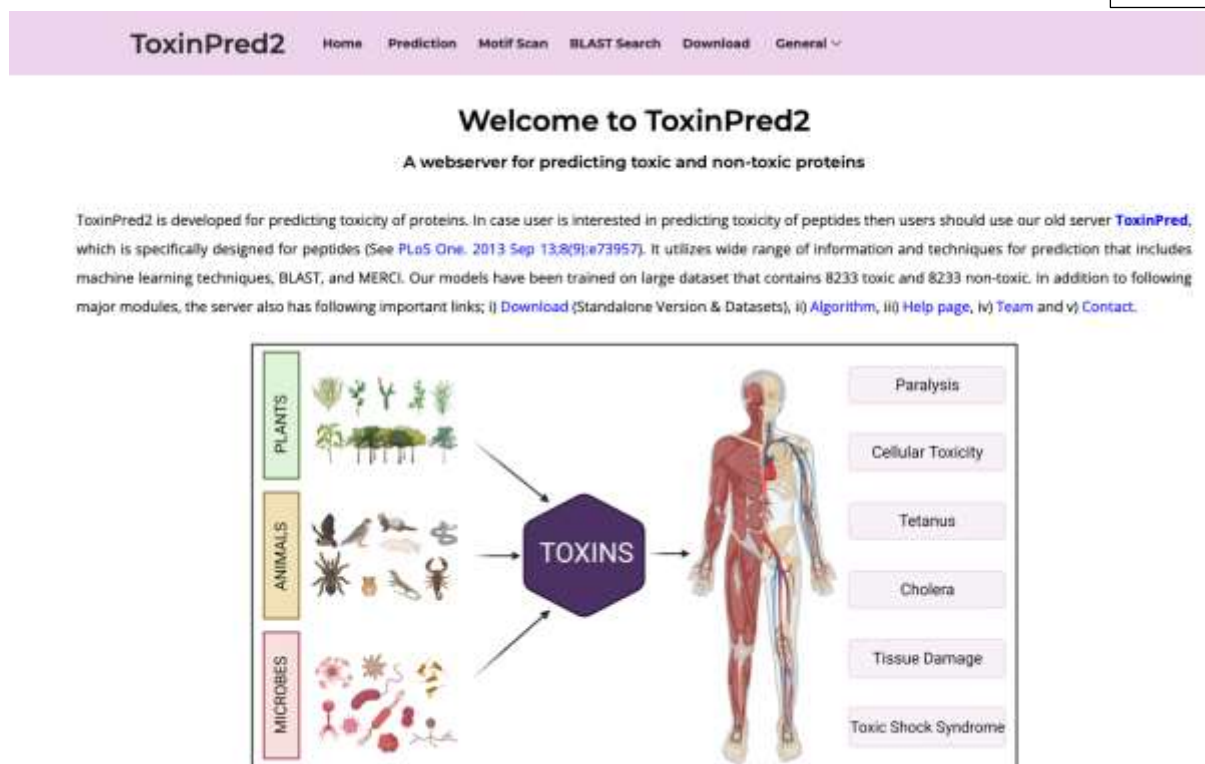
Methods	F1 score	MCC	auROC	auPRC
ToxIBTL	0.83	0.82	0.99	0.91

A.1.11. ToxinPred2 (Sharma et al., 2022)

Primary reference: Sharma N, Naorem LD, Jain S, Raghava GPS. *ToxinPred2: an improved method for predicting toxicity of proteins*. Brief Bioinform. 2022 May 21:bbac174.

Abstract: *Proteins/peptides have shown to be promising therapeutic agents for a variety of diseases. However, toxicity is one of the obstacles in protein/peptide-based therapy. The current study describes a web-based tool, ToxinPred2, developed for predicting the toxicity of proteins. This is an update of ToxinPred developed mainly for predicting toxicity of peptides and small proteins. The method has been trained, tested and evaluated on three datasets curated from the recent release of the SwissProt. To provide unbiased evaluation, we performed internal validation on 80% of the data and external validation on the remaining 20% of data. We have implemented the following techniques for predicting protein toxicity; (i) Basic Local Alignment Search Tool-based similarity, (ii) Motif-Emerging and with Classes-Identification-based motif search and (iii) Prediction models. Similarity and motif-based techniques achieved a high probability of correct prediction with poor sensitivity/coverage, whereas models based on machine-learning techniques achieved balance sensitivity and specificity with reasonably high accuracy. Finally, we developed a hybrid method that combined all three approaches and achieved a maximum area under receiver operating characteristic curve around 0.99 with Matthews correlation coefficient 0.91 on the validation dataset. In addition, we developed models on alternate and realistic datasets. The best machine learning models have been implemented in the web server named 'ToxinPred2', which is available at <https://webs.iiitd.edu.in/raghava/toxinpred2/> and a standalone version at <https://github.com/raghavagps/toxinpred2>. This is a general method developed for predicting the toxicity of proteins regardless of their source of origin.*

Link: The web server is available at <https://webs.iiitd.edu.in/raghava/toxinpred2/>, while the source code is available at <https://github.com/raghavagps/toxinpred2> .



ToxinPred2 Home Prediction Motif Scan BLAST Search Download General

Welcome to ToxinPred2

A webserver for predicting toxic and non-toxic proteins

ToxinPred2 is developed for predicting toxicity of proteins. In case user is interested in predicting toxicity of peptides then users should use our old server [ToxinPred](#), which is specifically designed for peptides (See *PLoS One*. 2013 Sep 13;8(9):e73957). It utilizes wide range of information and techniques for prediction that includes machine learning techniques, BLAST, and MERCI. Our models have been trained on large dataset that contains 8233 toxic and 8233 non-toxic. In addition to following major modules, the server also has following important links; i) [Download](#) (Standalone Version & Datasets), ii) [Algorithm](#), iii) [Help page](#), iv) [Team](#) and v) [Contact](#).

PLANTS
ANIMALS
MICROBES

TOXINS

Paralysis
Cellular Toxicity
Tetanus
Cholera
Tissue Damage
Toxic Shock Syndrome

Figure 20: ToxinPred2 homepage.

Citations: 0 – since it is not already available in Scopus.

Field of application: toxins

Training dataset: The dataset was retrieved from UniProt release 2021_03 (released on 2 June 2021) (Bateman et al., 2021) using different keywords for obtaining toxic and non-toxic proteins. 9,940 toxic proteins were extracted using the keyword 'toxin AND reviewed: yes'. All protein sequences comprising 'BJOUXZ', <35 amino acids and non-toxic sequences like toxic sequences were discarded, resulting in 8233 toxic sequences, which is referred to as a positive dataset. The negative dataset was extracted from UniProtKB/Swiss-Prot (Bateman et al., 2017) using keywords 'NOT toxin NOT allergen AND reviewed: yes' resulting in 554,145 proteins, from which the sequences with length <35 amino acids and with non-standard characters were discarded for a final number of 460,257 non-toxic sequences.

To remove sequence redundancy, CD-HIT software (Fu et al., 2012) was then applied to both datasets at 40% sequence identity resulting in a positive dataset reduced of 1924 sequences and a negative dataset reduced to 88,263 sequences from 460,257. Three datasets were assembled as follows:

- (a) Main Dataset: this dataset contains 8,233 toxic (not CD-HIT filtered) and 8,233 non-toxic (randomly selected among the 88,263 negative and CD-HIT filtered) protein sequences. In this dataset positive sequences are redundant.



(b) Alternate Dataset: this dataset contains 1,924 toxic (CD-HIT filtered) and 1,924 non-toxic (randomly selected from among the 88,263 negative and CD-HIT filtered) non-redundant protein sequences. In this dataset, no two proteins have >40% sequence similarity.

Realistic Dataset (the same as alternate but with 10 times Negative Dataset): this dataset consists of 1,924 toxic and 19,240 non-toxic protein sequences.

Predicting methods: Several machine learning techniques are used to discriminate toxic from non-toxic proteins. Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), DT, k-nearest neighbours (KNNs), XGBoost (XGB) and SVC were implemented to develop the classification models. These classifiers were optimized using various hyperparameters, and the best results were included. The prediction is based also on BLAST, Motif-Emerging and with Classes-Identification (MERCİ) tool, and Pfeature.

Results: Table 38 reports statistics declared in the primary research article.

Table 38: Performance of ToxIBDL motif-based approach on main dataset when combined with machine learning-based models developed using AAC, as declared by Authors.

ML	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
RF	0.84	0.89	0.87	0.74	0.85	0.88	0.86	0.73
SVC	0.85	0.82	0.84	0.68	0.85	0.80	0.82	0.65
XGB	0.83	0.84	0.84	0.68	0.82	0.83	0.83	0.65
KNN	0.82	0.83	0.82	0.65	0.82	0.84	0.83	0.66
DT	0.75	0.81	0.78	0.56	0.74	0.83	0.79	0.58
LR	0.74	0.79	0.76	0.53	0.74	0.77	0.76	0.51
GNB	0.72	0.79	0.76	0.52	0.73	0.79	0.76	0.52

The literature search reported also about three other tools, namely ClanTox, ATSE and SpiderP but none of these are accessible anymore (web page unavailable and source code not provided anywhere).

In addition to the predictive methods identified and described above, toxin databases were identified; these are not considered as “predictive tools” since they check the query against an internal datasets. For completeness of information these are reported below.

A.1.12. ConoServer (Kaas et al., 2012, 2008)

Primary references:

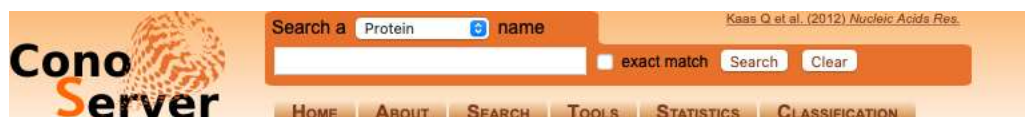


- 1) Kaas Q, Yu R, Jin AH, Dutertre S and Craik DJ. *ConoServer: updated content, knowledge, and discovery tools in the conopeptide database*. *Nucleic Acids Research* (2012) 40:D325-30
- 2) Kaas Q, Westermann JC, Halai R, Wang CK and Craik DJ. *ConoServer, a database for conopeptide sequences and structures*. *Bioinformatics* (2008) 24(3):445-6

Abstract:

- 1) *ConoServer* (<http://www.conoserver.org>) is a database specializing in the sequences and structures of conopeptides, which are toxins expressed by marine cone snails. Cone snails are carnivorous gastropods, which hunt their prey using a cocktail of toxins that potently subvert nervous system function. The ability of these toxins to specifically target receptors, channels and transporters of the nervous system has attracted considerable interest for their use in physiological research and as drug leads. Since the founding publication on *ConoServer* in 2008, the number of entries in the database has nearly doubled, the interface has been redesigned and new annotations have been added, including a more detailed description of cone snail species, biological activity measurements and information regarding the identification of each sequence. Automatically updated statistics on classification schemes, three-dimensional structures, conopeptide-bearing species and endoplasmic reticulum signal sequence conservation trends, provide a convenient overview of current knowledge on conopeptides. Transcriptomics and proteomics have begun generating massive numbers of new conopeptide sequences, and two dedicated tools have been recently implemented in *ConoServer* to standardize the analysis of conopeptide precursor sequences and to help in the identification by mass spectrometry of toxins whose sequences were predicted at the nucleic acid level.
- 2) *Summary: ConoServer is a new database dedicated to conopeptides, a large family of peptides found in the venom of marine snails of the genus Conus. These peptides have an exceptional diversity of sequences and chemical modifications and their ability to block ion channels makes them important as drug leads and tools for physiological studies. ConoServer uses standardized names and a genetic and structural classification scheme to present data retrieved from SwissProt, GenBank, the Protein DataBank and the literature. The ConoServer web site incorporates specialized features like the graphic display of post-translational modifications that are extensively present in conopeptides. Currently, ConoServer manages 1214 nucleic sequences (from 54 Conus species), 2258 proteic sequences (from 66 Conus species) and 99 3D structures.*

Link: <http://www.conoserver.org>



ConoServer, a database for conopeptides

ConoServer is a database specializing in the sequence and structures of conopeptides, which are peptides expressed by carnivorous marine cone snails. A fascinating feature of these peptides is their high specificity and affinity towards human ion channels, receptors and transporters of the nervous system. This makes conopeptides an interesting resource for the physiological studies of neuroreceptors and promising drug leads. Conopeptides are further described [here](#) and a selection of recent reviews on the subject can be found [here](#).



Classifications. Conopeptides are classified into disulfide rich (conotoxins) and several classes of disulfide poor peptides. The three classification schemes used in ConoServer, the gene superfamilies, the cysteine frameworks, and the pharmacological families are described and analyzed [here](#).



Post-translational modifications. Conopeptides are heavily post-translationally modified and the list of modifications found naturally or introduced artificially is provided [here](#).

Statistics. Statistics on the currently known conopeptides are provided [here](#). The statistics include: the relationship between conopeptide classification schemes, the sequence consensus between signal peptides of each gene superfamily, the number of entries for each cone snail species, the characteristics of conopeptides for which a three-dimensional structure was determined, the number of patented sequences, and the journals the most cited in the ConoServer.

	NUMBER OF ENTRIES	SPECIES COVERAGE	LINKS TO UNIPROT, GENBANK OR PDB
NUCLEOTIDES	2986	90	1814
PROTEINS	8364	123	4817
STRUCTURES	232	48	232

ConoNews

2021-11-11
The ConoPrec tools now features a comparison with disulfide poor representatives to predict the conopeptide class.

2021-11-10
All user file uploads are now limited to 50 Mb, which should be enough for any type of reasonable data sets (1,000 sequences in fasta format should be < 1Mb!).

2021-11-01
ConoServer is now hosted on a different webserver. Please do not hesitate to contact us if you notice something is not working properly (email: q.kaas@imb.uq.edu.au).

Figure 21: ConoServer homepage.

Citations: Reference 1: 240 citations; Reference 2: 163 citations

Field of application: conotoxins

Dataset: Since 2008, ConoServer is able to manage 1,214 nucleic sequences (from 54 Conus species), 2,258 protein sequences (from 66 species) and 99 3D structures. The protein sequences are split into 450 mature peptides, 615 prepro-peptides, 34 synthetic peptides and 1,159 sequences from patents. The 427 mature conotoxins are splitted into superfamilies as follows: 133 O, 104A, 58M, 51T, 31 I, 7 L, 6P, 6J, 6P, 3 D, 2S and 1G superfamily peptides.



The sequences and structures of conopeptides were extracted from public databases, GenBank (Benson et al., 2007), UniProtKB/Swiss-Prot (Boeckmann et al., 2003) and the Protein Data Bank (Berman et al., 2002), and by an extensive survey of the literature.

In September 2011, Authors updated ConoServer database. At the state of the art, ConoServer provides information on 1,180 mature conopeptides and contains information on 338 synthetic variants. ConoServer catalogues 95 three-dimensional structures of wild type conopeptides and 42 structures of synthetic variants. Finally, ConoServer describes 1,288 patented protein and 737 patented nucleic acid sequences. The sequences of 1,120 precursors are currently in ConoServer and 16 gene superfamilies are described.

Methods: ConoServer allows searches of nucleic acids, proteins and 3D structures of conopeptides based on their name, patent ID, sub-sequence, FASTA alignment, mass range, peptide mass fragments (fingerprints), classification, type (mature peptide, prepro-peptide, synthetic peptide or patent) and species. The name search simultaneously uses standard names and related names (historical names, non-standard names, trade names). Moreover, ConoServer allows comparison of sequences of entries selected from a result list. The sequences can be aligned with CLUSTALW (Larkin et al., 2007) and the alignment analysed with an residue-based colour scheme or with a LOGO representation (Schneider and Stephens, 1990) or with a distance tree computed with protdist and dnadist from the PHYLIP package.

A.1.13. KNOTTIN (Gracy et al., 2008; Postic et al., 2018)

Primary references:

- 1) Gelly JC, Gracy J, Kaas Q, LeNguyen D, Heitz A and Chiche L. *The KNOTTIN website and database: a new information system dedicated to the knottin scaffold*. Nucleic Acids Res., 2004;32, D156–D159.
- 2) Postic G, Gracy J, Périn C, Chiche L, Gelly JC. *KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling*. Nucleic Acids Res. 2018 Jan 4;46(D1):D454-D458.

Abstract:

- 1) *The KNOTTIN website and database organize information about knottins or inhibitor cystine knots, small disulfide-rich proteins with a knotted topology. Thanks to their small size and high stability, knottins provide appealing scaffolds for protein engineering and drug design. Static pages present the main historical and recent results about knottin discoveries, sequences, structures, folding, functions, applications and bibliography. Database searches provide dynamically generated tabular reports or sequence alignments for knottin three-dimensional structures or sequences. BLAST/HMM searches are also available. A simple nomenclature, based on loop lengths between cysteines, is proposed and is complemented by a uniform numbering scheme. This standardization is applied to all knottin structures in the database, facilitating comparisons. Renumbered and structurally fitted knottin PDB files are available for download. The standardized numbering is used for automatic*



drawing of two-dimensional Colliers de Perles. The KNOTTIN website and database are available at <http://knottin.cbs.cnrs.fr> and <http://knottin.com>.

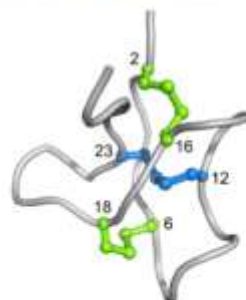
- 2) *Knottins, or inhibitor cystine knots (ICKs), are ultra-stable miniproteins with multiple applications in drug design and medical imaging. These widespread and functionally diverse proteins are characterized by the presence of three interwoven disulfide bridges in their structure, which form a unique pseudoknot. Since 2004, the KNOTTIN database (www.dsimb.inserm.fr/KNOTTIN/) has been gathering standardized information about knottin sequences, structures, functions and evolution. The website also provides access to bibliographic data and to computational tools that have been specifically developed for ICKs. Here, we present a major upgrade of our database, both in terms of data content and user interface. In addition to the new features, this article describes how KNOTTIN has seen its size multiplied over the past ten years (since its last publication), notably with the recent inclusion of predicted ICKs structures. Finally, we report how our web resource has proved usefulness for the researchers working on ICKs, and how the new version of the KNOTTIN website will continue to serve this active community.*

Link: <https://www.dsimb.inserm.fr/KNOTTIN/>

Welcome to the KNOTTIN database

What are knottins?

- Knottins are small disulfide-rich proteins characterized by a very special "disulfide through disulfide knot"
- This knot is achieved when one disulfide bridge crosses the macrocycle formed by the two other disulfides and the interconnecting backbone.
- The knot implies that knottins contain at least 3 disulfide bridges.
- The structural family of knottins have the disulfide between cysteines III and VI (blue) going through disulfides I-V and II-V (green).
- The growth factor cystine knots also contain a knot but the connectivity is different and they cannot be superimposed onto knottins. These proteins belong to a distinct structural family not described in this site.
- Knottins are sometime referred to as "inhibitor Cystine Knots".



Updates

- June 2017
Major update of the KNOTTIN database
+ Addition of theoretical 3D models
+ Visualization with Jmol
+ New design of the web interface
- February 2014
The KNOTTIN database is available
We had to stop the server for technical reasons. It is now available again.
Please let us know about any trouble you could encounter with this server.
Also note that we are currently working on the next update of the database.
Thank you for using KNOTTIN!

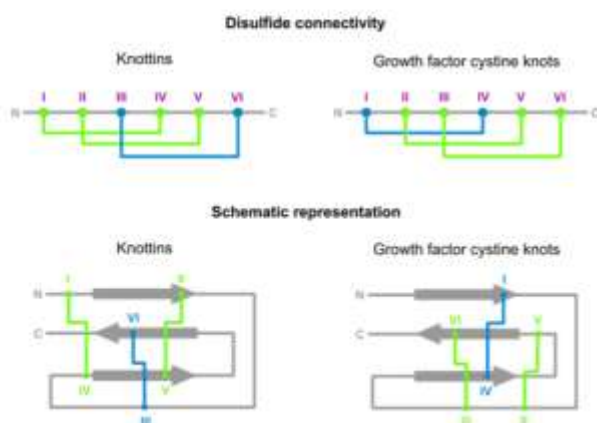


Figure 22: KNOTTIN homepage.

Citations: Reference 1: 91; Reference 2: 44

Field of application: Knottin.

Training dataset: Homologs of known knottins were searched in the SwissProt/TrEMBL database using BLAST (22) and HMMER (23) programs with low cut-offs followed by manual elimination of irrelevant hits. Cross-links between PDB IDs and SwissProt IDs were manually checked and extended when possible. Data are stored in several tables in a MySQL relational database management system. The first version (2004) of the KNOTTIN database contained 85 3D structures and 385 sequences of knottins, while with the 2018 update KNOTTIS grown up to 214 3D structures and 3,320 sequences of knottins.

Methods: KNOTTIN uses:

- BLAST/HMMER a sequence against the knottin database;

www.efsa.europa.eu/publications

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author(s).



- ii) STRIDE and PDBgeo for looking for structural properties such as torsion angles, secondary structures, solvent accessibility);
- iii) SWISS-MODEL and Mod-Base for modelling 3D unknown structures.

A.1.14. DBETH (Chakraborty et al., 2012)

Primary reference: Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S. *DBETH: a Database of Bacterial Exotoxins for Human*. Nucleic Acids Res. 2012 Jan;40(Database issue):D615-20.

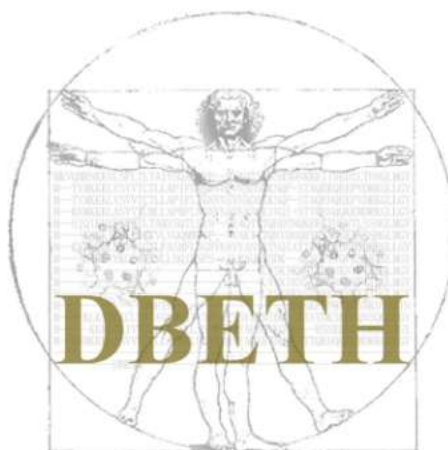
Abstract: *Pathogenic bacteria produce protein toxins to survive in the hostile environments defined by the host's defense systems and immune response. Recent progresses in high-throughput genome sequencing and structure determination techniques have contributed to a better understanding of mechanisms of action of the bacterial toxins at the cellular and molecular levels leading to pathogenicity. It is fair to assume that with time more and more unknown toxins will emerge not only by the discovery of newer species but also due to the genetic rearrangement of existing bacterial genomes. Hence, it is crucial to organize a systematic compilation and subsequent analyses of the inherent features of known bacterial toxins. We developed a Database for Bacterial ExoToxins (DBETH, <http://www.hpppi.iicb.res.in/btox/>), which contains sequence, structure, interaction network and analytical results for 229 toxins categorized within 24 mechanistic and activity types from 26 bacterial genres. The main objective of this database is to provide a comprehensive knowledgebase for human pathogenic bacterial toxins where various important sequence, structure and physico-chemical property based analyses are provided. Further, we have developed a prediction server attached to this database which aims to identify bacterial toxin like sequences either by establishing homology with known toxin sequences/ domains or by classifying bacterial toxin specific features using a support vector based machine learning techniques.*

Link: <http://www.hpppi.iicb.res.in/btox/>

Citations: 47 citations.



Database of Bacterial ExoToxins for Human is a database of sequences, structures, interaction networks and analytical results for 229 exotoxins, from 26 different human pathogenic bacterial genus. All toxins are classified into 24 different Toxin classes. The aim of DBETH is to provide a comprehensive database for human pathogenic bacterial exotoxins.



DBETH also provides a platform to its users to identify potential exotoxin like sequences through Homology based as well as Non-homology based methods. In homology based approach the users can identify potential exotoxin like sequences either running BLASTp against the toxin sequences or by running HMMER against toxin domains identified by DBETH from human pathogenic bacterial exotoxins. In Non-homology based part DBETH uses a machine learning approach to identify potential exotoxins (Toxin Prediction by Support Vector Machine based approach).

Figure 23: DBETH homepage.

Field of application: Bacterial ExoToxins

Training dataset: DBETH contains sequence, structure, interaction network information and analytical results for 229 toxins categorized within 24 mechanistic and activity types from 26 pathogenic bacterial genres. A total of 305 experimentally validated three dimensional (3D) structures and 55 *in silico* modeled 3D structures are available at DBETH database. Authors provide in Supplementary materials the complete data collection process.

Methods: DBETH aim is to identify the potential toxin sequences. The server is divided into two sub parts; the first part includes 'Homology based' toxin identification, which aims to identify toxin specific domains within a given protein sequence using HMMER (Potter et al., 2018) derived Hidden Markov Model (HMM) profile matching against a toxin domain HMM profile database. The HMM profile dataset is created by running the exotoxins against six different domain database including Pfam-A (Finn et al., 2016), Pfam-B, CDD (Marchler-Bauer et al., 2017), COG (Tatusov et al., 2001), SMART (Carnate and Ed, 2008) and TIGR (Chan et al., 2006). Users can also search their sequences against the toxin protein sequences and their homologues using conventional BLAST (Altschul et al., 1990) searching procedure. Users can also search a protein structure against the available DBETH structure database. Structural

www.efsa.europa.eu/publications



alignment using Mustang_v3.2.1 (Konagurthu et al., 2006) enables the user to identify structural similarity within a protein against toxin structures.

The second part of DBETH server includes a 'Non-Homology' based approach where a SVM (Wong et al., 2013) based method is employed to identify potential bacterial toxins. A total of 298 features based on peptide (di-peptide and tri-peptide) frequencies and combinations along with frequencies of amino acids' physico-chemical property groups were calculated to characterize the positive (toxins) and negative (non-toxins) samples. LibSVM (Chang and Lin, 2011) was used to build the classifier models. A training dataset comprising of 180 bacterial toxins and 1800 non-toxins (1:10 ratio for positive and negative sample) were developed to train the model using svm-train program of the LibSVM package. A Radial basis kernel function (RBF) has been used via a 10-fold cross validation of the training set to obtain the optimized gamma (0.5) and C parameter (2.0). Further a feature selection protocol was implemented to remove the possible redundant features from original feature set.

A.1.15. T1TAdb (Tourasse and Darfeuille, 2021)

Primary reference: Tourasse NJ, Darfeuille F. *T1TAdb: the database of type I toxin-antitoxin systems*. RNA. 2021 Dec;27(12):1471-1481.

Abstract: *Type I toxin-antitoxin (T1TA) systems constitute a large class of genetic modules with antisense RNA (asRNA)-mediated regulation of gene expression. They are widespread in bacteria and consist of an mRNA coding for a toxic protein and a noncoding asRNA that acts as an antitoxin preventing the synthesis of the toxin by directly basepairing to its cognate mRNA. The co- and post-transcriptional regulation of T1TA systems is intimately linked to RNA sequence and structure, therefore it is essential to have an accurate annotation of the mRNA and asRNA molecules to understand this regulation. However, most T1TA systems have been identified by means of bioinformatic analyses solely based on the toxin protein sequences, and there is no central repository of information on their specific RNA features. Here we present the first database dedicated to type I TA systems, named T1TAdb. It is an open-access web database (<https://d-lab.arna.cnrs.fr/t1tadb>) with a collection of ~1,900 loci in ~500 bacterial strains in which a toxin-coding sequence has been previously identified. RNA molecules were annotated with a bioinformatic procedure based on key determinants of the mRNA structure and the genetic organization of the T1TA loci. Besides RNA and protein secondary structure predictions, T1TAdb also identifies promoter, ribosome-binding, and mRNA-asRNA interaction sites. It also includes tools for comparative analysis, such as sequence similarity search and computation of structural multiple alignments, which are annotated with covariation information. To our knowledge, T1TAdb represents the largest collection of features, sequences, and structural annotations on this class of genetic modules.*

Link: <https://d-lab.arna.cnrs.fr/t1tadb>

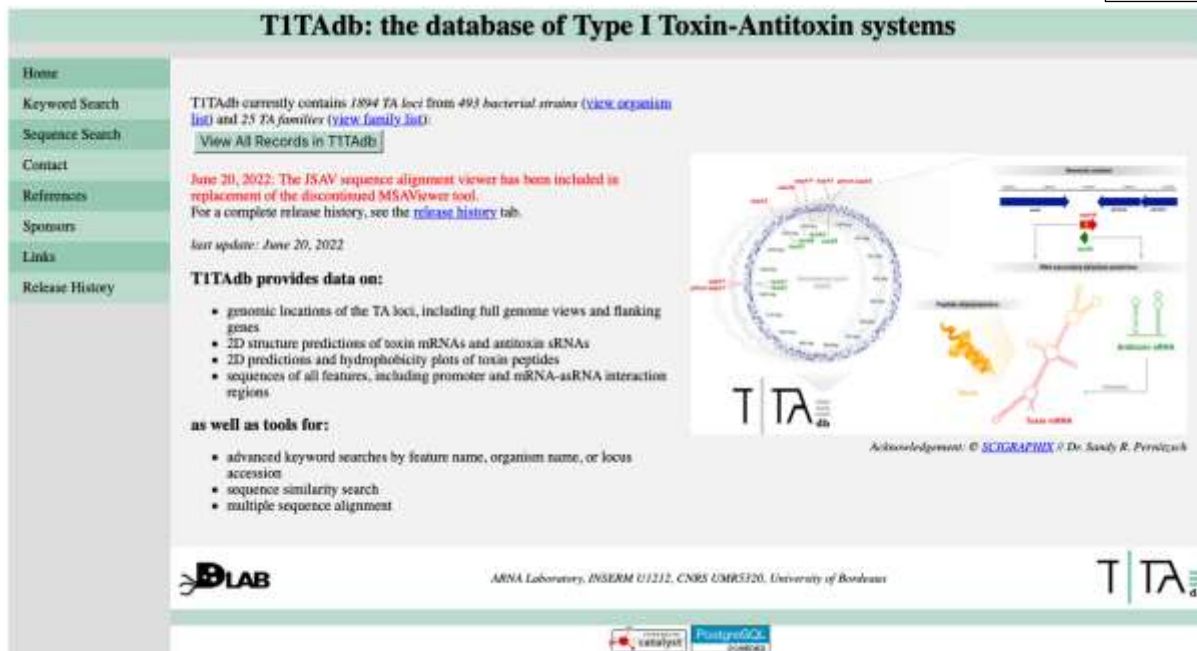


Figure 24: T1TAdb homepage.

Citations: 2 citations.

Field of application: Type I Toxin-Antitoxin System.

Training dataset: The majority of the data in T1TAdb are based on the genome-wide bioinformatic searches by (Fozo et al., 2010) who identified sequences of ORFs coding for type I toxin peptides of known and novel families in hundreds of bacterial strains. For toxins belonging to known TA families, computational analyses were performed to locate the coordinates of the toxin mRNA and antitoxin asRNA to identify the complete TA locus corresponding to each reported toxin ORF.

Methods: Secondary structures of RNA were predicted using MFOLD 3.6 (Zuker, 2003, 1989) and annotated diagrams highlighting the location of specific motifs (start/stop codon, SD sequence, interaction region) were generated with VARNA 3.93 (Darty et al., 2009). Secondary structures of toxin peptides were predicted using PSIPRED 4.02 (McGuffin et al., 2000) run with PSI-BLAST 2.2.26 against the UniRef90 protein sequence database (<https://www.uniprot.org/help/uniref>) and drawn with POLYVIEW-2D (Porollo et al., 2004). Hydrophobicity plots were computed with ProtScale (Gasteiger et al., 2005). Interactive genomic maps in SVG format showing the localizations of TA loci were drawn using CGView (Stothard and Wishart, 2005). Sequence similarity searches in T1TAdb are done with BLAST+ 2.2.31 (Camacho et al., 2009) and multiple sequence alignments are computed by MAFFT 7.407 (Kato and Standley, 2013). For peptide sequences MAFFT is run with the method "mafftinsi" and the option "--localpair", whereas for RNA sequences MAFFT is run with the method "mafft-xinsi" and the option "--scarnapair" to incorporate structure information and produce a structural alignment (Kato and Toh, 2008).

A.1.16. TOXiTAXi (Baranek et al., 2020)

Primary references: Baranek, J., Pogodziński, B., Szipluk, N. et al. *TOXiTAXi: a web resource for toxicity of Bacillus thuringiensis protein compositions towards species of various taxonomic groups*. Sci Rep 10, 19767 (2020)

Abstract: Bioinsecticides consisting of different sets of *Bacillus thuringiensis* (Bt) Cry, Cyt and Vip toxins are broadly used in pest control. Possible interactions (synergistic, additive or antagonistic) between these proteins can not only influence the overall efficacy of certain Bt-based bioinsecticide, but also raise questions regarding environmental safety. Here, we assemble, summarize and analyze the outcomes of experiments published over 30 years, investigating combinatorial effects among Bt Cry, Cyt and Vip toxins. We collected the results on 118 various two-to-five-component combinations that have been bioassayed against 38 invertebrate species. Synergism, additive effect and antagonism was indicated in 54%, 32% and 14% of experiments, respectively. Synergism was noted most frequently for Cry/Cyt combinations, followed by Cyt/Vip and Cry/Cry. In Cry/Vip combinations, antagonism is more frequent and higher in magnitude compared to other categories. Despite a significant number of tested Bt toxin combinations, most of them have been bioassayed only against one pest species. To aid the research on Bt pesticidal protein activity, we present TOXiTAXi (<http://www.combio.pl/toxitaxi/>), a universal database and a dedicated web tool to conveniently gather and analyze the existing and future bioassay results on biocidal activity of toxins against various taxonomic groups.

Link: <http://www.combio.pl/toxitaxi/>

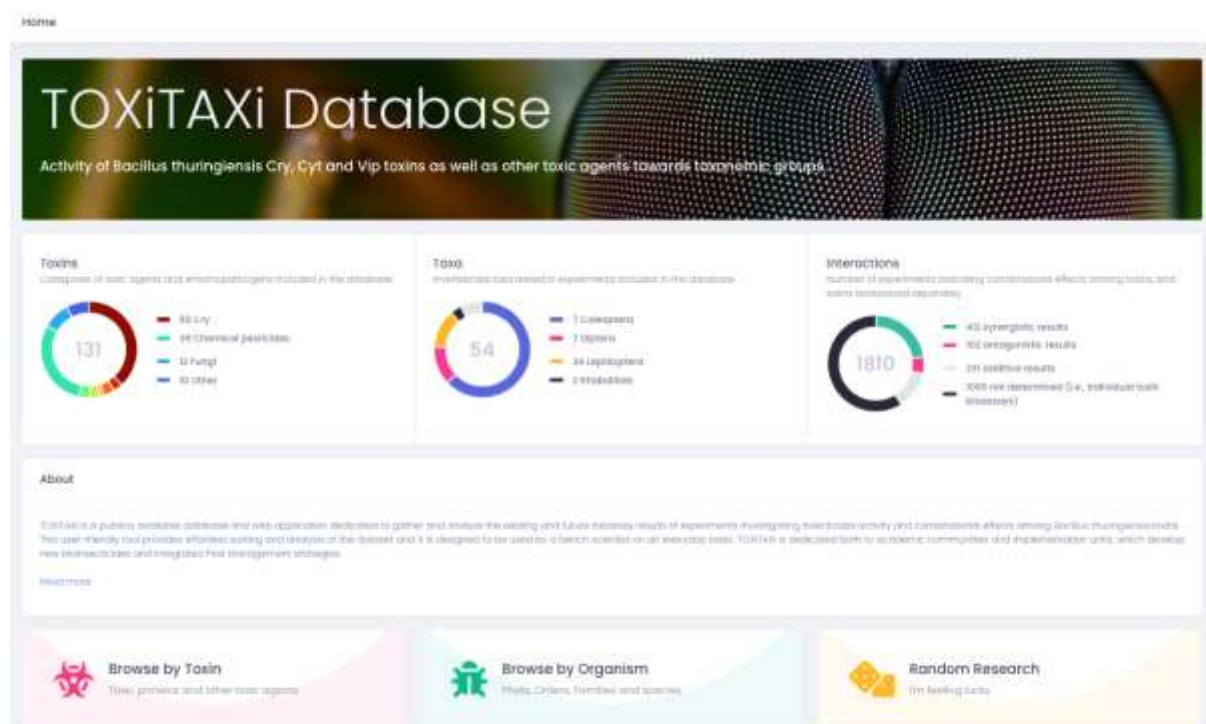


Figure 25: TOXiTAXi homepage.

www.efsa.europa.eu/publications

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author(s).

Citations: 1 citation

Field of application: *Bacillus thuringiensis* proteins.

Dataset: 1,810 separate experiments that have been performed since 1993 and published in 76 research articles were manually collected. Out of all collected experiments 973 test biocidal activity of single toxins and 837 investigate the potency of toxin compositions. Among these, 1,645 experiments (described in 59 manuscripts) investigate the activity of Bt proteins: 845 assess separate toxins and 800 concern Cry/Cyt/Vip toxin combinations.

Methods: TOXiTAXi is a manually curated database. It is not based on any predictive *in silico* methodology.

Appendix B Methodologies

B.1. SVM application to protein toxicity prediction

Primary citations: Bhosale H, Ramakrishnan V, Jayaraman VK. *Support vector machine-based prediction of pore-forming toxins (PFT) using distributed representation of reduced alphabets.* J Bioinform Comput Biol. 2021 Oct;19(5):2150028.

Abstract: *Bacterial virulence can be attributed to a wide variety of factors including toxins that harm the host. Pore-forming toxins are one class of toxins that confer virulence to the bacteria and are one of the promising targets for therapeutic intervention. In this work, we develop a sequence-based machine learning framework for the prediction of pore-forming toxins. For this, we have used distributed representation of the protein sequence encoded by reduced alphabet schemes based on conformational similarity and hydropathy index as input features to Support Vector Machines (SVMs). The choice of conformational similarity and hydropathy indices is based on the functional mechanism of pore-forming toxins. Our methodology achieves about 81% accuracy indicating that conformational similarity, an indicator of the flexibility of amino acids, along with hydrophobic index can capture the intrinsic features of pore-forming toxins that distinguish it from other types of transporter proteins. Increased understanding of the mechanisms of pore-forming toxins can further contribute to the use of such "mechanism-informed" features that may increase the prediction accuracy further.*

Citations: 1 citation.

Primary citations: Su MG, Huang CH, Lee TY, Chen YJ, Wu HY. *Incorporating amino acids composition and functional domains for identifying bacterial toxin proteins.* Biomed Res Int. 2014;2014:972692.

Abstract: *Aside from pathogenesis, bacterial toxins also have been used for medical purpose such as drugs for cancer and immune diseases. Correctly identifying bacterial toxins and their types (endotoxins and exotoxins) has great impact on the cell biology study and therapy development. However, experimental methods for bacterial toxins identification are time-*



consuming and labor-intensive, implying an urgent need for computational prediction. Thus, we are motivated to develop a method for computational identification of bacterial toxins based on amino acid sequences and functional domain information. In this study, a nonredundant dataset of 167 bacterial toxins including 77 exotoxins and 90 endotoxins is adopted to learn the predictive model by using support vector machines (SVMs). The cross-validation evaluation shows that the SVM models trained with amino acids and dipeptides composition could yield an accuracy of 96.07% and 92.50%, respectively. For discriminating endotoxins from exotoxins, the SVM models trained with amino acids and dipeptides composition have achieved an accuracy of 95.71% and 92.86%, respectively. After incorporating functional domain information, the predictive performance is further improved. The proposed method has been demonstrated to be able to more effectively identify and classify bacterial toxins than the other two features on independent dataset, which may aid in bacterial biomedical development.

Citations: 2 citations.

B.2. HMM application to protein toxicity prediction

Primary citations: Laht S, Koua D, Kaplinski L, Lisacek F, Stöcklin R, Remm M. *Identification and classification of conopeptides using profile Hidden Markov Models*. *Biochim Biophys Acta*. 2012 Mar;1824(3):488-92.

Abstract: *Conopeptides are small toxins produced by predatory marine snails of the genus Conus. They are studied with increasing intensity due to their potential in neurosciences and pharmacology. The number of existing conopeptides is estimated to be 1 million, but only about 1000 have been described to date. Thanks to new high-throughput sequencing technologies the number of known conopeptides is likely to increase exponentially in the near future. There is therefore a need for a fast and accurate computational method for identification and classification of the novel conopeptides in large data sets. 62 profile Hidden Markov Models (pHMMs) were built for prediction and classification of all described conopeptide superfamilies and families, based on the different parts of the corresponding protein sequences. These models showed very high specificity in detection of new peptides. 56 out of 62 models do not give a single false positive in a test with the entire UniProtKB/Swiss-Prot protein sequence database. Our study demonstrates the usefulness of mature peptide models for automatic classification with accuracy of 96% for the mature peptide models and 100% for the pro- and signal peptide models. Our conopeptide profile HMMs can be used for finding and annotation of new conopeptides from large datasets generated by transcriptome or genome sequencing. To our knowledge this is the first time this kind of computational method has been applied to predict all known conopeptide superfamilies and some conopeptide families.*

Citations: 16 citations.



Primary citations: Koua D, Laht S, Kaplinski L, Stöcklin R, Remm M, Favreau P, Lisacek F. *Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies*. *Biochim Biophys Acta*. 2013 Apr;1834(4):717-24.

Abstract: *Classified into 16 superfamilies, conopeptides are the main component of cone snail venoms that attract growing interest in pharmacology and drug discovery. The conventional approach to assigning a conopeptide to a superfamily is based on a consensus signal peptide of the precursor sequence. While this information is available at the genomic or transcriptomic levels, it is not present in amino acid sequences of mature bioactives generated by proteomic studies. As the number of conopeptide sequences is increasing exponentially with the improvement in sequencing techniques, there is a growing need for automating superfamily elucidation. To face this challenge we have defined distinct models of the signal sequence, propeptide region and mature peptides for each of the superfamilies containing more than 5 members (14 out of 16). These models rely on two robust techniques namely, Position-Specific Scoring Matrices (PSSM, also named generalized profiles) and hidden Markov models (HMM). A total of 50 PSSMs and 47 HMM profiles were generated. We confirm that propeptide and mature regions can be used to efficiently classify conopeptides lacking a signal sequence. Furthermore, the combination of all three-region models demonstrated improvement in the classification rates and results emphasise how PSSM and HMM approaches complement each other for superfamily determination. The 97 models were validated and offer a straightforward method applicable to large sequence datasets.*

Citations: 11 citations.