A Proposed Unified, Scalable Platform for Integrative Research on Venomous Species

Authors: Shaadi Mehr (0000-0008-8687-1232)[1,13], Todd Castoe (0000-0002-5912-1574)[2], Marymegan Daly (0000-0001-6182-0369)[3], Florence Jungo (0000-0002-1509-4332)[4], Kim N. Kirchhoff (0000-0002-0004-1529)[1], Ivan Koludarov (0000-0002-9073-4758)[5], Stephen P. Mackessy (0000-0003-4515-2545)[6], Jason Macrander (0000-0002-1509-4332)[7], Praveena Naidu (0000-0002-5053-6036)[1], Maria Vittoria Modica (0000-0003-2532-6763)[8], Elda E. Sanchez (0000-0001-5222-8873)[9], Giulia Zancolli (0000-0003-3060-2507)[10], Mandë Holford (0000-0001-9476-2687)[1,11,12*]

Affiliations:

1: Dept. of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave, Cambridge, MA 02138

2: Department of Biology, University of Texas at Arlington, Arlington, TX, USA

3: Dept of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus OH USA

4: SIB Swiss Institute of Bioinformatics, Swiss-Prot Group, 1211 Geneva, Switzerland.

5: TUM (Technical University of Munich), School of Computation, Information and Technology (CIT), Faculty of Informatics, Chair of Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

6: Department of Biological Sciences, University of Northern Colorado, Greeley, CO, USA

7: Biology Department, Florida Southern College, Lakeland, FL, USA

8: Stazione Zoologica Anton Dohrn, Department of Biology and Evolution of Marine Organisms, Roma, Italy

9: Department of Chemistry, National Natural Toxins Research Center, Kingsville, TX, USA

10: Department of Ecology and Evolution, University of Lausanne, Lausanne 1015, Switzerland

11: Department of Invertebrate Zoology, The American Museum of Natural History, NY, NY, USA

12: Programs in Biology, Biochemistry, and Chemistry at CUNY Graduate Center, NY, NY, USA

13: Department of Chemistry, Hunter College, City University of New York, NY, NY

*Corresponding authors: Shaadi Mehr; shaadimehr@g.harvard.edu, Mholford@fas.harvard.edu

Abstract

Venomous animal research is hampered by fragmented, specialized, and non-interoperable databases (isolated genomic, proteomic, and ecological data). Despite the immense promise of venomous organisms to yield novel bioactive compounds for pharmacological and evolutionary applications, the informatics landscape for such taxa has remained patchy, lacking macro-scale integration across species. We present VenomsBase, an integrated, modular resource that synthesizes multi-omics data, ecological metadata, and functional annotations for venom-bearing organisms. Following the FAIR guidelines, VenomsBase combines an ontology-driven architecture with big-data cloud workflows for sequence integration, motif clustering, 3D display, and linking ecological metadata. Standardized tools and training modules facilitate worldwide access to resources for both researchers in developed countries and in resource-limited areas. Its plug-and-play design allows for integration of additional analytical modules and extension to other species. One can also examine evolutionary trends and connect venom chemistry to ecological niches. VenomsBase would (i) accelerate the pace of venom discovery, whether for therapeutic purposes or evolutionary significance, by providing validated, cross-referenced data sets and community-driven curation, and (ii) foster an open, just, and innovation-ready venom research ecosystem.

Background and Motivation

Animal venoms and venom systems represent iconic evolutionary novelties with broad biological applications. From the aculei of scorpions to the neurotoxic conotoxins of cone snails and the enzymatic elixirs of snakes and spiders, venom systems play a crucial role in the ecology of venomous organisms through prey procurement, deterrence, and intrapopulation combat. Envenomation by these species has significant impacts on human health and can be fatal. Nevertheless, the components of venoms, which represent a multifaceted pharmacopeia, also offer extraordinary opportunities for the development of both basic and applied scientific knowledge. Modern medicine is, in fact, full of remedies inspired by venom. From the modulation of blood sugar levels with venom-derived GLP-1 receptor agonists, such as Ozempic, based on Gila monster venom, to the treatment of high blood pressure with ACE inhibitors, initially found in Brazilian pit viper venom [1,2]. Evolutionary innovations in venom molecules are of great importance to pharmacology but largely untapped for the discovery of new modulators of cellular physiology and drug development [3].

Despite their ecological, biological, and biomedical importance, resources for studying and harnessing the diverse potential of venomous animal species lack a centralized, comprehensive bioinformatics infrastructure. Investing in robust infrastructure to capture the ecological and biomedical diversity of venoms is both a scientific necessity and a translational opportunity.

Limitations of Existing Resources

Although the field of venom research is supported by pioneering databases such as ConoServer and ArachnoServer, the corresponding bioinformatics ecosystem remains largely fragmented (see Table 1 footnote, Figure 2). Only a handful of curated, venom-specific databases exist. While general repositories such as NCBI and UniProtKB, or specialized aggregators like VenomZone and Tox-Prot (see Table 1 footnote), have contributed significantly to the field, they exhibit several critical limitations, such as:.

- Limited coverage: Most specialized databases are restricted to a single taxon or data type.
- Data incoherence: The absence of standardized metadata and inconsistent formats results in siloed datasets that cannot be easily compared or integrated.
- Limited scalability: Few platforms can manage data across hundreds of species or thousands of sequences.
- High threshold for entry: Many tools require advanced bioinformatics expertise and computational infrastructure, limiting accessibility, particularly for researchers in biodiversity-rich but resource-limited regions.

These limitations restrict proper annotation, comparative analyses, and translational potential, especially in environments where high-performance computing and bioinformatics expertise are scarce.

Venomous animals are estimated to encompass at least 250,000 species across multiple lineages, including approximately 50,000 described spider species and about 850 species of cone snails (see Table 1 footnote). This vast diversity underscores the magnitude of data required for comprehensive comparative analyses and highlights why existing databases remain necessarily fragmented.

To quantify these limitations, we compared three major venom databases: ArachnoServer, ConoServer, and Tox-Prot. We reviewed nine key data dimensions, including total curated entries, species representation, sequence types, post-translational modifications (PTMs), and functional annotations (see Figure 2). The benchmarking analysis reveals substantial variation in data coverage and completeness. Tox-Prot exhibits the broadest taxonomic representation and higher inclusion of PTMs and structural data, whereas ConoServer and ArachnoServer provide taxon-specific depth but limited data diversity. Key information types, such as 3D structures, domain annotations, and analytical tools, remain incomplete or entirely absent across all platforms. These disparities underscore the ongoing fragmentation of venom informatics resources and highlight the need for an integrative framework to harmonize these datasets.

To address these challenges, we propose *VenomsBase,* a centrally managed, scalable platform designed to aggregate, integrate, and harmonize venom-specific data in alignment with FAIR (Findable, Accessible, Interoperable, and Reproducible) principles (see Figure 1). VenomsBase builds upon the complementary strengths of existing venom databases (see Table 1 footnote):

VenomKB v2.0, ArachnoServer, and ConoServer. *VenomKB v2.0* provides a semantic, ontology-driven framework for linking toxin-related biological and pharmacological entities [7]. *ArachnoServer* provides expert-curated sequence and functional data on spider toxins, serving as a benchmark for annotation quality and curation depth. *ConoServer* contributes detailed structural and pharmacological data on cone snail peptides, capturing their diversity and bioactivity profiles.

By integrating these strengths, VenomsBase collects, curates, and organizes venom compound nomenclature, bioactivities, sequences, and structural information within an ontology-based architecture (a framework that links data through shared biological concepts and relationships). This ensures consistent classification and facilitates interoperability across diverse data types, including genomic, transcriptomic, proteomic, pharmacological, ecological, and phylogenetic datasets. Such a curated and interoperable resource will serve as a foundational knowledge base for broad scientific inquiry and translational discovery within and beyond the venom research community.

Proposed Platform Architecture and Its Data Flow

The first phase of the VenomsBase project would focus on platinum-level species (i.e., species for which high-quality genomes, transcriptomes, venom gland proteomes, and detailed ecological metadata are available) [4]. Those anchors could serve as a scalable workflow approach and best practices that could later be applied to other species with poor-quality, fragmented data.

The envisioned data ingestion and validation pipeline would be semi-automated with downloads from databases such as NCBI, UniProtKB, SVDB, ConoServer, T3DB, and ArachnoServer (see Table 1 footnote). After data entry, the platform generates standard file formats, bundles metadata, and performs deduplication, providing this information for downstream analysis (see Figure 1).

VenomsBase would support multiple levels of functional and comparative analysis. Note, in this context, “motif clustering” refers to grouping toxin sequences that share conserved motifs or domains to reveal functional convergence, while “eco-evolutionary contextualization” integrates ecological traits and evolutionary relationships to interpret adaptive trends in venom evolution:.
:

Functional annotation: Use of state-of-the-art tools to predict signal peptides, motifs, and active sites, converting raw sequences into biologically meaningful information (see Table 1).

Motif annotation and co-expression Clustering: Identifying conserved motifs in venom genes with no global homology; clustering the expression patterns across species and tissue types (see Table 1 footnote).

OMICS integration: Building networks around venom-related entities, including gene isoforms, orthologs, and mobile genetic elements, to discover pipelines for new venom

compounds acting as e.g., ligands for molecular probes (ion channels, receptors) and therapeutic leads.

Eco-evolutionary contextualization: To support the contextualization of habitat, prey preference, phylogenetic and morphological information, and, thus, to act as a bridge that connects sequence discoveries with ecology and adaptation [5].

Terminology editor: Through dissemination, write the "glossary" of controlled vocabularies and annotation standards, incentivizing reproducibility, comparability between studies, often in different areas, and communication across the venom branches.

Interactive visualization and training: Genome browser, 3D structure viewer, and visual analytics. It has been supported with online tutorials and training manuals, especially for low-resource settings [6].

To support long-term scalability, the platform is envisioned to adopt a modular, cloud-based approach, leveraging plug-and-play toolkits that can grow in line with evolving research requirements. Here, we would learn from the GBIF (Global Biodiversity Information Facility) and i5K: Sequencing Five Thousand Arthropod Genomes initiatives; VenomsBase would benefit from community governance and transparent update cycles [7]. This model encourages continual input and ensures that infrastructure remains aligned with the rapid pace of discovery in genomics and venom biology.

Platinum-Level Species Definition

To provide clear reference standards for high-quality data integration, VenomsBase prioritizes Platinum-level species: organisms for which genomic, transcriptomic, proteomic, and ecological data meet consistent quality and completeness benchmarks.

Platinum-level genomes are defined as:

- Chromosome-level assemblies with scaffold N50 > 10 Mb
- BUSCO completeness >90%
- Fully annotated gene models integrating empirical RNA-seq (and ideally Iso-Seq) data from venom glands
- Curated venom loci with linked proteomic data
- Rich metadata, including geographic origin and tissue provenance

Platinum-level transcriptomes include:

- Full-length transcript inference via long-read Iso-Seq
- Complete annotation of gene models
- Metadata specifying collection origin and experimental conditions

Examples include *Crotalus viridis*, *Naja naja*, and *Conus geographus*, which exemplify the data standards guiding scalable inclusion of other taxa within VenomsBase.

Use Case Demonstration

To demonstrate the front-end functionality of VenomsBase, we applied its analytical workflow to venom-gland transcriptomic data from *Doryteuthis pealeii* (longfin inshore squid), a cephalopod of ecological and biomedical interest, with results and interactive visualizations available through the VenomsLanding/VenomView demonstration portal [8], which serves as the visualization layer of VenomsBase supporting community engagement, testing, and training. VenomView is registered with the Research Resource Identifier (RRID:SCR_027588). Raw RNA-seq reads were processed through quality control and de novo assembly, generating standardized reports that include N50, contig counts, and BUSCO completeness. Functional annotation identified candidate toxin transcripts using SignalP (see Table 1 footnote) for signal peptide prediction, InterProScan (see Table 1 footnote) for domain and motif assignments, and BLAST (see Table 1 footnote) for curated homology searches against ArachnoServer, ConoServer, and ToxProt. The resulting annotations include metalloproteinases and serine proteases typical of cephalopod venom families. VenomsBase links each annotation to sequences in the database. This demonstration highlights VenomsBase's end-to-end integration, from assembly evaluation to functional annotation and visualization, within a fully reproducible, FAIR-compliant workflow.

Use Cases and Community Impact

Unlike other general biodiversity databases, VenomsBase focuses solely on venomous organisms. Its design would support the particular data structures necessary to investigate venom biology, with species-level molecular integration and evolutionary history. Its centralization would ultimately help resolve data fragmentation by enabling the collection, standardization, and analysis of multi-species venom data through a unified portal. Scalability and adaptability are two key aspects of VenomsBase's design. It would be implemented as modular cloud software that could easily accommodate new species, data types, and analysis tools (VenomsBase). Future image-based analyses may leverage containerized workflows to ensure consistent analyses across different computational environments, as well as heuristic scoring to identify high-confidence records [9].

VenomsBase would also support diverse research applications across disciplines (see Figure 3). Its integrated data framework would enable comparative and translational discovery: evolutionary biologists could analyze venom compound diversification across species; pharmacologists could mine curated venom peptides for therapeutic leads; and computational biologists could reproduce and share workflows using standardized metadata. These examples demonstrate how VenomsBase could function not as a data repository, but as a knowledgebase that is an interactive, reproducible ecosystem that unites data exploration and collaboration.

VenomsBase would also provide tutorials and training modules to enhance its usability and sustainability, with a long-term educational and functional impact. The platform would continuously evaluate its effectiveness through indicators in research, training, species diversity, and community engagement. VenomsBase's network infrastructure would represent a leap in assisting global animal envenomation treatment and venom-related therapeutic development, with the potential to increase global bioeconomic output.

Future Roadmap

The VenomsBase initiative aims to expand both in scope and technical capacity through several core priorities:

1. **Platinum reference species expansion** – Curate high-quality, multi-omics datasets as benchmarks for annotation transfer and cross-species evolutionary studies.
2. **Scalable multi-Omics integration** – Incorporate genomic, transcriptomic, proteomic, and other functional genomic data, alongside population variation.
3. **Ecological data and metadata** – Incorporate ecological and geographic metadata for samples, along with ecological data on species (e.g., diet, habitat, range, etc.).
4. **Enhanced interoperability & FAIR compliance** – Employ persistent identifiers (DOIs) and ontology-based metadata standards to ensure reproducibility and long-term reuse.
5. **Reproducible & shareable workflows and applications** – Support community-contributed, containerized pipelines registered in SciCrunch or other workflow platforms, such as RRID: SCR_027588.
6. **Interactive applications** – Develop tools for dynamic visualization and cross-species filtering, from assembly metrics to gene family exploration.
7. **Biomedical integration & AI prediction** – Interface curated venom datasets with drug discovery and ML models for toxin activity prediction.
8. **Community governance** – Implement distributed curation and transparent update cycles for long-term sustainability.

Together, these steps define next-step goals for expanding VenomsBase, transforming it from a modular prototype into a comprehensive, globally integrated platform for venom systems biology and translational research.

Conclusion

The proprosed VenomsBase would not be an online library of data. Instead, it would be designed as a collaborative infrastructure and knowledgebase for stimulating venom research and the discovery of comparative venom biology and applications of venom compounds. Whether for the identification of new peptides and proteins, insight into evolutionary dynamics, fundamental cellular development and physiology, or potential pharmacological applications for advancing biomedical research, VenomsBase would offer a single, accessible, and user-friendly foundational resource to pursue these goals. By overcoming key barriers in existing venom repositories and data

coordination, VenomsBase would adopt a scalable infrastructure, standardized data strategies, and a community-focused approach. In summary, VenomsBase would establish the basis for an inclusive, global, and integrated FAIR ecosystem to drive the next era of transformative advances in venom biology.

References

1. Munawar A, Ali SA, Akrem A, Azhar A, Al-Sehemi AG, Rehman MU. (2022) Comparative study to characterize the pharmaceutical potential of synthesized snake venom bradykinin-potentiating peptides in vivo. *Curr Med Chem* 29(42):6422–6432. PMID: 35125079.
2. Drucker DJ. (2018) Mechanisms of action and therapeutic application of glucagon-like peptide-1. *Cell Metab* 27(4):740–756. PMID: 29606347.
3. Holford M, Daly M, King G, Norton R. (2018) Venoms to the rescue. *Science* 361(6405):842–844. PMID: 30166472.
4. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci U S A* 115(17):4325–4333. PMID: 29686065.
5. Casewell NR, Wagstaff SC, Wüster W, Cook DAN, Bolton FMS, King SI, et al. (2020) Ecological diversification and its effects on venom variation in snakes. *Proc Biol Sci* 287(1920):20200415. PMID: 32475354.
6. Goecks J, Nekrutenko A, Taylor J; Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11(8):R86. PMID: 20738864.
7. Nilsen TW, Graveley BR. (2017) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Trends Genet* 33(9):581–583. PMID: 28689651.
8. VenomsLanding/VenomView demonstration portal. Available at: https://venomsbase-dashboard.shinyapps.io/VenomLanding. Accessed 11 Dec 2025.
9. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol* 35(4):316–319. PMID: 28398311.

Figure Legends

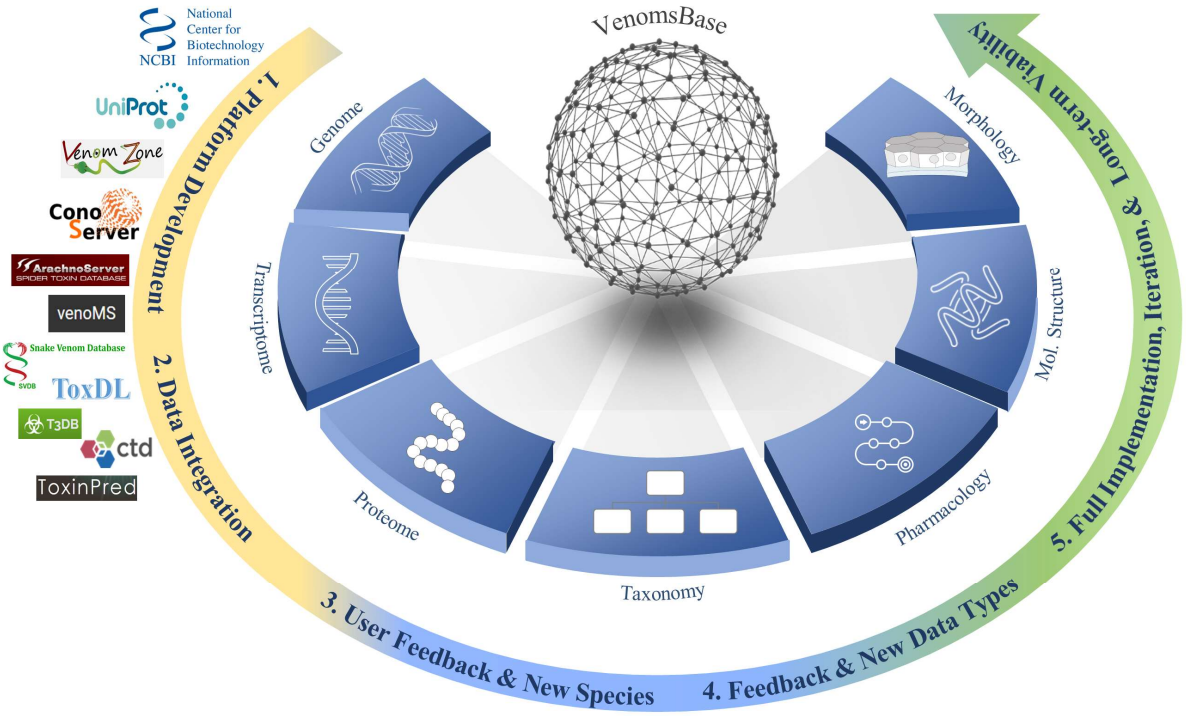

Figure 1. VenomsBase vision and development stages.
VenomsBase is envisioned as a centralized, standardized, and curated web-enabled knowledgebase to support the rapidly evolving field of venomics, with high potential for supporting fundamental and translational research including biotechnological applications in medicine, agriculture, and biomaterials. This vision is realized in three progressive stages: (I) platform development and data integration (yellow), (II) implementation of analytical tools (blue), and (III) acquisition of new data and community-driven expansion to meet research needs (green).
Short caption: Developmental stages and vision of the VenomsBase platform.

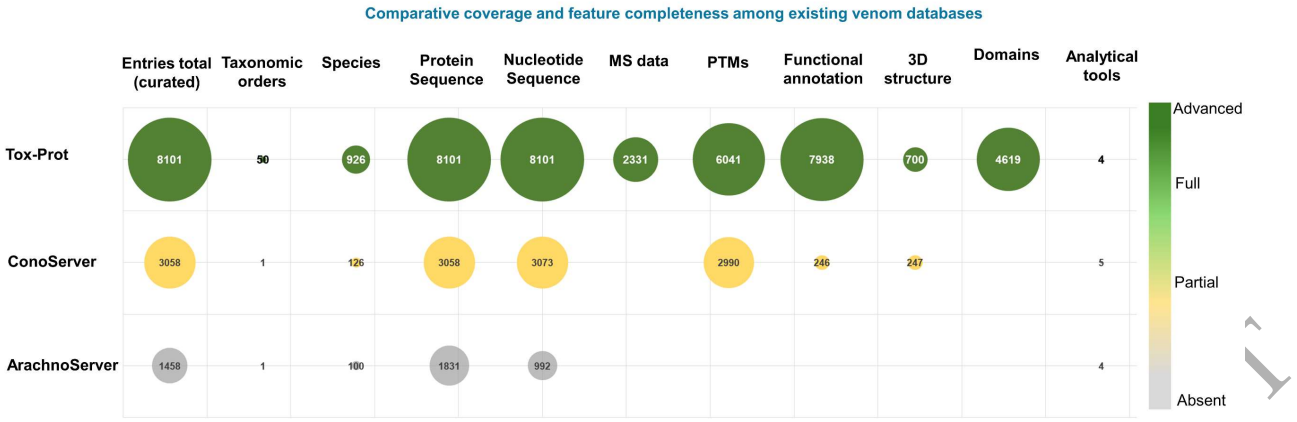

Figure 2. Comparative coverage and feature completeness among existing venom databases.

Relative data abundance and feature coverage across Tox-Prot, ConoServer, and ArachnoServer for nine core data categories: 1) total curated entries, 2) taxonomic diversity, 3) protein and nucleotide sequence coverage, 4) mass-spectrometry (MS) data, 5) post-translational modifications (PTMs), 6) functional annotation, 7) 3D structural information, 8) protein-family domain data, and 9) available analytical tools. Bubble size corresponds to entry abundance, while color shading indicates completeness (Advanced = dark green, Partial = yellow, Absent = gray). The figure highlights current coverage gaps and the need for an integrated, standardized platform such as VenomsBase.

Short caption: Benchmarking data coverage across existing major venom databases.

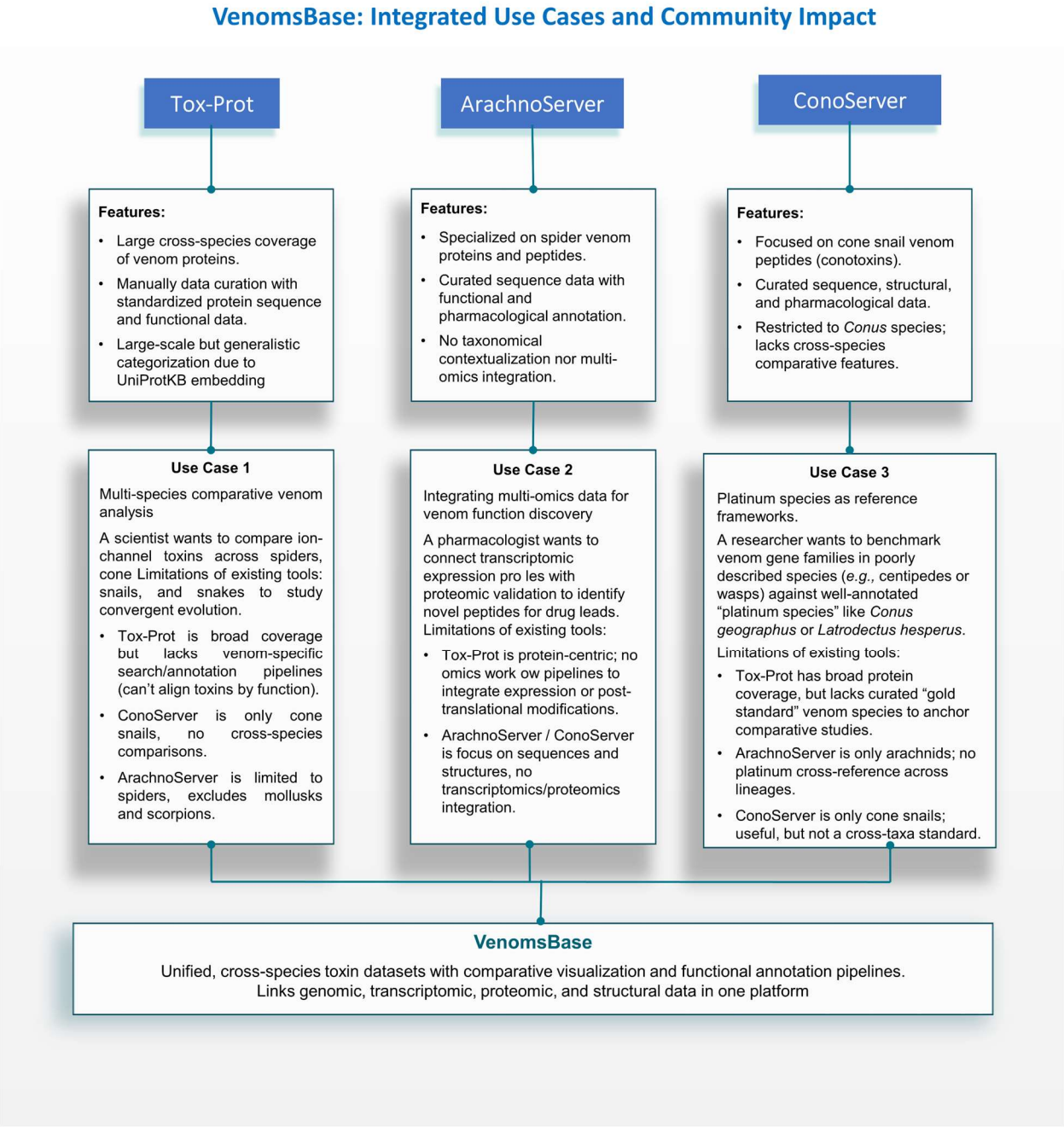

Figure 3. Comparative use cases and data integration overview in VenomsBase. VenomsBase unifies multi-species venom data and overcomes the taxonomic and functional limitations of existing databases. While ArachnoServer and ConoServer are restricted to specific venomous lineages (arachnids and cone snails), and Tox-Prot provides broad but generalist coverage, VenomsBase enables cross-species comparative analyses linking genomic, transcriptomic, proteomic, and structural data within a single FAIR-compliant framework. The

use case demonstrates functional comparison of ion-channel toxins across taxa through unified visualization and annotation pipelines.

Short caption: Integrated cross-species data and use cases enabled by VenomsBase.

Table 1. Key resources integrated into the VenomsBase infrastructure.

Tool / Database	URL	Role in VenomsBase
ConoServer[a]	https://www.conoserver.org	Reference database for conopeptide sequences and annotations
ArachnoServer[b]	http://www.arachnoserver.org	Spider toxin database and annotation reference
NCBI (GenBank)[c]	https://www.ncbi.nlm.nih.gov	Core repository for genomic, transcriptomic, and annotation data
UniProtKB[d]	https://www.uniprot.org	Protein knowledge base and functional annotation source
ToxProt[e]	https://www.uniprot.org/program/Toxins	Curated venom protein subset from Swiss-Prot
VenomZone[f]	https://www.venomzone.expasy.org	Structured taxonomy browser for venom-linked resources
VenomKB v2.0[g]	https://venomkb.org	Semantic resource linking venom proteins to biomedical context
EchinoBase[h]	https://www.echinobase.org	Genomics platform model informing backend architectural strategy
SVDB[i]	https://github.com/J35P312/SVDB	Structural variation schema and representation toolkit
T3DB[j]	http://www.t3db.ca	Toxic compound and exposure metadata reference
InterProScan[k]	https://www.ebi.ac.uk/interpro	Protein domain, motif, and functional annotation pipeline
DISOPRED3[l]	http://bioinf.cs.ucl.ac.uk/psipred/	Intrinsically disordered region prediction platform
HMMER[m]	http://hmmer.org	Domain and motif similarity search algorithm
Galaxy[n]	https://galaxyproject.org	Workflow platform supporting reproducible computation
SignalP[o]	https://services.healthtech.dtu.dk/service.php?SignalP-6.0	Prediction of signal peptides and secretion pathways for toxin annotation
BLAST[p]	https://blast.ncbi.nlm.nih.gov/Blast.cgi	Sequence similarity searches against reference toxin databases (ArachnoServer, ConoServer, ToxProt)

[a] Kaas Q et al., 2012, Nucleic Acids Res.; [b] Pineda SS et al., 2018, Bioinformatics; [c] Sayers EW et al., 2025, Nucleic Acids Res.; [d] UniProt Consortium, 2025, Nucleic Acids Res.; [e] Jungo F & Bairoch A., 2005, Toxicon; [f] Zancolli G et al., 2024, Gigascience; [g] Romano JD & Tatonetti NP., 2015, Sci Data; [h] Telmer CA et al., 2024, Genetics; [i] Eisfeldt J., 2017, GitHub (SVDB repository); [j] Wishart DS et al., 2015, Nucl Acids Res.; [k] Jones P et al., 2014, Bioinformatics; [l] Jones DT & Cozzetto D., 2015, Bioinformatics; [m] Eddy SR., 2011, PLOS Comput Biol; [n] Goecks J et al., 2010, Genome Biol; [o] Almagro-Armenteros JJ et al., 2019, Nat Biotechnol; [p] Altschul SF et al., 1990, J Mol Biol.**Context Footnote:** Koch TL et al., 2024, Mol Biol Evol.; Jenner RA et al., 2025, Trends Ecol Evol.; Dresler J et al., 2024, Front Arachn Sci.

Acknowledgements:

This work and the open accessibility were partially supported by National Institutes of Health – Director's Pioneer Award [5DP1AT012812] and the TUFCCC/HC Regional Comprehensive Cancer Health Disparity Partnership [U54CA221704(5)] to MH. TAC and SPM were supported by National Science Foundation Grant IOS-IOS-2307044 Manuscript contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.