TOXIFY: a deep learning approach to classify animal venom proteins

T Jeffrey Cole Corresponds. 1, Michael S Brewer Corresponds. 1

1 Department of Biology, East Carolina University, Greenville, NC, United States

Corresponding Authors: T Jeffrey Cole, Michael S Brewer
Email address: coleti16@students.ecu.edu, brewermi14@ecu.edu

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.

Availability: https://www.github.com/tijeco/toxify
**TOXIFY: a deep learning approach to classify animal venom proteins**

T. Jeffrey Cole and Michael S. Brewer

1 Department of Biology, East Carolina University, E 5th St, Greenville, NC 27858, USA

Corresponding author:
Michael S. Brewer

Email address: brewermi14@ecu.edu

**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. Availability: https://www.github.com/tijeco/toxify

**1 INTRODUCTION**

Venom is a complex trait utilized across numerous animal phyla and is an ideal source for novel drug discoveries with millions of peptides highly specific for therapeutically relevant ion channels, G-protein-coupled receptors, enzymes, and transporters (Prashanth et al., 2017). Despite recent advances in venom gland proteomics and transcriptomics that have yielded millions of potential venom peptides (Fry et al., 2009), the biological function of only approximately 8,000 entities have been empirically verified and deposited into UniProt to date. In lieu of empirical evidence, venom proteins are often identified via BLAST searches, which can be a problematic approach for two primary reasons. First, venom proteins frequently evolve via gene duplications and subsequent neofunctionalization and subfunctionalization of non-toxic physiological proteins, rendering BLAST unable to possess the sensitivity needed to distinguish venomous paralogs (Hargreaves et al., 2014; Duda and Palumbi, 1999). Second, venom peptides are often highly diverged to the extent that BLAST searches yield no strong matches. Gacesa et al. (2016) attempted to alleviate these difficulties through their software ToxClassifier with Generalized Linear Models (GLM) and by training a support vector machine (SVM) on venom and non-venom proteins encoded with Hidden Markov Models (HMM) of “tox-bit” motifs via BLAST searches against a positive venom database (Gacesa et al., 2016). However, SVM-based methods are restricted to linear representations, which may limit performance in regard to classifying protein sequences (Byvatov et al., 2003). Here we present TOXIFY, a deep-learning approach to distinguish animal venom proteins from non-toxic proteins by training neural networks on protein sequences encoded as numerical vectors.

**2 METHODS**

**2.1 Datasets**

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.
1. To train our models to classify venom proteins, we constrained our training sets to only include verified venom proteins from 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.

2. “Negative” datasets comprised 50,000 random, non-venom proteins from 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 AA), 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.

2.2 Sequence encoding

In order to encode proteins as numerical representations, the amino acid sequence for each protein was inscribed as a 5x500 matrix with values corresponding to each of the five Atchley factors per amino acid in the protein (Atchley et al., 2005). Atchley factors were chosen because they statistically summarize 500 AA attributes that reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. Matrices of proteins containing fewer than 500 amino acids were padded with zeros to fill the remainder of the matrix, which do not affect the resulting models and predictions.

2.3 Sequence classifier

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), we 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 (Figure 1).
2.4 Benchmark against TOXCLASSIFIER

While other machine learning approaches exist, such as TOXINPRED (Gupta et al., 2013) and CLANTOX (Naamati et al., 2009), TOXCLASSIFIER is the only one that supports large proteomic datasets, so we chose TOXCLASSIFIER as the standard to benchmark against TOXIFY. Test data used to compare performance of TOXIFY to TOXCLASSIFIER included all protein sequences from Swiss-Prot with ≤ 500 amino acids that have been uploaded since the publication of TOXCLASSIFIER (2016-2018) and not used in the training set of either TOXIFY or TOXCLASSIFIER. These were split into a positive dataset with all 274 verified venom protein sequences and a negative dataset with 274 randomly selected and verified non-venom protein sequences.

TOXCLASSIFIER uses nine SVM and GLM models and provides as output a nine dimensional vector of zeros and ones indicating the prediction of whether or not the protein is a venom from each model. The probability for TOXCLASSIFIER was calculated as the average for all models. Contrarily, the probability a protein is a venom from TOXIFY was determined as the probability calculated for each protein by the trained network. The performance of both models were measured using the following same criteria as Gacesa et al. (2016).

1. True positive (TP): This is the approximate number of positive sequences that were correctly predicted to be venom proteins. It is calculated as the product of the mean venom probability of the positive dataset and the total number of positive sequences.

2. True negative (TN): This is the approximate number of negative sequences that were correctly predicted to be non-venom proteins. This is calculated as the product of the mean venom complementary probability of the negative dataset and the total number of negative sequences.

3. False positive (FP): This is the approximate number of negative sequences that were incorrectly predicted to be venom proteins. This is calculated as the product of the mean venom probability of the negative dataset and the total number of negative sequences.

4. False negative (FN): This is the approximate number of positive sequences that were incorrectly predicted to be non-venom proteins. This is calculated as the product of the mean venom complementary probability of the positive dataset and the total number of positive sequences.

5. Accuracy (ACC): This is the proportion of proteins that were correctly predicted to be either a venom protein or a non-venom protein. This was calculated using the following expression; 

\[ ACC = \frac{TP + TN}{TP + TN + FP + FN} \]

6. Specificity (SPEC): This is the proportion of proteins that were correctly classified as a non-venom protein. This was calculated using the following expression; 

\[ SPEC = \frac{TN}{TP + TN} \]

7. Sensitivity (SENS): This is the proportion of proteins that were correctly classified as a venom protein. This was calculated using the following expression; 

\[ SENS = \frac{TP}{TP + FN} \]

8. Balanced accuracy (BACC): This is the average of specificity and sensitivity, which was calculated using the following expression; 

\[ BACC = \frac{SPEC + SENS}{2} \]

9. Negative predictive value (NPV): This is the proportion of negative sequences that were classified as true negatives, which was calculated using the following expression; 

\[ NPV = \frac{TN}{TN + FN} \]

10. Positive predictive value (PPV): This is the proportion of positive sequences that were classified as true positives, which was calculated using the following expression; 

\[ PPV = \frac{TP}{TP + FP} \]

11. F-score (F1): This is the harmonic mean of precision and sensitivity, which was calculated using the following expression; 

\[ F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \]

12. Matthew’s Correlation coefficient (MCC): This is a measure of the correlation between observed and predicted values (Matthews, 1975; Powers, 2011), which was calculated using the following expression; 

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \]

These analysis were carried out on the same Linux computer (Ubuntu 18.04; Dell PowerEdge-R910; 48 threads; 1 Tb RAM). The elapsed real time between invocation and termination of each process by the CPU was recorded in seconds(s), and the maximum resident set size of memory allocation (MEM) was recorded in megabytes (MB).
3 RESULTS

3.1 Validation
The overall accuracy of the model was assessed every 2nd epoch during training as recorded by the loss of accuracy, the training accuracy, and the test dataset accuracy (Figure 2). At the end of training, the model approached 99.9% training accuracy and 97.4% accuracy on the test dataset.

![Figure 2. Accuracy progression for RNN as training progressed for both training and test datasets.](image)

3.2 Benchmark against ToxClassifier
Of the 274 venom and non-venom proteins uploaded to Swiss-Prot since the publication of TOXCLASSIFIER, TOXIFY correctly classified 37% more venom proteins than TOXCLASSIFIER. TOXIFY also correctly classified 96.0% of non-venom proteins, which is only 2.7% less than the amount correctly classified by TOXCLASSIFIER. Further, TOXIFY completed these predictions an orders of magnitude faster than TOXCLASSIFIER, and used an order of magnitude less memory.

|                | TOXCLASSIFIER | TOXIFY |
|----------------|---------------|--------|
| TP*            | 152.8 (55.8%) | 209.6 (76.5%) |
| TN             | 270.2 (98.6%) | 263.0 (96.0%) |
| FP             | 3.8 (1.4%)    | 11.0 (4.0%)   |
| FN*            | 39.8 (44.2%)  | 64.4 (23.5%)  |
| ACC*           | 0.77          | 0.86        |
| SPEC           | 0.99          | 0.96        |
| SENS*          | 0.56          | 0.76        |
| BACC*          | 0.77          | 0.86        |
| NPV*           | 0.69          | 0.80        |
| PPV            | 0.98          | 0.95        |
| F1*            | 0.71          | 0.85        |
| MCC*           | 0.60          | 0.74        |
| CPU (s)*       | 100.18        | 4.05        |
| MEM (MB)*      | 6,824         | 293         |

Table 1. Benchmark metrics for TOXIFY compared to TOXCLASSIFIER. The top portion shows averages and percentages (in parenthesis) of true and false positives (TP & FP) as well as true and false negatives (TN & NF). Additionally, the proportions for accuracy, specificity, sensitivity, balanced accuracy, negative predictive value, positive predictive value, F-score, and Matthew’s correlation coefficient are listed. The bottom portion shows computational performance in terms of CPU time in seconds and memory usage in megabytes.
3.3 Implementation

3.3.1 Sensitivity against non-toxic homologs

To ensure that TOXIFY is able to discern between toxins and homologous physiological proteins, TOXIFY classified the top hits from a BLASTP search of the remaining 502,000 proteins in Swiss-Prot (not included in the training dataset) against the positive dataset keeping hits with an e-value cutoff of $1 \times 10^{-6}$. TOXIFY correctly assigned at least a 90% non-venom probability to 1,961 out of 2,183 (89.8%) of the non-toxic homologs, which is only 6.2 percentage points lower than TOXIFY’s true negative rate when calculated for non-toxic non-homologous proteins.

3.3.2 Transcriptome performance

To assess how TOXIFY performs on high-throughput venom cDNA sequence data across venomous taxa, we reassembled published venom gland transcriptomes available through NCBI Short Read Archive, including the following: eastern diamondback rattlesnake (*Crotalus adamanteus*), striped bark scorpion (*Centruroides hentzi*), cone snail (*Conus litteratus*), and Chinese bird spider (*Haplopelma hainanum*). Reads were trimmed and error corrected using the FASTP (v0.19.5) software with default parameters (Chen et al., 2018). All transcriptomes were assembled using TRINITY (v2.5.1) (Grabherr et al., 2011). Transcripts were then provided as input for TRANSDCODER (v5.3.0) (Haas et al., 2013) to infer complete open reading frames (ORFs). By implementing BLASTP searches against a database containing highly curated venom protein sequences from each respective taxon, we were able to partition the protein sequences into candidate venom proteins. The top BLASTP hits with an e-value cutoff of $1 \times 10^{-6}$ and at least a 90% sequence identity met our criteria for being venom proteins. We then provided those protein sequences as input for TOXIFY to classify the sequences. The results for this analysis are summarized in Table 2. In all transcriptomes, a relatively small percentage (<1%) of complete coding sequences had strong BLASTP hits with their respective Swiss-Prot database. In total, TOXIFY classified around 20% of all coding sequences from the transcriptomes as being venom proteins, despite the fact that less than 1% of those sequences met our BLAST criteria. Because many additional venoms are known to be present in these taxa beyond the number indicated by BLAST searches, this highlights TOXIFY’s greater predictive power compared to BLAST-based approaches. Further, 70% of those top BLASTP hits were correctly classified as venom proteins.

| Species                      | SRA number | coding sequences | toxify sequences | BLASTP hits | toxify verified |
|------------------------------|------------|------------------|------------------|-------------|----------------|
| *Crotalus adamanteus*        | SRR5259496 | 16,393           | 3,247            | 21          | 13             |
| *Centruroides hentzi*        | SRR6041834 | 26,315           | 5,825            | 4           | 2              |
| *Haplopelma hainanum*        | SRR2155568 | 61,012           | 6,131            | 4           | 3              |
| *Conus litteratus*           | SRR6381570 | 25,230           | 7,515            | 19          | 16             |

Table 2. Summary information for venom gland transcriptomes from four species analyzed by TOXIFY, including NCBI Short Read Archive accession number, total number of complete coding sequences, total number of sequences with a venom probability greater than 90%, total number of BLASTP hits with an e-value cutoff of $1 \times 10^{-6}$ in addition to at least 90% identity for reviewed proteins for each respective species from Swiss-Prot, and the number of those BLASTP hits that had a venom probability greater than 90%.

3.3.3 Case usage

Venom proteins from highly studied venomous animals made up the bulk of the training data. In order to assess the robustness of TOXIFY to other venomous animals, we classified a recently discovered venom protein expressed by a robber fly (Diptera: Asilidae: *Machimus arthriticus*), which belongs to a lineage of venomous flies whose venoms had not been previously studied and were not found in Swiss-Prot before our cutoff date. The neurotoxin described as asilidin1 was discovered through venom gland transcriptomics and venom proteomics, and its activity as a neurotoxin was verified via injection of the synthesized peptide into a honey bee (*Apis mellifera*), which caused subsequent paralysis. TOXIFY classified asilidin1 as a venom protein with a probability of 0.99.
4 DISCUSSION AND CONCLUSION

In this paper, we proposed TOXIFY, the first animal toxin classifier that utilizes deep learning. While other classifiers use slow BLAST-based methods to encode proteins as a set of features as training input for SVM models, TOXIFY quickly encodes amino acids as a 5-dimensional vector of Atchley factors and makes use of the sequential nature of proteins via neural networks with GRUs to classify sequences. While TOXIFY had a slightly poorer false positive rate and specificity scores when compared to TOXCLASSIFIER, it achieved a better accuracy, sensitivity score, F-score, and Matthew’s correlation coefficient – while performing with speeds and memory efficiency an order of magnitude better than TOXCLASSIFIER.

There are two obvious applications for TOXIFY. First of all, the prediction module of TOXIFY may be used to quickly classify large venom protein datasets. These datasets may be in the form of venom gland transcriptomes or venom proteomes and can stem from a wide variety of venomous taxa. Users may also find TOXIFY to be a helpful screening tool to reduce runtime of time-consuming annotation pipelines such as VENOMIX (Macrander et al., 2018). Secondly, the training module of TOXIFY has been setup for users to easily retrain models on custom protein datasets, while also being able to easily adjust hyper-parameters such as the number of training epochs, learning rate, and unit size for the GRUs.

While we were able to demonstrate TOXIFY’s utility in classifying venom proteins across the phylogenetic diversity of eumetazoans, improvements in the overall accuracy between taxa can still be made. Balancing the training dataset is still an issue due to verified venom protein sequences being outnumbered by non-venom proteins by several orders of magnitude. Improvements in the accuracy of the current iteration of TOXIFY will be accomplished by empirically verifying the functionality of additional venoms from an ever increasing number of sequenced venom proteins from a broader array of venomous taxa. In the meantime, a tool like TOXIFY in tandem with TOXCLASSIFIER (Gacesa et al., 2016), TOXINPRED (Gupta et al., 2013), CLANTOX (Naamati et al., 2009), and VENOMIX (Macrander et al., 2018) may expedite the process of narrowing down potentially novel venoms to be empirically verified and added to repositories such as Swiss-Prot.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Graduate Research Fellowship and the East Carolina University Department of Biology. We thank Chris Cohen for editing and improving the manuscript.
REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In OSDI, volume 16, pages 265–283.

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, 102(18):6395–6400.

Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences*, 43(6):1882–1889.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Duda, T. F. and Palumbi, S. R. (1999). Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod conus. *Proceedings of the National Academy of Sciences*, 96(12):6820–6823.

Fry, B. G., Roelants, K., Champagne, D. E., Scheib, H., Tyndall, J. D., King, G. F., Nevalainen, T. J., Norman, J. A., Lewis, R. J., Norton, R. S., et al. (2009). The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annual review of genomics and human genetics*, 10:483–511.

Gacesa, R., Barlow, D. J., and Long, P. F. (2016). Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Computer Science*, 2:e90.

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Raghava, G. P., Consortium, O. S. D. D., et al. (2013). In silico approach for predicting toxicity of peptides and proteins. *PLoS One*, 8(9):e73957.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494.

Hargreaves, A. D., Swain, M. T., Hegarty, M. J., Logan, D. W., and Mulley, J. F. (2014). Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. *Genome biology and evolution*, 6(8):2088–2095.

Macrander, J., Panda, J., Janies, D., Daly, M., and Reitzel, A. M. (2018). Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. *PeerJ*, 6:e5361.

Naamati, G., Askenazi, M., and Linial, M. (2009). Clantox: a classifier of short animal toxins. *Nucleic acids research*, 37(suppl.2):W363–W368.

Prashanth, J. R., Hasaballah, N., and Vetter, I. (2017). Pharmacological screening technologies for venom peptide discovery. *Neuropharmacology*, 127:4–19.