Predicting the Taxonomic Group of Venomous Animals Using Venom Proteins

Task 2: Proposal

Caleb Poock

Western Governors University

A: Project Overview

A1: Research Question or Organizational Need

The research question for this project is: "Can the taxonomic group of a venomous animal be predicted using only the protein composition of its venom?" As the literature review in section A3 will show, this project also addresses an organizational need within the multidisciplinary field of venom research: namely, the need for data analysis solutions and tools, both for the field in general, and also specifically for the development of safer and more effective antivenom treatments.

A2: Context & Background

There are 6 main taxonomic groups of animals that produce venom: Snakes, Spiders, Scorpions, Insects, Jellyfish + Relatives, and Cone Snails. The venom produced by certain species within these groups can cause lasting harm to humans, including death. Therefore, pharmaceutical research has produced many antivenoms to treat medical emergencies caused by these animals. Many compounds in these venoms have also been found to be useful in the development of new medicines to treat all sorts of issues (Morsy et al., 2023). However, this research faces significant difficulties.

The easiest antivenoms to produce are "monovalent" antivenoms, i.e. antivenoms that work to treat the venom from a single species. However, in many medical emergencies, the species of envenomating animal is not known, so a "polyvalent" (multi-species) antivenom must be used. This antivenom is generally tailored to the most common dangerous venomous animals present in the local area. Polyvalent antivenoms are more difficult to produce, and they have a higher rate of medical reactions and complications (Chanda et al., 2024). This project aims to

help this issue by providing the a machine learning model to predict the identity of a venomous animal based on the composition of its venom. The model could be used to develop tests to help medical professionals choose the right antivenom for an emergency, allowing for the use of the safer, simpler monovalent antivenoms.

On top of the difficulties that antivenom research faces on the medical side, the entire field of venom research suffers from scattered and disorganized data sources and tools. A centralized database and connected tools are necessary for venom research to reach its fullest potential (Zancolli et al., 2024). The workflow and model that will be produced by this project could be a small step toward such tools, providing a functional proof-of-concept for the usefulness of consolidated data and data analytics methods for venom research.

A3 & A3A: Summary of Works & Relation to Project

Work 1. Web of venom: exploration of big data resources in animal toxin research https://doi.org/10.1093/gigascience/giae054

Summary of work:

This paper summarizes the resources currently available to venom researchers and highlights the current fragmented and disconnected nature of such resources. It gives an overview of the tools used by different fields within venom research, including genomics, transcriptomics, proteomics, and other fields. A primary focus of the paper is the need for a centralized tool to consolidate and interact with the available data, and it discusses the various types of data that would need to be aggregated for such a project. However, the paper also notes that the funding, construction, and maintenance of a centralized tool would require significant resources and international collaboration and support. (Zancolli et al., 2024)

Relevance to project:

This capstone project will be a small step toward the kind of centralized resource that the paper argues for. It builds on the VenomZone and Tox-Prot resources developed and referenced by the authors, and aims to provide the framework for a pipeline to interact with, transform, and use the centralized data for model training and prediction. It should be both useful for building a larger, more comprehensive tool, and for providing a proof-of-concept for the usefulness and usability of VenomZone and future consolidated data sources.

Work 2. Venoms classification and therapeutic uses: A narrative review

https://doi.org/10.26355/eurrev_202302_31408

Summary of work:

The paper provides an overview of the various components found in venoms, and the different taxonomic groups that produce such venoms. It describes the effects that different kinds of venom proteins have on their victims, and includes instances where the effect or purpose is not entirely understood. This paper also summarizes some current therapeutic applications that venom research has produced, and highlights the need for "high-throughput screening" for future developments. (Morsy et al., 2023)

Relevance to project:

As a whole, the paper is an excellent overview on the importance of venom research, not only for emergency response to injuries from venomous animals, but also for the discovery and development of new medicines. As the first paper (Zancolli et al., 2024) showed, there is a wealth of venom information publicly available, and this paper highlights the benefits of conducting venom research for medical applications. Together, these two papers make a solid case for the need for data analysis in the world of venom research. This project is a step in that

direction, providing another tool in a venom researcher's toolbox to analyze the data and draw insights from it.

Work 3. Supplementation of polyclonal antibodies, developed against epitope-string toxin-specific peptide immunogens, to commercial polyvalent antivenom, shows improved neutralization of Indian Big Four and Naja kaouthia snake venoms

https://doi.org/10.1016/j.toxcx.2024.100210

Summary of work:

This paper is primarily focused on a method of developing polyvalent (multiple-species applicable) antivenom for snakebites. It gives context to the difficulties involved in treating venom-related medical emergencies, especially in rural areas and undeveloped countries. The paper also highlights the medical complications that can result from using a polyvalent (multi-species) antivenom, which are more common than when monovalent (single-species) antivenoms are used. However, because the species is not often known in snakebite cases (and other cases of envenomation), better methods for developing polyvalent antivenoms, such as the one presented in the paper, are needed. (Chanda et al., 2024)

Relevance to project:

The main applicability of this paper to this capstone project is that it highlights the difficulty in producing antivenoms that are useful in emergency situations when the identity of the venomous species is not known. The trained model this project will produce will hopefully be a valuable step toward an alternative solution; a way to identify the offending animal by the venom composition itself. While a test to analyze every venom component is unreasonable for this application, producing a trained model will provide valuable information on which

compounds are most useful for this diagnostic. By producing a trained model and then investigating which proteins the model relies on for classification, key proteins can be identified, which could then be used to develop a blood test that could give emergency medical personnel vital information on which antivenom to administer.

A4: Summary of Data Analytics Solution

This project's goal is to produce a trained random forest classifier model that can accurately predict the taxonomic group an animal belongs to, using only information about the proteins found in the animal's venom. It will use the only consolidated animal venom database, Tox-Prot (see section D1 for more information about the dataset) to train and test this model, which will require significant data transformation. The original format of the dataset presents each protein as a separate record. For the purposes of this project, the dataset will need to be transformed so that each species represented in the dataset is a single record, with column features representing the presence or absence of each protein family in that species' venom. Once the data is transformed and the model is trained, the performance of the model will be evaluated, with a target accuracy score of above 0.90. Python 3 will be the main development language for the project, with the pandas and scikit-learn libraries doing the heavy lifting for data transformation, analysis, and model training and testing. See section C3 for more information on the tools used.

A5: Benefits & Support of Decision-Making Process

As discussed in sections A1-A3, the benefits this project will produce are twofold:

1. The trained model could help emergency medical personnel in determining the correct antivenom to administer in cases where the identity of the venomous animal is not known. By predicting the animal's identity, a monovalent antivenom can be used, which is safer and cheaper than a polyvalent antivenom.
2. It will be a tool for venom researchers to discover relationships between the venoms of different species, as well as providing a model workflow for training a model on the Tox-Prot dataset.

B: Data Analytics Project Plan

B1: Goals, Objectives, & Deliverables

Goal: To determine whether venom composition can be used to identify animal groups.

- Objective: Transform the dataset to a useable form.
  - Deliverable: Standardize protein family names for grouping of proteins.
  - Deliverable: Simplify taxonomic lineage into 6 classes (Snakes, Spiders, Scorpions, Insects, Jellyfish + Relatives, Cone Snails).
  - Deliverable: Pivot table from protein-level records to organism-level records, with each protein family as a separate column feature.
- Objective: Predict placement in classes based on venom composition.
  - Deliverable: Trained and tested random forest model.
  - Deliverable: Accuracy score of model.
- Objective: Show distribution of protein families in each class.
  - Deliverable: A graph for each class showing the protein families present in that class, and the rates at which each family is present.

B2: Scope of Project

In-Scope:

This project will take the Tox-Prot dataset, transform it from protein-level records to organism-level records, train and test a random forest model for predicting class based on distribution of protein families, and produce protein family distribution graphs for each class.

Not-In-Scope:

The point of the project is to provide proof-of-concept for predicting taxonomic class based on venom composition; any further analysis is out of scope and more suited to future projects. While prediction to a more specific level than the 6 classes selected would be beneficial for practical applications, this would require a larger and more complete dataset, and so is out of scope for the data used. This project also will not prescribe the method by which its results could be used for developing a medical test, since that requires medical and pharmaceutical expertise beyond the ability of the author.

B3: Standard Methodology

This project will use the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining). This is well suited to a research project whose progress is mostly linear (like Waterfall methodologies), but will require some amount of iteration and return to previous steps (like Agile methodologies). The methodology is structured enough to provide guidelines and clear progression, but flexible enough that the design of the methodology does not hinder iteration where necessary. CRISP-DM has 6 major steps:

1. Business Understanding

   Identify a question or problem in an industry or area of study, and examine what data is available that could help answer that question or address the problem. For this project, this involved identifying the need in general for more data analysis and centralization of data within venom research, and in specific the usefulness of tools to predict venomous animal identity using venom for emergency medical applications. This step also involved the initial identification of the Tox-Prot dataset as a useful source of data.

2. Data Understanding

   Evaluate the data selected for content, quality, and usefulness/applicability to the project intent. This involves standard data exploration to discover issues that will need to be addressed in the next step, as well as identifying transformations necessary to fit the data to the methodology/model that will be used.

3. Data Preparation

   Transform the data into the form that will be used for analysis. This includes addressing quality and completeness issues, as well as transformations such as pivoting, melting, or one-hot encoding to make the dataset more suitable for use by the selected model.

4. Modeling

   Once the data is transformed, split it into training and testing data, and train the model on the test set.

5. Evaluation

   Once the model is trained, evaluate its performance and the implications of its success or failure.

6. Deployment

   Present the finished project and provide the workflow for others to use and modify to produce their own trained models. Generate charts that provide insight into the discoveries of the project and direction for further research and analysis.

B4: Timeline & Milestones

Milestone/Deliverable	**Duration**	**Timeframe**
Establish Project Question and Identify Dataset	3 days	11/22/25 – 11/24/25
Data Exploration	2 days	11/25/25 - 11/26/25
Submit Topic Approval	3 days	11/27/25 - 11/29/25
Data Cleaning and Transformation	3 days	11/30/25 – 12/2/25
Training and Testing Model	1 day	12/3/25
Generating Charts	2 days	12/4/25 – 12/5/25
Write Project Proposal	4 days	12/6/25 – 12/9/2025
Write Project Report	4 days	12/10/2025 – 12/13/25
Record Project Video	2 days	12/14/25 – 12/15/25

B5: Resources & Costs

1. Python Language, VSCode IDE, and Other Software: Free
2. Tox-Prot Dataset: Free
3. 128 Work Hours: $3,840 (at $30 / hour)

B6: Criteria for Success

The criteria for success for this project are:

1. A successfully cleaned and transformed dataset:

   The dataset is successfully cleaned, with missing values imputed or dropped. The "Protein families" and "Taxonomic lineage" columns are properly simplified for use in the machine learning model. The dataset is successfully pivoted from protein-level records to organism-level records.

2. A trained model and the accuracy score of the model on the test set:

   The random forest model is successfully trained on the training dataset. Once trained, the model is able to predict values from the test dataset. The model will be considered successful if it achieves an overall accuracy score above 80%. Below this threshold, the model's performance will not be considered successful, but the information gained from investigating the model's performance will still be considered a success for the project in terms of discovery of new information and avenues for further research.

3. A protein family distribution graph for each class:

   A protein family distribution graph (horizontal bar chart) is successfully generated for each of the six classes. Each graph shows both the raw count of protein family appearances in the class, and the percentage of the class this count represents. The graphs are also readable, without overlapping labels or bars that blend together.

C: Design of Data Analytics Solution

C1: Hypothesis

The proteins found in venoms are different enough between taxonomic groups (snakes, spiders, scorpions, insects, jellyfish and relatives, and cone snails) that the composition of the venom of a species can be used to classify it in one of these groups via a machine learning classifier model, without other information on the species.

C2: Analytical Method

The method used to test the hypothesis will be a random forest classifier model from the scikit-learn Python library. The model will be trained and tested on a stratified 80% / 20% split of the data, respectively. The performance of the trained model will be evaluated using the included .score() method to measure mean class accuracy on the test split of the dataset.

C2A: Justification

Random forest is a popular model for multi-class classification tasks. Because a random forest model breaks the classification task down into multiple smaller decision trees, it avoids the overfitting that more monolithic models might be prone to for multi-class prediction tasks. The scikit-learn library also provides methods for evaluating feature importance, which, while outside the scope of this project, is useful for the intended applications.

C3: Tools & Environment

The language used for this project is Python 3. The project will be developed in the VSCode IDE, and version control will be handled by GitHub. The coding environment is handled by Conda within a WSL Ubuntu virtual machine. The Conda environment will include

the following dependencies: matplotlib, pandas, jupyterlab, scipy, numpy, scikit-learn, pip, and pickle. The primary code for the project will be written in a jupyter notebook .ipynb file, with libraries and classes from the above modules imported ass necessary. Pandas will be the main module used to explore, clean, and transform the raw data, scikit-learn will be used for the training and testing of the random forest model, and matplotlib will be used for generating visuals.

C4: Methods & Metrics to Evaluate Statistical Significance

The random forest model is a supervised classification model. Its performance will be evaluated using the accuracy score returned by the scikit-learn .score() method, which gives the mean of the subset accuracies. If this score is below 0.8, accuracy scores for each class will be individually assessed. Scores lower than 0.7 will be considered a failure of the model, but this does not constitute a failure of the project goal. If the model is unable to achieve reasonably high accuracy, this is likely an indicator that there is more overlap in venom composition between taxonomic groups than expected by the hypothesis, which is itself useful information for this project's goals and intended application.

C4A: Justification of Selected Evaluation Methods & Metrics

As an ensemble method, the random forest algorithm is well-suited for a multi-class classification task such as the one presented in this project. The use of many individual voting decision trees reduces the chance of overfitting that a single decision tree or other method might be prone to. Scikit-learn's provided accuracy score method (.score()) gives a simple evaluation of the model's success or failure at the classification task, and there are methods and tools that can be used to investigate accuracy on a per-class basis if the overall accuracy is lower than targeted.

C5: Practical Significance

There are a few useful applications for this model. The simplest is as a tool to help medical professionals identify which antivenom to administer in cases where the identity of the envenomating animal is unknown. Antivenoms can either be developed to treat the venom of a single species, or to treat the venom of multiple species. However, the multi-species antivenoms (called polyvalent antivenom) have a higher chance of allergic reactions and other medical issues in the patient. Providing a tool that would help identify the species from the venom composition would allow for the accurate administration of single-species (monovalent) antivenom would reduce the chance of medical complications. The model produced in this project would not be directly used for this purpose; rather it is a proof of concept and a tool for developing the necessary medical tests. The workflow provided in this project could be tailored to the venomous animals present in a particular region to produce a model useful for medical workers, aiding in the development of regionally-specific diagnostic tests for use by EMTs and emergency rooms.

C6: Visual Communication

The matplotlib pyplot library will be used to generate a series of horizontal bar charts representing the presence and frequency of protein families within each class. These will show which protein families are widespread within and between the different classes, offering insights into the unique venom compositions of each group and inter-class overlap. There will be six graphs, one for each class, which will be titled “Distribution of Protein Families in [Class].” The graphs will have protein family names along the y-axis and the number of times they appear within the class on the x-axis. Because the classes are different sizes, the percentage of protein family occurrence count to total class size will also be added as a data label to each bar.

D: Description of Dataset

D1: Source of Data

The data used comes from the UniProtKB/Swiss-Prot Tox-Prot database, which contains expert-reviewed records of proteins found in animal venoms. This dataset is a subset of the larger UniProtKB/Swiss-Prot protein database. At the time of retrieval (November 19, 2025), the dataset contained 7,794 venom protein records. The dataset used can be found here: https://www.uniprot.org/uniprotkb/?query=taxonomy_id%3A33208+AND+%28cc_tissue_specificity%3Avenom+OR+cc_scl_term%3Anematocyst%29+AND+reviewed%3Atrue&facets=reviewed%3Atrue.

D2: Appropriateness of Dataset

This project seeks to assess the viability of predicting taxonomic group based on the protein composition of venom from an animal species. The dataset includes species name, protein name, protein family, and higher-order taxonomic classification information for each protein in the database, which will be all the information needed to assess the research question and train the corresponding model.

D3: Data Collection Methods

The dataset was downloaded as a .tsv using UniProt's download feature. Columns selected for download were Entry, Reviewed, Entry Name, Protein names, Organism, Length, Protein families, and Taxonomic lineage. Protein names, Organism, Protein families, and Taxonomic lineage are the only four columns that are necessary for the analysis and model

training; the rest were included to preserve data quality and provide easy data verification and reference to the original web-based dataset during exploratory analysis.

D4: Observations on Quality & Completeness of Data

The data is meticulously curated and intended for use in research, so it is very clean, complete, and has few missing values or issues. There are a little over 500 of the nearly 8,000 records which do not have a protein family listed, but this is due to incomplete or inconclusive research, rather than a failing on the part of the dataset. Since the dataset is small, artificial protein families may be generated based on protein name similarities in order to preserve information, but most of these records will likely be dropped.

D5: Data Governance, Privacy, Security, Ethical, Legal, & Regulatory Compliances

This dataset does not include personal or otherwise sensitive information, and is publicly available and intended for use in research, so privacy, security, and ethical, legal, and regulatory concerns are irrelevant. Data governance for the purpose of project integrity and version management will be handled as follows:

D5A: Precautions

- Project Integrity: The computer used to conduct this project is password protected and only accessible by the author, and the same is true for the GitHub repository. This ensures that only the author may contribute to the project, and the progress on the project will be clearly visible.
- Version Management: A GitHub repository will be used to track changes and backup work to protect against data loss.

References

Chanda, A., Salvi, N. C., Shelke, P. V., Kalita, B., Patra, A., Puzari, U., Khadilkar, M. V., & Mukherjee, A. K. (2024). Supplementation of polyclonal antibodies, developed against epitope-string toxin-specific peptide immunogens, to commercial polyvalent antivenom, shows improved neutralization of Indian big four and Naja Kaouthia snake venoms. *Toxicon: X*, *24*. https://doi.org/10.1016/j.toxcx.2024.100210

Morsy, M. A., Gupta, S., Dora, C. P., Jhawat, V., Dhanawat, M., Mehta, D., Gupta, K., Nair, A. A., & El-Daly, M. (2023). Venoms classification and therapeutic uses: A narrative review. *Eur Rev Med Pharmacol Sci*, *27*(4), 1633–1653. https://doi.org/10.26355/eurrev_202302_31408

Zancolli, G., von Reumont, B. M., Anderluh, G., Caliskan, F., Chiusano, M. L., Fröhlich, J., Hapeshi, E., Hempel, B.-F., Ikonomopoulou, M. P., Jungo, F., Marchot, P., de Farias, T. M., Modica, M. V., Moran, Y., Nalbantsoy, A., Procházka, J., Tarallo, A., Tonello, F., Vitorino, R., … Antunes, A. (2024). Web of venom: Exploration of big data resources in animal toxin research. GigaScience, 13. https://doi.org/10.1093/gigascience/giae054