biology and life sciences forum

MDPI

Proceeding paper

Machine Learning Analysis Suggest Relative Protein Abundance is Weakly Correlated with Snake Venom Toxicity †

Anas Bedraoui[1*], Salim El Mejjad[1], Zakaria Alouani[1], Salwa Enezari[1], Jacob A. Galan[2], Tariq Daouda[1*]

[1] Institute of Biological Sciences (ISSB), Faculty of Medical Sciences (FMS), Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco
[2] Department of Human Genetics, University of Texas Rio Grande Valley, Brownsville, TX, USA.
* Correspondence: Tariq.Daouda@um6p.ma, Anas.bedraoui@um6p.ma
† The 2nd International Electronic Conference on Toxins, online, 14-28 July 2023

Abstract: Snakebite is a neglected public health issue in many tropical and subtropical countries. Each year, about 5.4 million snake bites occur, resulting in 138,000 deaths, and over 400,000 amputations and other permanent disabilities. The bite causes severe neurotoxic, hemorrhagic and myotoxic damage, the extent of which depends on the toxicity and venom composition of the snake species. Therefore, predicting the toxicity from snake venom composition would vastly improve diagnosis, antivenom treatment, saving lives and limbs. Herein, we investigate the potential of Machine Learning (ML) in venomics, by training several models to predict Lethal Dose (LD50) from venom composition. The analysis was conducted on 130 snake species (15% of all venomous species), using five ML models: Support Vector Machine (SVM), Multilayer Perceptron (MLP), Linear Regression, Decision Tree, Random Forest, and four ensemble learning methods: Stacking, Voting, Bagging, AdaBoost, trained to predict LD50 from relative protein abundance. Although, data from 13 proteins and enzymes were combined, results showed an overall weak correlation between model prediction and LD50 (Spearman correlations ranging from 0.49 to 0.53, and Pearson correlations ranging from 0.30 to 0.47), even when considering only the highly significant proteins and enzymes: SVMP, SVSP, 3FTx, and PLA2. These results, challenge the assumption that relative protein abundance is the main driver of toxicity. They suggest that toxicity is a multi-factor phenomenon influenced by different biological aspects, such as protein 3D structure and potential binding sites. This in turn highlights the need for high quality multi-modal venomics databases, combining toxicity with several biological factors such as protein structure and metabolic data to better understand the nature of snake venom toxicity.

Keywords: Snakebite; Venom composition; Machine Learning; Lethal Dose (LD50); LD50 prediction

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

1. Introduction

Snakebite is a neglected tropical disease that causes extensive neurotoxic, hemorrhagic, and myotoxic damage, the severity of which is determined by the snake's toxicity and venom composition. Accurate prediction of the toxicity of snake venom, particularly the lethal dose (LD50) values, is crucial for the development of effective antivenoms and other therapies [1]. While *in vivo* animal studies remain the gold standard for determining LD50 values, this approach is hampered by ethical, temporal, and financial challenges [2].

Machine learning (ML) methods provide a promising approach for predicting LD50 values and identifying potential leads for developing effective antivenoms [3]. This approach offers the ability to evaluate the toxicity of venom compounds before performing animal studies, thereby decreasing time and costs associated with antivenom development while also reducing the need for animal testing [4].

Here, we used five ML algorithms and four ensemble learning methods to predict LD50 values of 13 relative protein abundances using data from 130 snake species (15% of all venomous species) belonging to the Viperidae, Elapidae, and Crotalidae families (Figure 1).

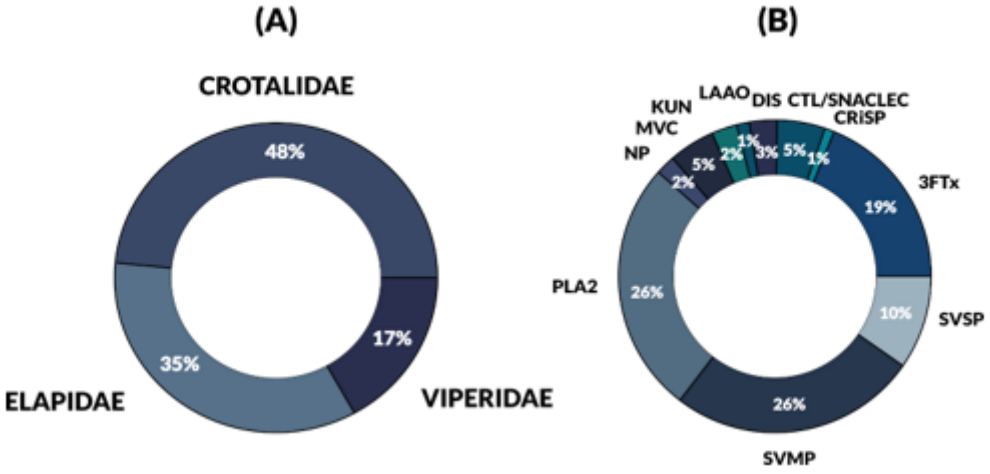

Figure 1. Relative distribution of the most characterized protein families for Crotalidae, Viperidae, and Elapidae. **(A)** Percentage distribution of snake species among Viperidae, Elapidae, and Crotalinae families. **(B)** Distribution of protein families from Crotalidae, Viperidae, and Elapidae. 3FTx, three finger toxins; SVMP, snake venom metalloproteinases; SVSP, snake venom serine proteases; PLA2, phospholipase A2; LAAO, l-amino acid oxidases; NP, natriuretic peptides; KUN, Kunitz peptides; DIS, disintegrins; CTL/SNACLEC, C-type lectins, and CTL-like proteins; MVC, minor venom components; CRiSP, Cysteine-Rich Secretory Proteins.

2. Methods

All models were trained to predict LD50 from the relative protein abundance of 13 proteins using a five-fold cross-validation with the python module Scikit-learn [5]. Results for five ML algorithms were evaluated: Linear Regression, Decision Tree, Support Vector Regression, Random Forest, and MLP (Figure 2). These models were also combined using four ensemble learning methods (Stacking [6], Voting [7], Bagging, [8]. and AdaBoost [9].) (Figure 2). The efficiency of each model was evaluated using Pearson and Spearman correlation coefficient (Figure 3,4).

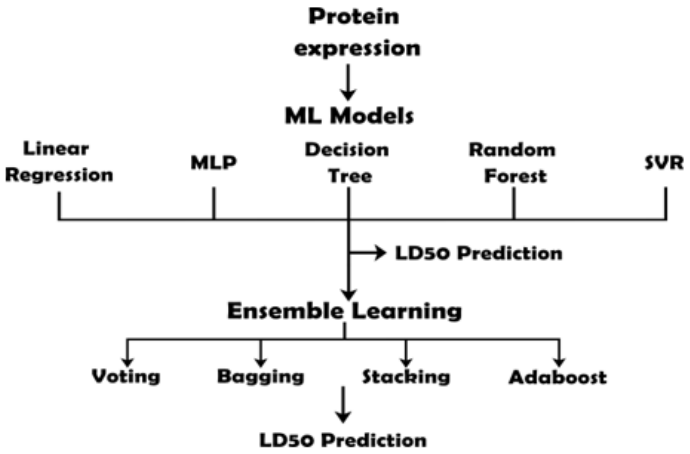

Figure 2. Five ML algorithms and four Ensemble learning methods to predict LD50 correlation scores. ML models were trained using 13 protein expressions as an input to predict correlations with LD50 as a target. Four ensemble learning methods were used to validate accuracy and enhance predictive performance of the ML models.

3. Results

Pearson correlations scores of the five ML regression models ranged from 0.30 to 0.47. Performances for the four ensemble learning methods ranged from: 0.49 to 0.53. Highest Spearman score was obtained using the AdaBoost method followed by Bagging, suggesting that there is value in combining the strengths of several machine learning models for the prediction of LD50s. However, prediction results remain relatively low, suggesting that relative protein abundance is but one of the factors behind snake venom toxicity.

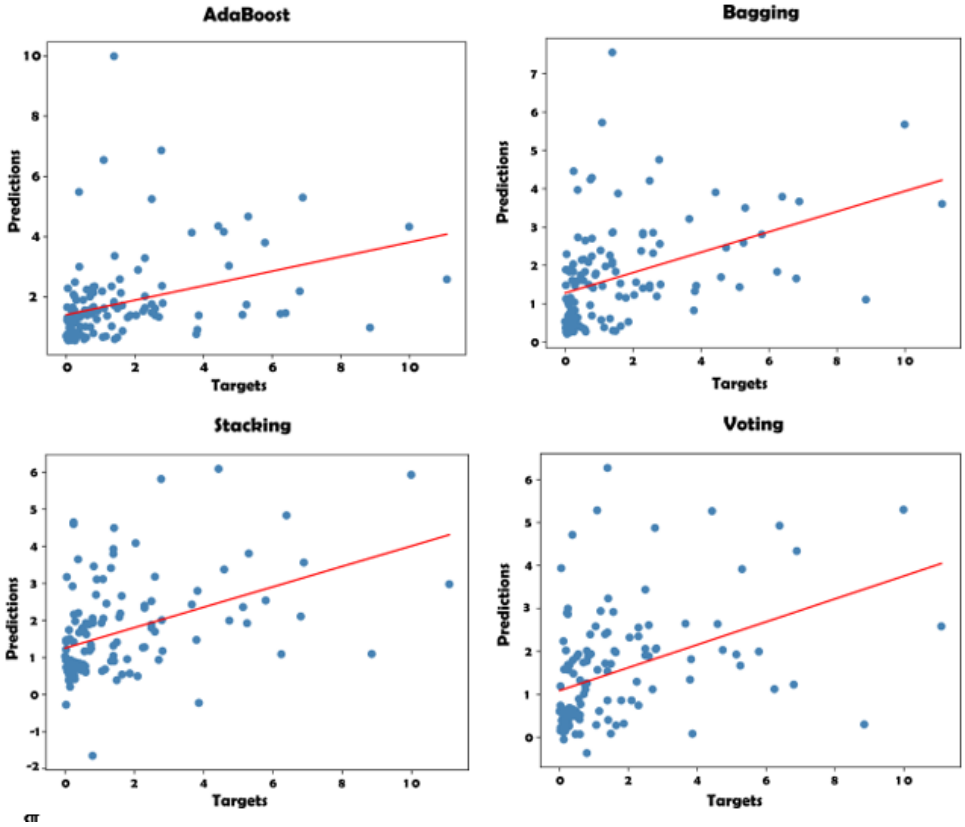

Figure 3. Spearman Correlation scores of four ensemble learning methods. AdaBoost achieved the highest performance with a score of 0.53. The other three methods, Bagging, Voting, and Stacking, showed similar low scores ranging from 0.49 to 0.51.

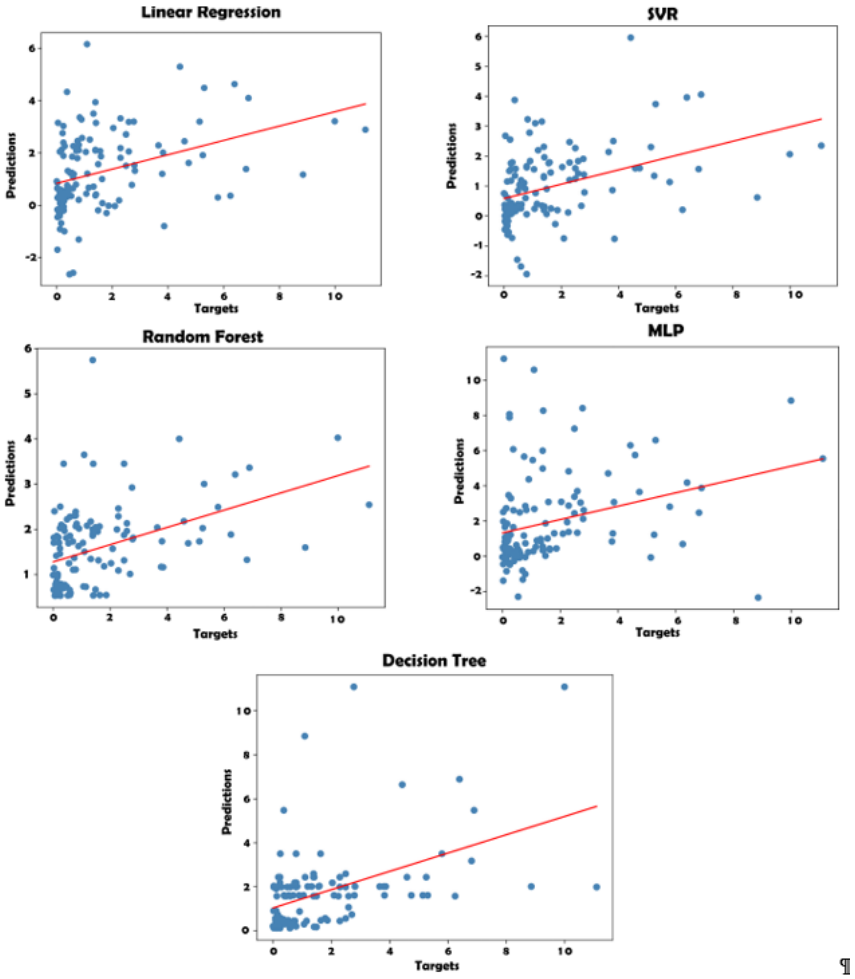

Figure 4. Pearson Correlation scores of five ML regressors. Decision Three achieved the highest performance with a score of 0.47. MLP had the lowest score of 0.30 due to the limited amount of data. The other three models, Linear Regression, SVR, and Random Forest, showed similar low scores ranging from 0.38 to 0.43.

4. Conclusions

This study combined data from 13 proteins and enzymes to predict LD50 values, but the results revealed a low correlation. Rather than supporting the hypothesis that relative protein abundance is the primary determinant of toxicity, the findings suggest a weak correlation between model prediction and LD50. Despite the relatively limited size of the dataset (13 protein families for 130 snakes), these results suggest that snake venom toxicity is a multifaceted phenomenon induced by several biological factors besides relative protein abundance, factors such as protein 3D structure and potential binding sites. This study highlights the necessity for high-quality, multi-modal venomics databases combining abundance with other modalities such as protein structure and metabolic data to help understand the intricate mechanisms involved in snake venom toxicity.

Author Contributions: Conceptualization, A.B., T.D., J.A.G.; methodology, A.B., T.D.; software, A.B.; validation, J.A.G, T.D.; formal analysis, A.B., S.E., Z.A., S.E.; investigation, T.D., J.A.G.; resources, T.D., J.A.G.; data curation, J.A.G.; writing—original draft preparation, A.B.; writing—review and editing, A.B., T.D., J.A.G.; visualization, A.B. S.E, Z.A., S.E., T.D., J.A.G.; supervision, T.D., J.A.G.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Albulescu, L.-O.; Xie, C.; Ainsworth, S.; Alsolaiss, J.; Crittenden, E.; Dawson, C.A.; Softley, R.; Bartlett, K.E.; Harrison, R.A.; Kool, J.; et al. A Therapeutic Combination of Two Small Molecule Toxin Inhibitors Provides Broad Preclinical Efficacy against Viper Snakebite. *Nat Commun* **2020**, *11*, 6094, doi:10.1038/s41467-020-19981-6.
2. Erhirhie, E.O.; Ihekwereme, C.P.; Ilodigwe, E.E. Advances in Acute Toxicity Testing: Strengths, Weaknesses and Regulatory Acceptance. *Interdiscip Toxicol* **2018**, *11*, 5–12, doi:10.2478/intox-2018-0001.
3. Lane, T.R.; Harris, J.; Urbina, F.; Ekins, S. Comparing LD50/LC50 Machine Learning Models for Multiple Species. *ACS Chem. Health Saf.* **2023**, doi:10.1021/acs.chas.2c00088.
4. Galati, S.; Di Stefano, M.; Martinelli, E.; Macchia, M.; Martinelli, A.; Poli, G.; Tuccinardi, T. VenomPred: A Machine Learning Based Platform for Molecular Toxicity Predictions. *Int J Mol Sci* **2022**, *23*, 2105, doi:10.3390/ijms23042105.
5. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project 2013.
6. Breiman, L. Stacked Regressions. *Mach Learn* **1996**, *24*, 49–64, doi:10.1007/BF00117832.
7. Bartlett, P.; Freund, Y.; Lee, W.S.; Schapire, R.E. Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *The Annals of Statistics* **1998**, *26*, 1651–1686, doi:10.1214/aos/1024691352.
8. Breiman, L. Bagging Predictors. *Mach Learn* **1996**, *24*, 123–140, doi:10.1007/BF00058655.
9. Schapire, R.E. Explaining Adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* **2013**, 37–52, doi:10.1007/978-3-642-41136-6_5.