Multidisciplinary Cancer Investigation **Original Article**

April 2025, Volume 9, Issue 1

Language Model–Based Representation Learning for Venom Protein Identification and Therapeutic Target Discovery in Cancer

Meisam Ahmadi[1], Mohammad Reza Jahed-Motlagh[1*], Ehsaneddin Asgari[2], Adel Torkaman Rahmani[1]

[1] *Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran*

[2] *Qatar Computing Research Institute, Doha, Qatar*

***Corresponding Author:** Mohammad Reza Jahed-Motlagh, Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran, E-mail: jahedmr@iust.ac.ir*

Submitted: 06 October 2025
Revised: 12 October 2025
Accepted: 28 October 2025
e-Published: 01 december 2025

Keywords:
Protein Language Models
Representation Learning
Venom-inspired Drug Discovery

Introduction: Venom is a complex mixture of bioactive molecules produced by venomous organisms for predation, defense, or intraspecific competition, often leading to specific physiological responses in target organisms. Venom-derived peptides and proteins have recently garnered attention in biomedical research for their potential therapeutic applications, including the discovery of anticancer drugs. However, venom sequences constitute a highly divergent class of proteins, making their identification using machine learning and homology-based methods particularly challenging. To address this, we propose *ToxVec*, a transfer learning-based framework for the automatic representation learning of protein sequences, designed to improve venom identification.

Materials and Methods: Our approach leverages pre-trained protein language models to capture sequence-level information without manual feature engineering.

Results: *ToxVec* outperforms existing feature-based models, achieving a macro-F1 score of 0.89. Furthermore, an ensemble model trained on multiple balanced subsets achieves a macro-F1 score of 0.93, representing a 7% improvement over the state of the art. Beyond benchmark performance, screening of experimentally validated anticancer peptides from the CancerPPD2 dataset revealed that many exhibit high venom-like signatures according to *ToxVec*, supporting the notion that toxin-inspired molecular architectures may underlie anticancer bioactivity.

Conclusions: We further discuss how language model–based representation learning embodies a Cognitive Mind–Body–Inspired interpretation, linking abstract sequence semantics (the "mind") to biological function (the "body"). By enabling more accurate large-scale identification of venom proteins, *ToxVec* provides a foundation for systematically exploring venom-derived bioactive peptides as potential therapeutic candidates, including those targeting pathways implicated in breast cancer progression and metastasis. This automated approach thus bridges computational protein informatics with translational oncology, supporting future efforts in bioactive peptide-based anticancer research.

INTRODUCTION

Venom is a complex mixture of enzymatic and non-enzymatic substances produced by venomous organisms for predation, defense, or intraspecific competition [1], often resulting in the immobilization or paralysis of target organisms. Venom has evolved independently multiple times throughout the tree of life, making it a topic of considerable interest in both evolutionary and biomedical fields [2]. Rich in peptides that interact with ion channels, G-protein-coupled receptors, and transporters, venom constitutes a valuable reservoir for therapeutic discovery [3, 4]. Several venom-derived peptides have demonstrated promising pharmacological activities, including cytotoxic, antiangiogenic, and antimetastatic effects relevant to cancer therapy [5,6]. These properties underscore the growing importance of venom-based compounds in the search for novel anticancer agents.

Despite their anticancer potential, only a small fraction of proteins in major databases, such as UniProt/SwissProt, are annotated as venom or toxin proteins [7]. Currently, only 8,101 of 573,661 Swiss-Prot sequences are labeled as such, highlighting the need for computational approaches that can automatically and accurately identify venom peptides in large proteomic datasets. This task remains challenging because venom and non-venom proteins are not easily separable by sequence similarity alone. BLAST-based homology searches often fail to discriminate between the two, as venoms (i) have frequently evolved from non-toxic precursors [8], resulting in primary sequences similar to non-venom counterparts, and (ii) have diverged extensively during evolution [9], such that homologous venom proteins may exhibit low sequence identity.

To address these challenges, several computational and machine learning- based methods have been proposed for venom prediction [10, 11, 12, 13, 14, 15, 16]. For example, **ClanTox** [13] employs boosted-stump classifiers over 545 handcrafted sequence features, while **ToxClassifier** [12] integrates multiple models, including SVM, GBM, and GLM, trained on combinations of amino acid frequencies, tox-bit motifs [17], and homology-based features. More recently, **toxify** [10] utilized a recurrent neural network (RNN) architecture based on Gated Recurrent Units (GRUs) [18] with Atchley factor encoding [19]. Although these methods achieved improved predictive accuracy, they are constrained by their dependence on manually engineered features, which limits their generalizability and scalability.

In this study, we present *ToxVec*, a language modeling-based framework for automatic representation learning of protein sequences designed to enhance venom identification. Instead of relying on handcrafted features, *ToxVec* leverages pre-trained protein language models to capture rich contextual information embedded in amino acid sequences. Building on skip-gram architectures [20, 21] and earlier protein embeddings such as ProtVec [22], *ToxVec* fine-tunes learned representations to distinguish venom from non-venom proteins effectively.

Over the last decade, transfer learning and, in particular, language modeling has revolutionized various domains of machine learning, particularly where annotated data are scarce [23, 24, 25, 26, 27]. Neural language models trained on large unlabeled corpora can serve as universal sequence encoders and be repurposed for diverse downstream biological prediction tasks [28, 29]. In bioinformatics, this paradigm enables the extraction of generalizable protein sequence representations that can be adapted to specialized applications, including venom protein identification and therapeutic peptide discovery [30].

Our main contributions are as follows: **(i)** We propose *ToxVec*, a transfer learning-based framework for venom protein identification that automatically learns biologically meaningful representations directly from raw sequence data.

(ii) We demonstrate that fine-tuning pre-trained protein language models significantly outperforms existing state-of-the-art approaches in venom peptide classification. **(iii)** We introduce ensemble learning strategies based on resampled negative instances to enhance classification robustness and improve macroF1 performance. **(iv)** We discuss the therapeutic relevance of venom-derived peptides identified through *ToxVec*, with particular emphasis on their potential applications in cancer research. **(v)** Finally,

we provide a conceptual interpreta- tion of our framework through a *Cognitive Mind–Body–Inspired AI* perspective, linking abstract sequence representations ("mind") to their functional biological manifestations ("body").

Materials and Methods

Datasets

1. Toxin/Venom Protein Dataset

For benchmarking, we adopt the dataset introduced by **Toxify** [10], which provides a standardized and curated benchmark for venom protein prediction using high-quality Swiss-Prot annotations. The dataset includes nonoverlapping training and test splits designed to evaluate model generalization to newly annotated sequences.

Training set: **(i) Positive examples:** 6,133 venom protein sequences extracted from Swiss-Prot entries annotated with the annotation term "tissue specificity: venom". These sequences span multiple venomous taxa, including snakes, scorpions, spiders, cone snails, and cnidarians. **(ii) Negative examples:** 50,000 non-venom protein sequences randomly selected from Swiss-Prot entries lacking the annotation term "tissue specificity: venom". To prevent data leakage, only entries uploaded prior to June 2016 were included in the analysis.

Test set: 274 verified venom protein sequences and 94 verified non-venom protein sequences from Swiss-Prot entries uploaded between 2016 and 2018. These data were not used during training, ensuring that the evaluation reflects the model's ability to generalize to unseen and temporally newer proteins. This benchmark offers a realistic class distribution and biological diversity, making it suitable for evaluating transfer learning frameworks, such as *ToxVec*.

Anticancer Peptide Dataset

To explore potential intersections between venom peptides and anticancer bioactivity, we additionally considered data from **CancerPPD2** [31], an updated and comprehensive repository of experimentally validated anticancer peptides and proteins1). The database contains 6,521 entries, each annotated with peptide origin, target cancer cell line, tissue type, and sequence information. These peptides have been evaluated across 392 cancer cell lines and 28 cancer-associated tissues.

Skip-gram Analogous to Language Modeling

Language modeling aims to assign a probability $P(w_1, w_2, \ldots, w_N)$ *to a sequence of elements, such as words, phrases, or amino acids,* $w_1, w_2, \ldots, w_N$. It is a fundamental task in natural language processing (NLP), underpinning applications such as text generation, translation, and conversational systems. The joint probability of a sequence can be factorized using the chain rule:

$$P(w_1, w_2, \ldots, w_N) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times \ldots \times P(w_N|w_1, \ldots, w_{N-1})$$

Because it requires only raw sequence data and captures contextual dependencies, language modeling serves as an ideal foundation for self-supervised pretraining and transfer learning. Such approaches have achieved state-of-the-art performance in both NLP and bioinformatics [22, 23, 26, 29, 32].

Among the various architectures proposed, we focus on the *Skip-gram* neural network (illustrated in Figure 1.1), whose objective is analogous to language modeling but with the roles of input and output reversed: it predicts surrounding (context) tokens for a given central token. The objective of *Skip-gram* is to Maximize the following log-likelihood:

$$\sum_{t=1}^{M} \sum_{c \in [t-N, t+N]} \log p(w_c|w_t),$$

Where N is the context window size around a target word w_t, c indexes the context positions, and M is the total number of term occurrences in the corpus. To model the likelihood of a context word given a target,

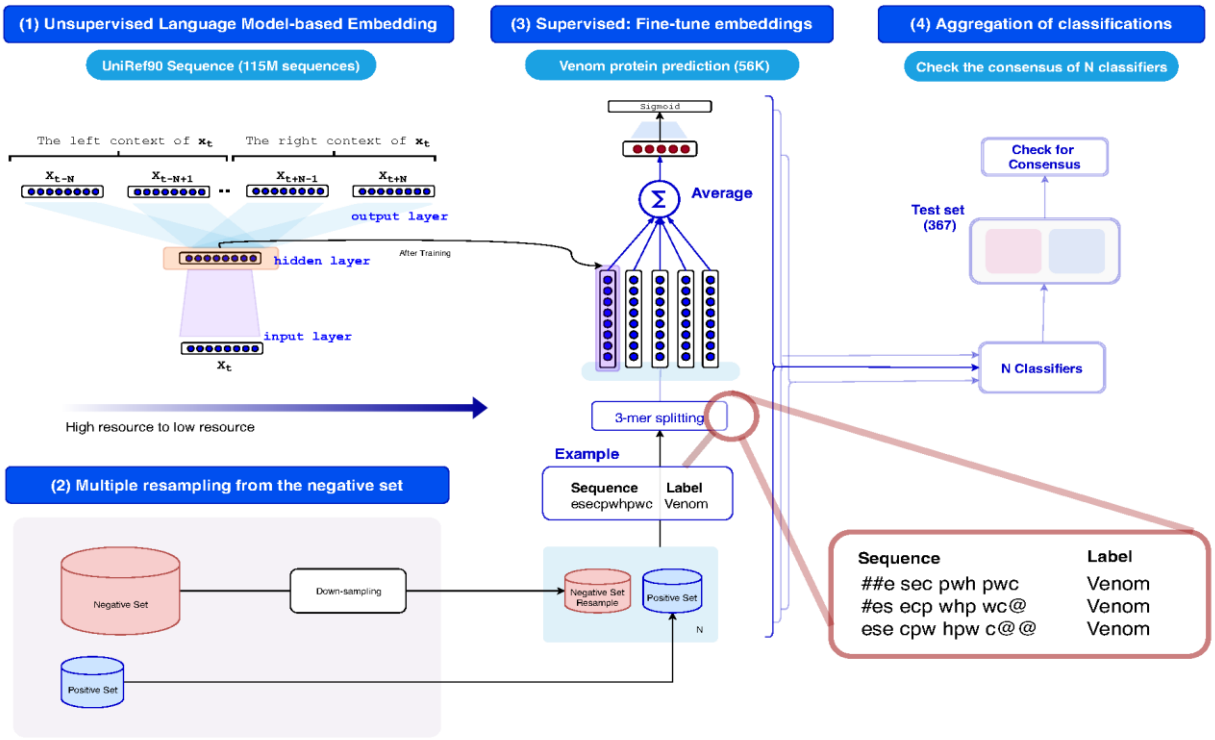

Figure 1. Overview of the *ToxVec* approach for venom protein detection using fine-tuned language model representations. The workflow consists of: **(Training phase)** (1) unsupervised training of *Skip-gram* embeddings on UniRef90, (2) multiple resamplings of the negative class, and (3) supervised fine-tuning for venom classification; and **(Inference phase)** (4) ensemble aggregation across $N = 10$ classifiers to produce the final prediction.

the skip-gram framework employs a softmax function over the vector space of possible contexts. Specifically, the conditional probability of observing a context word w_c given the target w_t is parameterized as:

$$P(w_c|w_t;\theta) = \frac{e^{v_c \cdot v_t}}{\sum_{c' \in V} e^{v_{c'} \cdot v_t}}$$

Where V denotes the vocabulary of possible context tokens. Since computing this full softmax is computationally expensive, negative sampling is employed as an efficient approximation. Under negative sampling, Equation 1 becomes:

$$\sum_{t=1}^{T} \left[\sum_{c \in [t-N, t+N]} \log\left(1 + e^{-s(w_t, w_c)}\right) + \sum_{w_r \in N_{t,c}} \log\left(1 + e^{s(w_t, w_r)}\right) \right]$$

Where $N_{t,c}$ is a set of randomly sampled negative examples (non-context words) for target w_t, and $s(w_t, w_c) = v_t v_c$ denotes their embedding similarity [33].

The use of *Skip-gram* for protein sequences and its application to transfer learning in protein informatics have been demonstrated in several studies [22, 34, 35]. Here, we extend this paradigm to the identification of venom proteins, where learned embeddings capture both sequence-level semantics and latent biochemical context.

Overview of Approach

We introduce *ToxVec*, a transfer learning–based framework that applies language model representations to classify venom peptides. The overall computational workflow is shown in Figure 1 and consists of four main steps:

1. **Unsupervised Training of Language Model–Based Embeddings.** We begin by training a protein k-mer representation using the ProtVec approach [22]. For this study, we utilized an

updated version of ProtVec, trained on the UniRef90 database (approximately 115 million sequences), which extends beyond the earlier Swiss-Prot dataset (*approximately 0.5 million* sequences). Each protein sequence is divided into nonoverlapping k-mers (k=3), with two start symbols (##) and two end symbols (@@) appended. To capture all positional variations, three offsets of splitting are performed, effectively tripling the corpus size to ≈ 345M k-mer (k=3) sequences. The *Skip-gram* model is then trained with a context window size of 20 and an embedding dimension of 3000.

2. **Multiple Resampling of the Major (Negative) Class.** In the benchmark dataset from [10], the negative class is approximately eight times larger than the positive class, resulting in a significant class imbalance. To mitigate this, we downsample the negative examples to match the number of positive sequences and generate N = 10 independent resamples. Each resampled dataset is used to train a separate classifier, thereby enhancing robustness and reducing sampling bias.
3. **Supervised Fine-Tuning of Embeddings for Venom Classification.** For each resampled training set, we fine-tune the pretrained embeddings within a classification network based on the fastText architecture [21]. The input k-mer embeddings are averaged and then passed through a feed-forward layer, followed by a sigmoid activation function, to yield class probabilities. To evaluate the contribution of pretraining, we compare fine-tuned ProtVec-initialized embeddings against randomly initialized embeddings trained from scratch.

   During inference, test sequences are also divided into three offset-based k-mer (k=3) representations (e.g., esecpwhpwc → (1) ##e, sec, pwh, pwc; (2) #es, ecp, whp, wc@; (3) ese, cpw, hpw, c@@). Each segmentation produces an independent classification, and the final label for a sequence is determined by majority voting across the three outcomes.
4. **Ensemble of Multiple Resampled Classifiers.** As described above, we train N = 10 classifiers, each on a different downsampled subset of negatives. The final ensemble prediction assigns a positive label to a sequence only if all N models agree, ensuring high precision in venom identification.
5. **Screening Anticancer Peptides using ToxVec.** To examine the intersection between venom peptides and anticancer bioactivity, we screened sequences from CancerPPD2 [31], a curated repository of experimentally validated anticancer peptides and proteins. Using the trained ToxVec classifier, we evaluated these peptides to determine the proportion exhibiting venom-like molecular signatures. This analysis provides a computational perspective on the structural and functional convergence between venom-derived and anticancer peptides, highlighting candidates that may mimic or evolve from toxin-related bioactivities.

RESULTS

Performance Summary

ToxVec demonstrated improved classification performance for both venom (pos- itive) and non-venom (negative) classes, achieving a macro-F1 score of 0.89, which further increased to 0.93 with ensembling. In screening contexts, this balance between sensitivity and specificity is essential: a higher negative-class F1 score reduces false positives when prioritizing peptides for experimental validation, while a strong positive-class F1 score maintains sensitivity to bioactive motifs potentially associated with genotoxic or proliferative mechanisms. Moreover, the organization of *k*-mers in the learned embedding space (e.g., gradients in hydrophobicity and flexibility) provides a mechanistic basis for interpreting model predictions and communicating model behavior to domain experts in biomedical research and oncology.

Table 1 summarizes the venom protein classification results on the *Toxify* test set across four models: *ToxVec*, *Toxify*, *ToxClassifier*, and *ClanTox*. We report accuracy, F1 scores for both positive and negative classes, and their macro-average. *ToxVec* outperformed all baselines, improving the macro-F1 score by approximately three percentage points (from 0.86 to 0.89) compared to *Toxify*. Incorporating

negative-set resampling and ensembling further increased the macro-F1 to 0.93, representing a 7% absolute improvement over the previous state of the art.

Figure 2 visualizes test and training cases as well as misclassified instances across different models. The figure indicates that many test sequences differ substantially from typical training instances, underscoring the non-trivial nature of the venom classification task within the learned embedding space.

Visualization of Learned Embeddings

To better understand the biochemical structure of the learned embedding space, we utilized t-SNE [36] to visualize protein trimer embeddings (Figure 3). Trimers represented by 3000-dimensional vectors were projected into two dimensions, revealing coherent organization according to physicochemical properties.

Each trimer was annotated with averaged biophysical attributes, including mean molecular weight, mean surface accessibility [37], Kyte– Doolittle hydrophilicity [38], mean flexibility [39], and Janin interior-to-surface transfer energy [40]. All scales were standardized (zero mean, unit variance) for comparability. As shown, *k*-mers with similar physicochemical characteristics cluster closely in the embedding space, suggesting that the learned representation captures intrinsic biochemical relationships.

We then projected the *Toxify* training instances by averaging their constituent trimer vectors and mapping them into the same t-SNE space. The bottom-right panel of Figure 3 displays venom sequences in red and non-venom sequences in blue. The visualization indicates that venom peptides are biophysically diverse, occupying a broader manifold compared to non-venom peptides. This diversity aligns with previous findings that venom composition

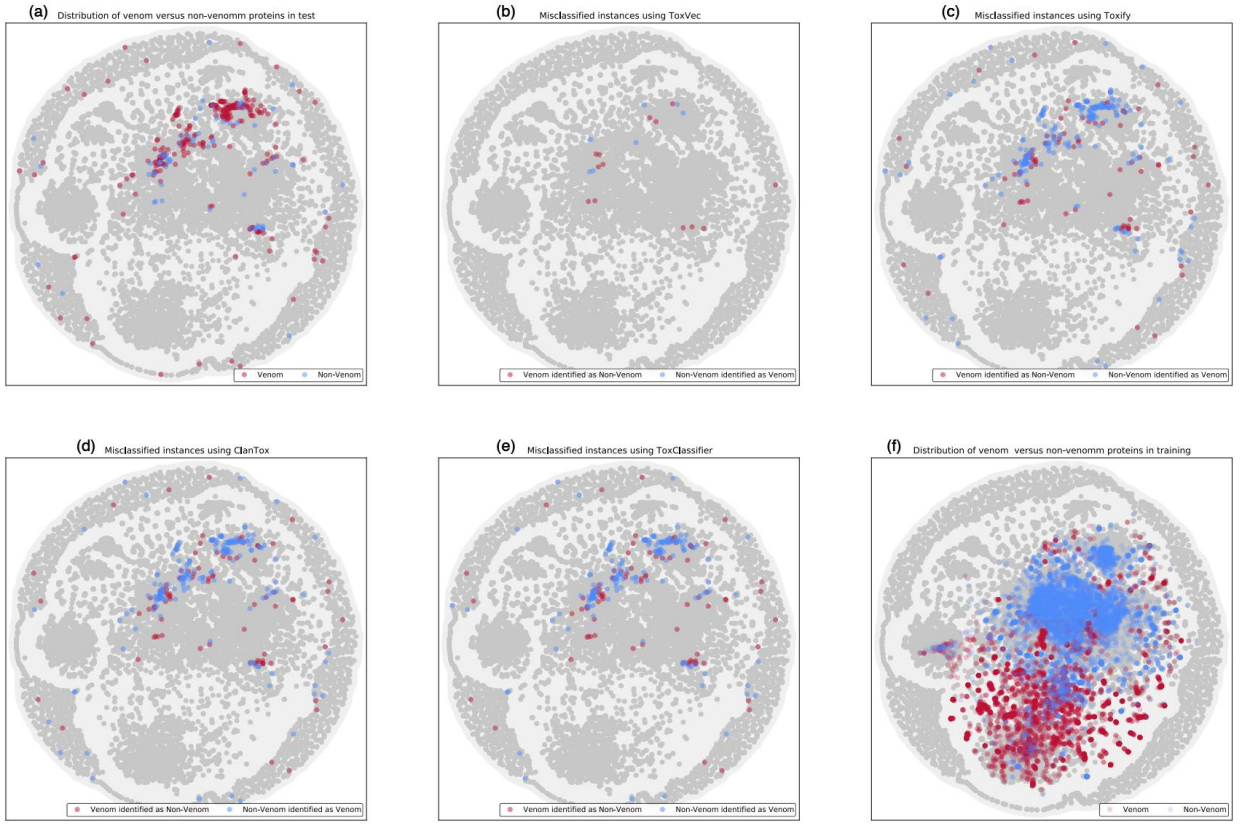

Figure 2. Visualization of test (a), training (f), and misclassified instances (b–e) in the embedding space for various venom prediction models. Results are compared across *ToxVec* (b), *Toxify* (c), *ClanTox* (d), and *ToxClassifier.* (e). Red points indicate venom sequences misclassified as non-venom, and blue points indicate non-venom sequences misclassified as venom. Panels (a) and (f).

Table 1. Performance comparison for venom protein classification on the *Toxify* test set. Accuracy, F1 scores for both classes, and their macro-average are reported. Results are shown for *ToxVec* (with random and UniRef90-based initializations) and baseline methods (*ClanTox*, *ToxClassifier*, and *Toxify*).

Method	Accuracy	F1-positive	F1-negative	Macro-F1
ClanTox	0.79	0.84	0.69	0.77
ToxClassifier	0.73	0.78	0.65	0.72
Toxify	0.86	0.85	0.87	0.86
ToxVec (Random-init)	0.90	0.82	0.93	0.88
ToxVec-Ensemble (Random-init)	0.94	0.87	0.96	0.92
ToxVec (UniRef90-Emb)	0.91	0.84	0.94	0.89
ToxVec-Ensemble (UniRef90-Emb)	**0.95**	**0.89**	**0.96**	**0.93**

can vary widely even within closely related species or snake families [41]. The resulting embedding space thus provides both discriminative and interpretable structure for downstream classification and the generation of biological hypotheses.

Venom-like Signatures in Anticancer Peptides

Screening of the 6,521 anticancer peptides from the CancerPPD2 dataset revealed that a substantial subset displayed high venomity scores when evaluated with *ToxVec*, indicating strong resemblance to venom-derived molecular profiles. The distribution of venomity scores (Figure 4) was bimodal, suggesting the presence of distinct peptide classes with either clear venom-like or non-venom-like characteristics. These results indicate that many anticancer peptides share core physicochemical and structural motifs with venom peptides. This overlap reinforces the hypothesis that toxin-inspired molecular architectures contribute to anticancer efficacy, supporting the broader conceptual connection between venom bioactivity and therapeutic peptide design.

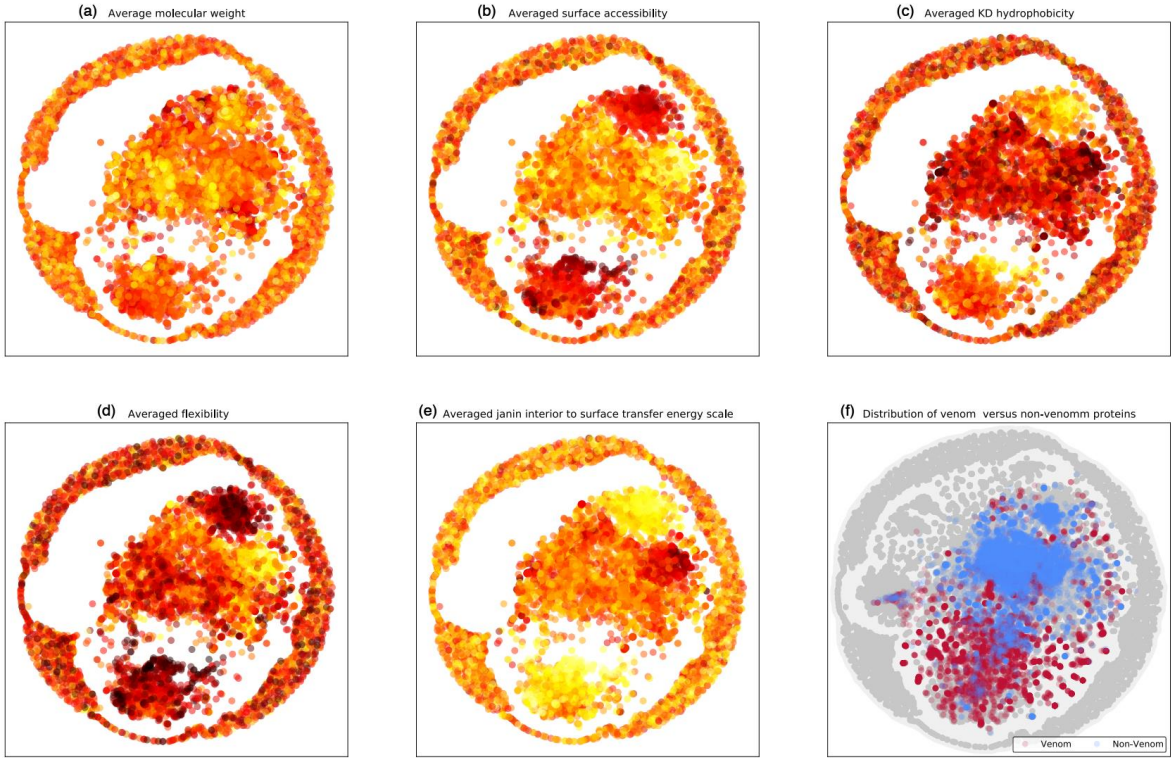

Figure 3. Distribution of biophysical and biochemical properties in the protein trimers (a–e) and distribution of venom versus non-venom sequences (f) visualized in the embedding space using t-SNE. The heatmaps in (a–e) show standardized biophysical property scales averaged across each trimer. Panel (f) highlights venom sequences (red) and non-venom sequences (blue). Higher intensity (lighter color) indicates higher values along each scale.

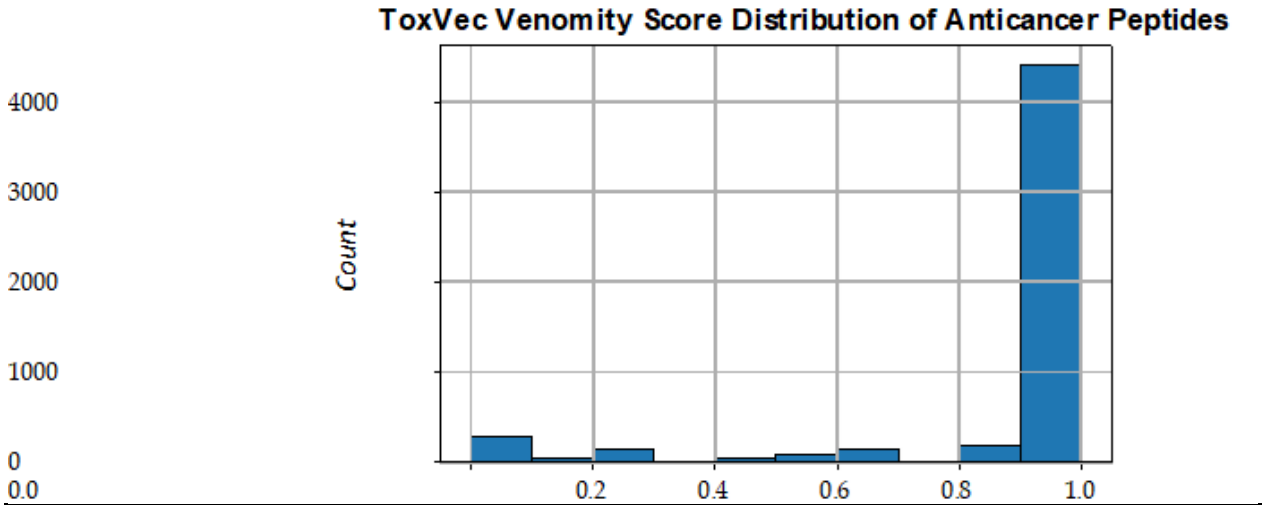

Figure 4. Distribution of *ToxVec*-predicted venomity scores for 6,521 anticancer peptides from the CancerPPD2 dataset. A notable fraction of peptides exhibits high venom-like probabilities (> 0.5), suggesting functional and structural convergence between venom-derived and anticancer peptides.

A Cognitive Mind–Body–Inspired Interpreta- tion of *ToxVec*

Beyond its computational performance, *ToxVec* can be conceptually interpreted through a Cognitive Mind–Body–Inspired AI framework that parallels cognitive representation and biological embodiment. In this view, the latent sequences embedded in the language model correspond to the mind—an abstract, symbolic, and contextual representation of molecular meaning—while the biochemical realization of those sequences in cellular systems corresponds to the *body*. This dual perspective provides a useful lens for understanding how computational representations capture biologically grounded functionality.

From a cognitive standpoint, the protein language model constructs internal representations that resemble semantic encoding in the human brain [20, 42, 43]. Each *k*-mer embedding aggregates contextual information from neighboring residues, forming a distributed "mental space" that encodes physicochemical regularities, evolutionary constraints, and functional cues. These distributed representations enable the model to generalize across diverse protein families, much like human cognition abstracts underlying patterns from sensory inputs.

On the biological side, these representations have direct implications for embodied molecular function. When interpreted in the context of venom and toxin biology, the emergent embedding clusters correspond to distinct physicochemical gradients (e.g., hydrophobicity, charge, flexibility) that influence real molecular behavior such as membrane binding, channel modulation, or receptor interaction [28, 38]. In this sense, the embedding space constitutes a computational "body schema" that mirrors the physical affordances and constraints of peptide structure and activity.

Integrating these views, *ToxVec*, as an instance of deep learning models, exemplifies a form of *computational embodiment*: abstract sequence-level cognition (the mind) gives rise to predictions about concrete biochemical outcomes (the body). This interpretive bridge aligns with emerging ideas in cognitive science and neuroscience, suggesting that cognition is inherently embodied, where symbolic processes emerge from and are constrained by sensorimotor interaction and physical structure [44, 45, 46]. By drawing on this analogy, we situate protein representation learning within a broader framework of embodied intelligence, connecting abstract sequence understanding to molecular function prediction.

Such a cognitive–embodied framework has practical implications for biomedical discovery. Viewing embeddings as a form of "molecular cognition"

provides an interpretable structure for reasoning about latent features relevant to carcinogenicity, membrane disruption, and receptor-mediated signaling—processes central to venom bioactivity and cancer progression. As such, *ToxVec*, as well as all models using latent space embeddings, not only advances protein informatics but also contributes to a conceptual synthesis between artificial intelligence, cognitive neuroscience, and molecular oncology.

CONCLUSIONS AND DISCUSSION

In this work, we presented **ToxVec**, a deep learning framework based on protein language model representations for venom protein identification. We compared the performance of *ToxVec* with recent supervised methods and demonstrated that fine-tuning pre-trained protein language model representations achieves state-of-the-art performance in this task. To address class imbalance in the training data, we employed ensemble learning with multiple models trained on downsampled subsets of the majority (non-venom) class, which further improved the macro-F1 score by 4 percentage points, reaching a final macro-F1 score of 0.93.

Overall, the analysis highlights the inherent difficulty of venom classification due to distributional shifts between training and test data, while demonstrating that *ToxVec* outperforms existing methods by improving F1 scores for both venom and non-venom classes. The lowest macro-F1 achieved by *ToxVec* (0.88), obtained when *k*-mer embeddings were trained from scratch, still exceeded the best baseline score (0.86). Ensembling ten classifiers trained on distinct downsampled subsets of the negative class further increased performance to 0.92. When pre-trained *Skip-gram* embeddings from UniRef90 were used, performance improved again to a macro-F1 of 0.93.

These results demonstrate that automatic feature learning, whether through supervised training from random initialization or fine-tuning of self-supervised embeddings, substantially enhances venom identification compared with methods relying on manually engineered features. Consistent with trends in natural language processing, fine-tuning language model–based representations has improved downstream supervised task performance, particularly in limited-data regimes.

We also applied *ToxVec* to the CancerPPD2 dataset of validated anticancer peptides. The analysis revealed that a notable fraction of anticancer peptides exhibit high venomity scores, suggesting that many therapeutic peptides share physicochemical and structural motifs with natural toxins. This finding supports the hypothesis that toxin-inspired molecular scaffolds may underlie anticancer bioactivity, further emphasizing the translational potential of venom-based computational models.

The success of automatic representation learning in this study motivates future exploration of contextualized embedding architectures, such as transformer-based ones, to further improve interpretability and predictive power in computational toxicology and related biomedical domains.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICS APPROVAL

Not applicable.

REFERENCES

1. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning. 2012;27:17–36. [Available from: https://proceedings.mlr.press/v27/bengio12a.html]

2. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025;16(5):1–72. doi:10.1145/3652427

3. Howard J, Ruder S. Universal Language Model Fine-Tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 328–339. doi:10.18653/v1/P18-1031

4. Goldberg Y, Levy O. Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint. 2014;arXiv:1402.3722. Available from: https://arxiv.org/abs/1402.3722

5. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol. 1988;204(1):155–164. doi:10.1016/0022-2836(88)90606-3

6. Hassabis D, Kumaran D, Summerfield C, Botvinick M. Neuroscience-inspired artificial intelligence. Neuron. 2017;95(2):245–258. doi:10.1016/j.neuron.2017.06.011

7. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins. 1994;19(2):141–149. doi:10.1002/prot.340190207

8. Ahmadi S, Knerr JM, Argemí L, Bordon KCF, Pucca M, Cerni F, et al. Scorpion venom: Detriments and benefits. Biomedicines. 2020;8(5):118. doi:10.3390/biomedicines8050118

9. Jenner RA, von Reumont BM, Campbell LI, Undheim EAB. Parallel evolution of complex centipede venoms revealed by comparative proteotranscriptomic analyses. Mol Biol Evol. 2019;36(12):2748–2763. doi:10.1093/molbev/msz176

10. Maaten Lvd, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–2605. Available from: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

11. Lewis RJ, Garcia ML. Therapeutic potential of venom peptides. Nat Rev Drug Discov. 2003;2(10):790–802. doi:10.1038/nrd1197

12. Gallese V. Embodied simulation: From neurons to phenomenal experience. Phenom Cogn Sci. 2005;4(1):23–48. doi:10.1007/s11097-005-4737-z

13. Wong ES, Hardy MC, Wood D, Bailey T, King GF. SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. PLoS One. 2013;8(7):e66279. doi:10.1371/journal.pone.0066279

14. Chauhan M, Gupta A, Tomer R, Raghava GP. CancerPPD2: an updated repository of anticancer peptides and proteins. Database (Oxford). 2025;2025:baaf030. doi:10.1093/database/baaf030

15. Chen JY, Wang JF, Hu Y, Li XH, Qian YR, Song CL. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. Front Bioeng Biotechnol. 2025;13:1506508. doi:10.3389/fbioe.2025.1506508

16. Pan X, Zuallaert J, Wang X, Shen HB, Campos EP, Marushchak DO, et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. Bioinformatics. 2020;36(13):4222–4231. doi:10.1093/bioinformatics/btaa518

17. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724–1734. doi:10.3115/v1/D14-1179

18. Xiao Y, Zhao W, Zhang J, Jin Y, Zhang H, Ren Z, et al. Protein large language models: A comprehensive survey. arXiv preprint. 2025;arXiv:2502.17504. Available from: https://arxiv.org/abs/2502.17504

19. Naamati G, Askenazi M, Linial M. ClanTox: a classifier of short animal toxins. Nucleic Acids Res. 2009;37(suppl_2):W363–W368. doi:10.1093/nar/gkp240

20. Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol Evol. 2013;28(4):219–229. doi:10.1016/j.tree.2012.10.020

21. Wan F, Zeng JM. Deep learning with feature embedding for compound-protein interaction prediction. bioRxiv. 2016;086033. doi:10.1101/086033

22. Linial M, Rappoport N, Ofer D. Overlooked short toxin-like proteins: a shortcut to drug design. Toxins (Basel). 2017;9(11):350. doi:10.3390/toxins9110350

23. Starcevic A, Moura-da Silva AM, Cullum J, Hranueli D, Long PF. Combinations of long peptide sequence blocks can be used to describe toxin diversification in venomous animals. Toxicon. 2015;95:84–92. doi:10.1016/j.toxicon.2014.12.005

24. Asgari E, McHardy AC, Mofrad MR. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx). Sci Rep. 2019;9(1):3577. doi:10.1038/s41598-019-39813-0

25. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint. 2019;arXiv:1910.03771. Available from: https://arxiv.org/abs/1910.03771

26. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint. 2023;arXiv:2308.08708. [Available from: https://arxiv.org/abs/2308.08708]

27. Prashanth JR, Hasaballah N, Vetter I. Pharmacological screening technologies for venom peptide discovery. Neuropharmacology. 2017;127:4–19. doi:10.1016/j.neuropharm.2017.03.008

28. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One. 2015;10(11):e0141287. doi:10.1371/journal.pone.0141287

29. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on the evaluation of large language models. ACM Trans Intell Syst Technol. 2024;15(3):1–45. doi:10.1145/3639376

30. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc

Comput Linguist. 2017;5:135–146. doi:10.1162/tacl_a_00051

31. Ojeda PG, Ramírez D, Alzate-Morales J, Caballero J, Kaas Q, González W. Computational studies of snake venom toxins. Toxins (Basel). 2018;10(1):8. doi:10.3390/toxins10010008

32. Hargreaves AD, Swain MT, Hegarty MJ, Logan DW, Mulley JF. Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. Genome Biol Evol. 2014;6(8):2088–2095. doi:10.1093/gbe/evu166

33. Clark A. Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford: Oxford University Press; 2010. doi:10.1093/acprof:oso/9780195333213.001.0001

34. Johnson M. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books; 1999.

35. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Cham: Springer; 2018. p. 270–279. doi:10.1007/978-3-030-01424-7_27

36. Cole TJ, Brewer MS. Toxify: a deep learning approach to classify animal venom proteins. PeerJ. 2019;7:e7200. doi:10.7717/peerj.7200

37. Dao FY, Yang H, Su ZD, Yang W, Wu Y, Hui D, et al. Recent advances in conotoxin classification by using machine learning methods. Molecules. 2017;22(7):1057. doi:10.3390/molecules22071057

38. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 2005;102(18):6395–6400. doi:10.1073/pnas.0408677102

39. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS). 2013;26:3111–3119.

40. Gacesa R, Barlow DJ, Long PF. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. PeerJ Comput Sci. 2016;2:e90. doi:10.7717/peerj-cs.90

41. Jungo F, Bougueleret L, Xenarios I, Poux S. The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. Toxicon. 2012;60(4):551–557. doi:10.1016/j.toxicon.2012.03.010

42. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating protein transfer learning with TAPE. In: Advances in Neural Information Processing Systems (NeurIPS). 2019;32:9689–9701.

43. Nawarak J, Sinchaikul S, Wu CY, Liau MY, Phutrakul S, Chen ST. Proteomics of snake venoms from elapidae and viperidae families by multidimensional chromatographic methods. Electrophoresis. 2003;24(16):2838–2854. doi:10.1002/elps.200305517

44. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–132. doi:10.1016/0022-2836(82)90515-0

45. Emini EA, Hughes JV, Perlow D, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol. 1985;55(3):836–839. doi:10.1128/jvi.55.3.836-839.1985

46. Ma R, Mahadevappa R, Kwok H. Venom-based peptide therapy: insights into anticancer mechanism. Oncotarget. 2017;8:100908–100930. doi:10.18632/oncotarget.21757