

Web of venom: exploration of big data resources in animal toxin research

Giulia Zancolli^{1,2,*†}, Björn Marcus von Reumont^{3,4,†}, Gregor Anderluh⁵, Figen Caliskan⁶, Maria Luisa Chiusano^{7,8}, Jacob Fröhlich⁹, Evroula Hapeshi¹⁰, Benjamin-Florian Hempel¹¹, Maria P. Ikonomopoulou¹², Florence Jungo¹³, Pascale Marchot¹⁴, Tarcisio Mendes de Farias¹⁵, Maria Vittoria Modica¹⁶, Yehu Moran¹⁷, Ayse Nalbantsoy¹⁸, Jan Procházka¹⁹, Andrea Tarallo²⁰, Fiorella Tonello²¹, Rui Vitorino²², Mark Lawrence Zammit^{21,22}, and Agostinho Antunes^{23,24,*}

¹Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

²SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

³Goethe University Frankfurt, Faculty of Biological Sciences, 60438 Frankfurt, Germany

⁴LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany

⁵Department of Molecular Biology and Nanobiotechnology, National Institute of Chemistry, 1000 Ljubljana, Slovenia

⁶Department of Biology, Faculty of Science, Eskisehir Osmangazi University, 26040 Eskisehir, Turkey

⁷Department of Agricultural Sciences, University Federico II of Naples, 80055 Portici, Naples, Italy

⁸Department of Research Infrastructures for Marine Biological Resources, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

⁹Veterinary Center for Resistance Research (TZR), Freie Universität Berlin, 14163 Berlin, Germany

¹⁰Department of Health Sciences, School of Life and Health Sciences, University of Nicosia, 1700 Nicosia, Cyprus

¹¹Madrid Institute of Advanced Studies in Food, Precision Nutrition & Aging Program, 28049 Madrid, Spain

¹²SIB Swiss Institute of Bioinformatics, Swiss-Prot Group, 1211 Geneva, Switzerland

¹³Laboratory Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille University, Centre National de la Recherche Scientifique, Faculté des Sciences, Campus Luminy, 13288 Marseille, France

¹⁴Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, 00198 Rome, Italy

¹⁵Department of Ecology, Evolution and Behavior, Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, 9190401 Jerusalem, Israel

¹⁶Engineering Faculty, Bioengineering Department, Ege University, 35100 Bornova-Izmir, Turkey

¹⁷Laboratory of Transgenic Models of Diseases, Institute of Molecular Genetics of the Czech Academy of Sciences, 252 50 Vestec, Czech Republic

¹⁸Institute of Research on Terrestrial Ecosystems (IRET), National Research Council (CNR), 73100 Lecce, Italy

¹⁹Neuroscience Institute, National Research Council (CNR), 35131 Padua, Italy

²⁰Department of Medical Sciences, iBiMED, University of Aveiro, 3810-193 Aveiro, Portugal

²¹Department of Clinical Pharmacology & Therapeutics, Faculty of Medicine & Surgery, University of Malta, 2090 Msida, Malta

²²Malta National Poisons Centre, Malta Life Sciences Park, 3000 San Gwann, Malta

²³CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, 4450-208 Porto, Portugal

²⁴Department of Biology, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

*Correspondence address. Giulia Zancolli, Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

E-mail: giulia.zancolli@gmail.com; Agostinho Antunes, CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto,

Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208 Porto, Portugal. E-mail: aantunes@ciimar.up.pt

†First co-authors.

Abstract

Research on animal venoms and their components spans multiple disciplines, including biology, biochemistry, bioinformatics, pharmacology, medicine, and more. Manipulating and analyzing the diverse array of data required for venom research can be challenging, and relevant tools and resources are often dispersed across different online platforms, making them less accessible to nonexperts. In this article, we address the multifaceted needs of the scientific community involved in venom and toxin-related research by identifying and discussing web resources, databases, and tools commonly used in this field. We have compiled these resources into a comprehensive table available on the VenomZone website (<https://venomzone.expasy.org/10897>). Furthermore, we highlight the challenges currently faced by researchers in accessing and using these resources and emphasize the importance of community-driven interdisciplinary approaches. We conclude by underscoring the significance of enhancing standards, promoting interoperability, and encouraging data and method sharing within the venom research community.

Keywords: venom resources, toxin databases, machine learning, drug discovery, antivenom, proteomics, peptidomics, transcriptomics, genomics

Background

Venomous organisms possess the remarkable ability to synthesize and deliver potent cocktails of bioactive compounds known as venoms, which can elicit profound physiological effects in

other organisms. These complex mixtures of proteins, peptides, small organic molecules, and inorganic elements have undergone millions of years of evolution, primarily driven by selective pressure such as predation or defense [1]. Animal venoms have

Received: May 14, 2024. Revised: July 1, 2024. Accepted: July 13, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

captivated human curiosity for centuries, and recently, technological advancements in diverse research fields, especially in molecular biology, have propelled an increasing interest within the scientific community. This has attracted attention from industry, which recognizes the opportunities presented by animal toxins as drug candidates [2–7], diagnostic tools [8, 9], biopesticides, antimicrobial and antiparasitic agents [10, 11], and biological markers to study human physiology [12, 13].

Modern venom research is thus highly multidisciplinary, and it requires the ability to manipulate and analyze a heterogeneous array of data [14]. The emergence and integration of multiomics technologies such as proteomics, transcriptomics, and, more recently, whole-genome data has revolutionized the characterization of venom components and highlighted their biotechnological potential [14]. Despite the abundance of venom research methods, tools, and resources, their scattered nature limits their comprehensive utilization. Addressing this challenge requires centralized and coordinated web-based resources that could serve as repositories of data and knowledge, facilitating the seamless utilization of analytical tools, bioinformatics pipelines, and related databases, ultimately driving cutting-edge venom research.

In this article, we address the multifaceted requirements of the scientific community by discussing web resources, databases, and tools generally used in venom- and toxin-related research. We compiled them into a comprehensive, interactive table freely available on VenomZone [15]. To gather insights into the most prevalent resources used by both novice and seasoned venom researchers, we carried out a survey targeting the members of the European Venom Network (EUVEN) COST Action CA19144 [16] and the participants of the First International Congress of the EUVEN held virtually in September 2021. While this survey primarily focused on European researchers, limiting its comprehensiveness of the global venom research landscape, it served as a springboard to populate our resource list. More important, it enabled us to identify the key challenges and needs faced by venom scientists. Here, we highlight these challenges and discuss the necessity for user-friendly tools and innovative, community-driven approaches. Furthermore, we emphasize the importance of raising standards, enhancing interoperability, and promoting data and method sharing within the field of venom research. Lastly, we spur the idea to compose, curate, and mine a unified venom-specific database that would report venoms and toxins of diverse animal species, including genome architecture and function, whole proteome composition, toxin targets, mechanism of action, and ecological and evolutionary data.

Resources in Venom Research: State-of-the-Art

Overview of main web resources

The cornerstone of virtually any venom research endeavors entails the identification of venom compounds, encompassing their compositional diversity (e.g., protein families), variability (e.g., intra- and interspecies, sex-linked, seasonal, environmental), evolutionary traits, mode of action, and toxicity attributes (e.g., neurotoxicity, hemolytic potency, enzymatic activity, LD50, ED50, clearance rates). This initial step heavily relies on information found in several biological databases (Fig. 1).

The raw data are generally deposited in generalist repositories such as the Proteomics IDentification (PRIDE) database for mass spectrometry data [17] or the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA), and the NCBI GenBank

for nucleic acid data. Nucleotide sequences can also be found in venom-specific databases like ArachnoServer [18] and ConoServer [19], which additionally provide protein sequences, classification of gene superfamilies, cysteine frameworks, information on pharmacological activities of toxins, and sequence analysis tools (see following section). Amino acid sequences derived from direct sequencing or from translated nucleotide sequences are mostly available in 2 generalist databases, UniProtKB and NCBI protein. The Tox-Prot annotation project of UniProtKB/Swiss-Prot provides access to venom protein sequences and links to additional web resources [20]. Considering tools, UniProtKB supports BLAST searches (otherwise directly available on the NCBI website), sequence alignment, searches for similar proteins, and links to various features in the ExPasy Resource Portal [21]. The species from which the data originate are generally reported in the metadata and linked to taxonomy databases such as NCBI or UniProtKB Taxonomy (Supplementary Information Table S1).

The 3-dimensional (3D) structure of peptides and proteins is important to understand their function and mode of interaction with their molecular targets. The most comprehensive databases holding structural information are the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB) [22], the Biological Magnetic Resonance Data Bank (BMRB) [23], and the Electron Microscopy Data Bank (EMDB) [24]. The structures in PDB are primarily determined through X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy and increasingly by cryo-electron microscopy (cryo-EM), although the latter is only from molecules or molecular complexes with masses less than 100 kDa. BMRB is a database of NMR spectroscopic data from peptides, proteins, nucleic acids, and other biologically relevant molecules, while EMDb archives 3D maps of biological specimens from transmission electron microscopy experiments. Cryo-EM holds great potential for investigating toxin-receptor binding [25, 26]. Visualization of toxin 3D structures is provided in ArachnoServer and ConoServer, as well as in UniProtKB. Additionally, the AlphaFold Protein Structure Database [27] provides access to over 200 million 3D structures predicted by AlphaFold, an artificial intelligence (AI) system developed by Google DeepMind based on a neural network model [28].

A wide array of specialized databases for researchers interested in exploring biological pathways (e.g., the KEGG [29]), gene function classification (e.g., the Gene Ontology [GO] Resource [30]), or more specific information on compounds (e.g., PubChem [31], KalliumDB [32], SrepeYard, KNOTTIN [33, 34]) is discussed in the sections below and listed in Supplementary Information Table S1.

Currently, information on venoms and toxins is dispersed across a multitude of resources, both generalists and specialists, each offering varying types of data and occasionally resulting in redundancy. This scenario presents both advantages and disadvantages. On one hand, the proliferation of openly accessible data represents a goldmine for basic as well as applied research. Conversely, differences in data formats and content between disparate sources make it challenging to aggregate information and sometimes result in inconsistencies. For instance, the annotation related to the mature and precursor sequence of a toxin might differ between a generalist database like UniProtKB, which provides the amino acid sequence of a whole gene, and a venom-specialist database like Arachnoserver or ConoServer, which is instead focused on reporting the active, mature sequence [35].

An additional inconvenience in the database landscape is that some have become obsolete (e.g., SCORPION2 [36]), while others offer limited utility (e.g., ATDB [37] primarily available in Chinese) or are at times unavailable (e.g., ArachnoServer),

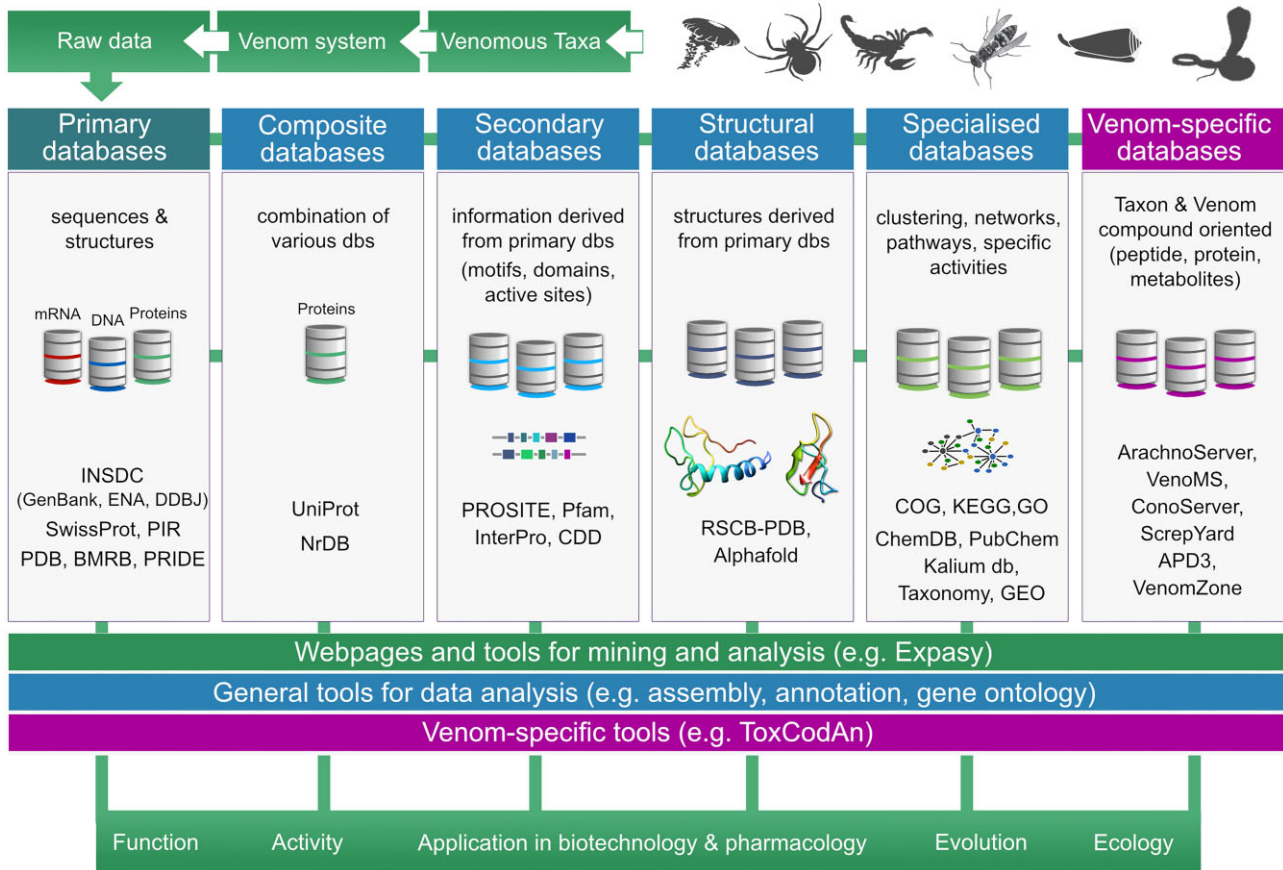


Figure 1: Specialized and generalist web resources, databases, and tools used in venom research. In a typical venom research workflow, raw data from venoms or venom glands are deposited in primary databases, and information is generally subsequently stored in secondary and specialized databases. Such information can be accessed and analyzed using different tools for a variety of research purposes. Dbs = databases.

highlighting the need to constantly curate the available databases [35]. Nonetheless, enduring venom-specific databases and resources include ConoServer, VenoMS [38], T3DB [39], or Tox-Prot. Furthermore, VenomZone [40] is a free web resource that provides information on venoms from 6 major venomous taxa (i.e., snakes, scorpions, spiders, cone snails, sea anemones, and insects), as well as on their molecular targets. Information is structured and accessible through pages on taxonomy (~170 pages), activity (~50 pages), and venom protein families (~40 pages). Each page also provides links to the corresponding proteins in Tox-Prot, classified by species or protein family. Importantly, VenomZone is consulted by around 2,000 visitors every month (average from January to May 2024) and has been regularly updated since its creation in 2015.

Many of the aforementioned websites include some tools for predicting mature peptide boundaries, pharmacological activity, theoretical molecular mass, and so on, while generalist web-based portals (e.g., Expaty [21] and Galaxy [41]) provide comprehensive resources for the analysis of gene expression data, structural biology, text mining, machine learning, and more.

Resources in genomics

Genomics is increasingly playing a central role in venom research. The advancements and decreasing costs of sequencing technologies have facilitated the availability of genome data from venomous species; consequently, genomics has become indispensable for elucidating the complexity of venom-related genes. In-

deed, genomic information is crucial to assess whether divergence in venom composition among species or populations arises from variation in gene copy number, nucleotide sequence, or regulation of gene expression [42–45].

One major advantage of genome data is that it eliminates artifacts from *de novo* proteo-transcriptomics, providing highly accurate results for predicting venom genes and identifying gene and protein variants, including all related transcript and protein-based modifications [46]. To achieve this, transcriptomic data can be assembled using a genome-guided transcriptome assembly approach (e.g., Trinity assembler [47]). Typically, the preferred method for creating genomes is to map transcripts against the genome sequences (scaffolds) with aligners such as BOWTIE2 [48] and splice-aware tools like HISAT2 [49], STAR, [50] and Tophat2 [51, 52] (although no longer supported). High-quality or reference genomes are generally annotated using transcriptomes from multiple tissue samples, comprehensively identifying most gene variants, which is especially relevant to properly characterize multi-gene families like many venom proteins [46].

Generating genomes involves using a plethora of tools and software, primarily command line based due to the specificity, computational demands, and challenges associated with genome analysis [53, 54]. Several pipelines have been developed by genome consortia, and the recently developed automated pipeline in Galaxy is expected to revolutionize the pace of reference genome production and annotation [55].

Resources and tools related to genomic data are currently not widely available in a venom-related context. However, there are

several web resources for accessing genomes, with NCBI Genome being the primary platform that provides genomic data in conjunction with their respective publications. While NCBI offers a comprehensive collection of genomes, some thematic databases, such as Ensembl Metazoa, focus specifically on metazoan reference genomes and offer more tailored data and information [56]. Additionally, Ensembl provides cross-genome resources, annotations, synteny, and other features, with the added benefit of being more accessible than NCBI through a server and application program interface (API) service. Many genome sequencing consortia, such as G10K, GIGA, i5K, B10K, VGP, EBP, DTOL, T2T, and ERGA, provide prepublication information on their planned genomes through dedicated websites, often including unpublished data [57–63]. For example, GenomeArk houses hundreds of high-quality reference genomes and assembly data.

Arguably, the venomous organisms benefiting from the richest genomic resources are Cnidaria (sea anemones, corals, hydroids, and jellyfish). The original reason for the construction of these datasets was the use of several cnidarian species as models for evolutionary developmental biology (“evo-devo”) [64–67] and the specific importance of reef-building corals for marine ecology [68–70]. The availability of these chromosome-scale assemblies, along with rich datasets on small RNA sequencing [71, 72], Chromatin immunoprecipitation sequencing of histone modification marks, and transcriptional regulator proteins [73, 74] for several key species, makes them an excellent resource for studying venom regulatory genomics and evolution. Some of these data can be easily accessed through the SIMRbase genome portal of the Stowers Institute for Medical Research and the *Hydra* 2.0 Genome Project Portal of the National Institutes of Health (NIH).

Despite these advancements, challenges persist in annotating and analyzing toxin-coding genes, as many venom components are part of large, multigene families, and gene comparison tools typically perform better for single-copy genes. Recent studies have demonstrated that analyzing the genome structure and arrangements of genes and their flanking regions across multiple species, known as micro-synteny, is the most effective method for unambiguously unraveling the origin and evolution of many understudied multigene venom protein families or short toxin genes [46, 75–77]. Another challenge is that many venom gene families are poorly studied and functionally characterized, with misleading naming conventions often implying phylogenetic relationships based on similar allergenic responses in bioactivity tests (e.g., venom allergens). Therefore, availability of a dedicated database based on phylogenetic relationships rather than naming conventions would be valuable for analyzing venom gene families. An example of a similar database is PhylomeDB, a catalog of gene phylogenies (phylomes) with multisequence alignments, phylogenetic trees, and ortholog predictions [78]. A promising specialized new resource is ToxCodAn-Genome, an automated pipeline for annotating toxin genes in genomes [79]. While it relies on prior knowledge of venom genes, and it has been tested on a set of well-known venomous lineages, it still overlooks rare venomous taxa and more species-specific gene families.

A branch of biology that is increasingly being explored for insights into venom production and phenotype changes is epigenetics [42, 80], the study of heritable traits occurring without DNA change (e.g., DNA methylation, histone modifications, chromatin architecture, noncoding RNA). Such changes are not erased by cell division, regulating gene expression, and altering cellular/physiological phenotypic traits influenced by environmental factors. A popular web-based genomic data exploration tool that provides visualization, integration, and analysis of epigenomic

datasets is the WashU Epigenome Browser [81]. This browser enables the interaction of 1D (genomic features), 2D (Hi-C data), 3D (chromatin structure), and 4D (gene/genomic regions as a function of time) data assessment, serving and expanding the data hubs from large consortia such as 4DN, Roadmap Epigenomics, TaRGET, and ENCODE. However, it currently does not include any venomous taxa.

Resources in transcriptomics

RNA sequencing is one of the most widely employed strategies used to characterize venom components by sequencing mRNA from dissected venom glands. This technique enables the acquisition of complete precursor sequences, which can then be used to build a custom database for mass spectrometry (MS)-based searches of crude venom (proteome-transcriptomics). Although genomes from venomous organisms are now becoming available, *de novo* transcriptome assembly (often coupled with subsequent proteome analysis) remains the most common method to describe venom compositions and for identifying novel toxin isoforms.

Due to the high computational demands of this process, most transcriptomics analyses are conducted on workstation computers, high-performance clusters, or via cloud computing and therefore use command-line tools. The most widely used assembler for venom gland transcriptomes is undoubtedly Trinity [47] and its companion Trinotate pipeline [82], which predicts coding regions and searches for homology against multiple databases. However, as of March 2024, Trinotate is no longer under active development or support. There are also more bioinformatics knowledge-demanding multiassembly pipelines that combine different assemblers and cover a larger space of gene models and reconstructed transcripts; 1 example is the Oyster River pipeline [83].

Functional annotation is typically performed manually through BLAST searches of translated amino acid sequences against UniProtKB, NCBI RefSeq, and other relevant databases (Supplementary Information Table S1), along with domain searches using tools like HMMER [84] or InterProScan [85] against Pfam [86], CDD [87], or own custom databases (e.g., [88]). To facilitate the identification of toxins, several predictor tools have been developed specifically for venom components. Some of these pipelines can be run locally from the command line (e.g., Venomix [89], ToxClassifier [90], TOXIFY [91], and DeTox [92]), while others, such as ToxDL [93] and ConoPrec on ConoServer [19], among others (Supplementary Information Table S1), can be run online through web interfaces where the translated amino acid sequences can be directly uploaded. Additionally, transcripts can be functionally annotated with GO terms using online deep learning approaches such as Pannzer2 [94] and filtering for transcripts annotated with terms like “toxin activity” or “modulation of process of another organism.”

While most transcriptomics studies on venomous animals focus on venom glands, comparative transcriptomics, which compares gene expression between venom glands and other tissues, provides further valuable insights. For instance, this approach can help with the annotation of a transcript as a venom protein, since toxin genes are generally uniquely or predominantly expressed in venom glands. Additionally, it helps identify pathways and genes involved in venom component biosynthesis and secretion [95–97]. After transcript quantification using command-line tools like Kallisto [98], differential expression analysis can be performed in R using various packages (e.g., edgeR [99]). The resulting list of venom gland upregulated genes can be subjected to enrichment

analysis for GO terms and KEGG pathways, revealing chaperones and other proteins important for protein folding and maturation or those secreted with toxins to facilitate their targeting.

RNA sequencing data, both raw and processed, can be archived in NCBI. Raw reads are deposited directly in the SRA archive or through the ENA either interactively or through the command line, while assemblies can be archived in the Transcriptome Shotgun Assembly (TSA) sequence database, although it does not accept sequences below 200 bp. Unlike the compulsory raw data submission, assemblies are not mandatory in most journals and are therefore often not uploaded or published as supplementary data (e.g., [100–104]). Gene expression quantifications can be uploaded on the NCBI Gene Expression Omnibus (GEO) archive.

Archiving sequencing and gene expression data is crucial and highly recommended for ensuring their accessibility and reproducibility. By making the assemblies and the expression levels of the corresponding transcripts freely available, researchers can prevent the duplication of effort and unnecessary reassembly and mapping of raw reads, allowing others to readily access and use this essential information for their own studies.

Resources in proteomics and peptidomics

Proteomics analysis plays a crucial role in venom research, as animal venoms are mostly composed of peptides and proteins. MS methods are commonly employed to identify venom components using 2 main approaches: bottom-up and top-down proteomics [14]. In bottom-up proteomics, venom components are enzymatically digested, and the resulting peptides are individually analyzed by tandem MS. Conversely, top-down approaches analyze intact venom proteins without any prior fragmentation, necessitating high-resolution MS instruments. In both approaches, peptides and proteins are identified through database-based or *de novo* searches. A database-based search matches spectra against an existing database, often derived from venom gland *de novo* transcriptome assembly or from other aforementioned datasets, while a *de novo* search infers peptide sequences directly from the mass spectra without relying on prior genomics or transcriptomics data [105]. Advancements in bottom-up proteomics have led to the development of user-friendly tools, democratizing complex data analysis. Similar to genomics and transcriptomics, proteomics analyses on the raw data are mostly performed locally or on a computer cluster, while online resources are applied for downstream analyses.

For bottom-up proteomics, prominent proprietary database search engines like Mascot [106] and PEAKS DB [107] are commonly used for venom protein identification. Additionally, software tools like ProteomeDiscoverer [108] integrate multiple search algorithms such as Sequest [109], Mascot, and Byonic [110] for peptide identification and protein characterization. Freely available platforms, including pFind 3 [111], MSFragger [112], and PeptideShaker [113], offer powerful tools for identifying venom components and characterizing posttranslational modifications (PTMs). Other software solutions like MaxQuant [114] and Skyline [115] enable identification and quantification of venom proteins using data-dependent acquisition (DDA) methods. To overcome the limitation of DDA, platforms such as DIA-NN [116] and Max-DIA [117] use data-independent acquisition (DIA) methods [118]. In contrast to database-based searches, *de novo* sequencing software like Novor [119] and pNovo [120] facilitate fast and accurate peptide sequencing, although it can be challenging for complex spectra and peptides with extensive PTMs.

Top-down approaches aim to characterize entire toxins, including their isoforms and PTMs, and have recently been applied to venom research [121]. In database-based searches, software such as OpenMS [122], MZmine [123], MS-Deconv [124], and Msconvert [125] are commonly used for deconvoluting complex data. Additionally, MS-Align+ [126], MASH Suite [127], pTop [128], and TopMG [129] allow for high-throughput and automated protein sequence matching of multiple isoforms with high confidence. For *de novo* searches, license-based software like PEAKS (Bioinformatics Solutions Inc.) and ProSight PC (Thermo Fisher Scientific) are generally used, as well as free academic licenses for TopPIC [130] and Informed-Proteomics [131].

AI tools are emerging in proteomics to predict protein structures, pharmacological properties, and interaction partners. Toxin-specific web server tools include ToxinPred [132], ToxinPred2 [133], and ToxClassifier [90] (although unavailable as of April 2024), while non-toxin-specific platforms include Peptide Ranker [134] and PEP-FOLD3 [135], which use machine learning algorithms to predict and design peptides from amino acid sequences. The newest version of PEP-FOLD4 [136] accounts for pH conditions and salt concentration conformations, which are critical parameters for accurate structure prediction. Well-known servers based on machine learning approaches include AlphaFold2 [28], available in ColabFold [137], RoseTTAFold [138], and RaptorX [139], which are based on PDB structures, multiple sequence alignments, and specific algorithms to learn the backbone conformations and side chain–side chain contacts. However, limitations exist, particularly with the accuracy of predictions when signal peptides, pro-peptides, or PTM positions are not specified in the input amino acid sequence. Despite challenges, AI tools offer promising capabilities in predicting unknown protein structures.

Raw proteomics data can be deposited in repositories like PRIDE [17] and Mass Spectrometry Interactive Virtual Environment (MassIVE) [140], which play a crucial role in facilitating collaboration and reproducibility. Additionally, MassIVE offers tools for reanalyzing spectral datasets, comparing results, and more.

Resources in metabolomics

The main objective of metabolomics is to identify and quantify the metabolites that exist in biological fluids, cells, and tissues. Amines, organic acids, steroids, alkaloids, and sugars are considered the substances of the metabolome. To date, the elucidation of metabolite structures is mainly performed by studying the literature and comparing the MS/MS spectra of related metabolites. Comprehensive databases include the Human Metabolome Database (HMDB) [141] and KEGG [29], which offer different qualitative and quantitative data for human metabolites and information about metabolomic pathways. HMDB is currently the database containing the largest data collection of MS/MS fragmentation spectra of metabolites [141, 142]. An interesting tool is offered by the Global Natural Products Social Molecular Networking (GNPS) [143], a web-based mass spectrometry ecosystem that aims to be an open-source and open-access knowledge base for community-wide organization and sharing of raw, processed, or identified tandem mass (MS/MS) spectrometry data. GNPS aids in identification and discovery throughout the entire life cycle of data, from initial data acquisition to postpublication.

The only existing venom-specialist metabolite database is Venoms [38], which focuses on low molecular mass metabolites from spider venoms. Venoms gathers known structures of spider venom metabolites and offers a fragment ion calculator (FRIOC) for the prediction of fragment ions for the linear

polyamine derivatives. This website can be considered complementary to *ArachnoServer*. Despite its usefulness, this resource is limited to spiders and is not included in the typical automated MS analyses.

A suggestion for a future endeavor could be to extend the content of *VenoMS* to other venomous organisms and create a more comprehensive online database of venom metabolites. As venom metabolomics is still in its infancy, challenges rely mostly in the chemical identification of metabolites and the integration with data from other omics platforms.

Resources in translational research

The vast biotechnological and biomedical potential of animal venoms and toxins is undeniable, with well-documented bioactivities ranging from analgesic, immunomodulatory, anticancer, antimicrobial, and antiparasitic properties [2–5, 144]. This potential translates into a growing number of venom-derived drugs, with already 11 approved by the US Food and Drug Administration and European Medicines Agency, and many more in preclinical or clinical development. Beyond medicine, venom toxins hold promise for diagnostics, nanopore-based sensing, agrochemicals, and cosmetics [8, 10–13, 145]. However, despite the evident opportunities, the translation of basic research into concrete applications is a lengthy process that requires the generation of a variety of data and access to a wide array of different tools and databases. In this section, we provide an overview of the available resources pertinent to venom and toxin research from a biomedical and translational perspective.

In a typical workflow for venom component discovery, the first step involves candidate identification. This can be achieved by generating new data by means of genomics or proteo-transcriptomics analysis or by mining existing databases. Typical databases include *ArachnoServer*, *Toxin* and *Toxin Target Database (T3DB)*, *PubChem*, and *UniProtKB/Swiss-Prot*, among others ([Supplementary Information Table S1](#)). *T3DB* is particularly useful as it combines detailed toxin data with comprehensive receptor information, molecular and biological properties, toxin effects, and potential therapeutic applications [39]. For peptide-based cancer research, *CancerPPD4* [146], *canSAR* [147], *ApInAPDB* [148], *PaccMann* [149], and *EviCor* [150] provide platforms for the exploration of the mechanism of action, function, binding target, affinity, structural information, and other physicochemical features of peptides. Furthermore, they offer AI-based predictions of anticancer compound sensitivity and other properties to inform drug discovery. A comprehensive database useful in translational research was the discontinued *VenomKB* [151], which included data on venom's molecular components and their potential applications in drug discovery and development.

The databases can be mined manually to select a list of potential candidates, which can be further screened using the prediction tools mentioned earlier. Alternatively, databases can be used to build machine learning models based on random forest, support vector machine, or artificial neural network algorithms, which can process a vast amount of data and identify patterns to predict potential drug targets. This first crucial step of target identification poses a challenge in venom research as the toxin information is scattered across several databases. Thanks to the advent of the Semantic Web (SW), the tedious process to manually mine different life science databases can be significantly reduced [152]. SW provides a common framework that enables data to be shared and reused across different data sources. Combining and querying these data sources are possible by using a standard se-

mantic query language like SPARQL. A solution to meaningfully access the databases containing animal venom information is to federate them by applying SW technologies that enable semantic queries across them [153]. For instance, currently *UniProtKB* and *PubChem* can be jointly queried by writing a single federated SPARQL query [154].

Once potential candidates are characterized, further steps include prediction of molecular targets and interactions with the toxins. Databases such as the *mousephenotype.org* for mammals [155], *zfin.org* for zebrafish [156], and *flybase.org* for insects [157] can be explored for predicting the effects of toxin intervention on a systemic level and specific regulatory functions, as well as identifying promising pharmaceutical or bioinsecticides targets. Web-based prediction tools for molecular docking include *Swiss-Dock* [158], the more recently developed *PPI-Affinity model* [159], and the *CAMP model* [160] to elaborate on target predictions for peptides and proteins. Molecular docking and molecular dynamics simulation models such as quantitative structure–activity relationship (QSAR), quantitative structure–property relationship (QSPR) analysis, pharmacophore modeling, and *iBitter-SCM* are frequently used to decipher peptide and protein interactions [161].

Once a lead compound has been identified and selected, it can be modified to have unique and desirable properties, for instance, to modulate their target selectively and induce a therapeutic rather than a harmful toxic effect [162]. *ToxinPred* and *ToxinPred2* include tools to design all possible single mutant analogues of a peptide and predict whether they are toxic or not, as well as to optimize the peptide sequence to get maximum, minimum, and desired toxicity. In addition, *ToxinPred* offers users to calculate various physicochemical properties.

While the approaches delineated above facilitate the search among known venom compounds, enduring challenges remain for the prediction of toxins with undescribed new mechanisms of action and the identification of the potential off-target effects that might limit the usefulness of the molecule as a putative therapeutic drug [163], although current machine learning algorithms present a promising avenue for the discovery of molecules with novel activities. Despite the potential benefits, it is important to acknowledge that the principles of open science may not always be guaranteed in translational and applied research, often due to confidentiality agreements associated with preliminary studies on toxin activity prediction and application.

Resources in antivenom production and administration

Scientists working in the field of antivenom research are typically interested in a variety of information spanning from the geographic distribution of the venomous species to their venom composition and variation, toxin structure, and bioactivity, which all impact antivenom efficiency. Most of the resources related to this kind of information have been already discussed in previous sections and are listed in [Supplementary Information Table S1](#); therefore, here we focus on the resources available for antivenom producers.

A first important resource is represented by the World Health Organization (WHO) guidelines, which provides comprehensive and important manuals for antivenom manufacturers on the design, production, control, and regulation of high-quality antivenom immunoglobulins. These guidelines are regularly updated to provide framework guidance to national regulatory bodies for securing the products they offer. Technical bulletins, reports, and documents are also available on the WHO website.

Within the scope of WHO web resources, in addition to pharmacopoeia requirements, current antidote production, especially the improvement of studies and technologies carried out under GMP quality system conditions, is ensured. Additionally, WHO manages the snakebite information and data platform as part of the 2019–2030 global strategy for the prevention and control of snakebite envenoming, which is within the scope of neglected tropical diseases by WHO. This web source platform is part of a collaboration between the departments for the control of Neglected Tropical Diseases (WHO/NTD) and the Dissemination of Data for Impact and analytics (WHO/DDI). Another data source created for easy access to antivenom in cases of envenoming caused by poisonous animals is the Munich AntiVenom INdex (MAVIN) created by the Munich Poison Center. MAVIN gathers a list of venomous animals, antivenom holding centers, antivenoms, and correlated information.

In addition to international web resources such as WHO and MAVIN, some countries have developed national web resources to help staff at zoos and aquariums managing the supply of antivenom and finding the right antivenom when they need it. For instance, an online Antivenom Index was created in 2006 by the Association of Zoos and Aquariums (AZA) and America's Poison Centers (previously known as American Association of Poison Control Centers). The University of Arizona College of Pharmacy is currently responsible for maintaining, updating, and hosting this index. However, only representatives of poison control centers and AZA-accredited institutions have access to the Antivenom Index.

Resources in clinical toxicology

Several freely available resources offer information on venoms and venomous animals, which are relevant to clinical toxicologists and toxinologists. A central resource is the “Clinical Toxicology Resources” website, which provides comprehensive information on venomous and poisonous animals, plants, and mushrooms from around the world ([Supplementary Information Table S1](#)). This repository receives support from experts around the world, and it features a searchable database that allows users to find specific organisms by common or scientific names, family, country, or region. Another useful resource is PubChem, which gathers information on chemical structure, chemical and physical properties, biological activity, toxicity, and medical management guidance, among others.

Most clinical toxicology and toxicology databases cater specifically to poison centers and are accessible only to registered health care professionals. Nonetheless, some are reachable upon subscription fees and may offer free or reduced-cost access, particularly for users in low-income countries. For instance, AfriTox offers online and offline versions, primarily for registered health care professionals, with subscription-based access. This database focuses on substances, including venomous exposures, from an African perspective. The Merative Micromedex® POISINDEX® System is widely used worldwide, especially in North America, and provides both summary and in-depth clinical toxicology information, including details on venomous animals, through subscription-based access. Another useful resource is TOXBASE, produced by poison specialists and medical toxicologists, which offers advice on toxin features and exposure management to toxins and venomous animals. While primarily accessible to UK health care professionals, TOXBASE is also used internationally, with special arrangements for certain countries. Lastly, TOXINZ provides information and treatment guidelines, including venomous animal exposures. While primarily designed for use in New

Zealand, TOXINZ is accessible in other countries through paid subscriptions.

Challenges, Needs, and Perspectives of Web Resources in Venom Research

The survey that we conducted within the framework of the EU-VEN COST Action [16], although representing only a sample of the worldwide venom research community, provided important insights into the challenges and needs of researchers and clinicians working with animal venoms or toxins. Here, we have summarized and discussed them.

Challenges

Many scientists in the venom research community expressed disappointment due to the bottleneck caused by the limited expertise in bioinformatics and data management, especially concerning the handling of complex “-omics” pipelines essential for cutting-edge research. Despite the improvements in accessibility offered by databases, there is still a demand for more user-friendly interfaces that seamlessly integrate data and tools into existing pipelines, facilitating the translation of research findings into clinical applications. However, achieving a unified graphical user interface (GUI) software is not easy due to the variety, volume, and complexity of current data, requesting storage on servers alongside the necessary analysis tools. Toxinologists are encouraged to collaborate with bioinformaticians and relevant technology experts in cross-disciplinary projects. Initiatives like EUVEN and organizations such as the Swiss Institute of Bioinformatics provide support and facilitate collaborations by offering access to databases of researchers and their corresponding expertise.

Another challenge faced by venom researchers, particularly those involved in applied aspects like drug discovery, was related to the scattered and diverse nature of information about venoms and toxins across several databases. This issue is not unique to venom researchers but is prevalent among biologists. As the production of biological and health data continues to exponentially grow, so does the number of databases [164]. However, querying is still largely limited to a single database at a time, making it difficult to integrate multiple data types to answer complex biological questions [152]. A step forward in addressing this challenge is the adoption of query languages like SPARQL to search across different databases and perform data manipulation tasks such as exploration, extraction, and annotation. Furthermore, to effectively manage and analyze datasets, standardized terminologies and classification systems are essential. Ontologies and glossaries serve as structured vocabularies that provide a common language for annotating and organizing biological information (e.g., GO, UniProtKB/Swiss-Prot controlled vocabularies). Even though the use of such resources is generally well consolidated along the research pipelines, often different terms are employed to denote the same concept, or conversely, the same term is used to represent multiple concepts across web resources, thereby hindering interoperability [165]. For instance, in Ontobee [166], a catalog and web-based linked data server for semantic terminologies, the term “venom” is described differently in 8 ontologies. This highlights the need for mapping terms between the semantic resources commonly used in the field.

To access the wealth of data, a reliable database needs to be regularly maintained and updated. Its longevity depends on several factors, including the underlying technology and system, the frequency of data updates, its ability to handle growing data

volumes, their regular backup, and the database capacity to continue to meet user needs. Ultimately, the decision to maintain a database largely depends on the funding required to support the work of developers and curators, which in turn depends on the size of the database and the number of users.

In terms of data analysis, as venom omics data accumulate, the challenge evolves from basic descriptive comparative findings to the more sophisticated task of integrating multiomics data. This approach ultimately aims to gain a comprehensive understanding of the complexity of biological systems and their underlying mechanisms. To this end, data standardization, advanced computational methods (e.g., machine learning techniques), and interpretation of diverse data types are key to provide meaningful insights. While multiomics integration tools are currently applied in studying complex human diseases [167], they hold great promise for deciphering equally complex venom phenotypes.

Needs

Despite the abundance of databases containing information on animal toxins, some data remain disorganized and inaccessible due to a lack of structured datasets. For the data to be accessible through query languages, databases need to be machine-readable, meaning they must be formatted in a way that can be processed by software tools. This is also crucial for full implementation of the Findable Accessible Interoperable Reusable (FAIR) principles [168]. The Resource Description Framework (RDF), for instance, is a SW standard data model adopted by many databases for sharing and linking data. Data in RDF can be queried, retrieved, and manipulated using the SPARQL language, which has the advantage that it is graph-based, thus allowing users to join data from multiple, diverse sources (in contrast to SQL, which is a table-based query language). Therefore, there is a need to standardize the structure of databases to run queries on animal venoms and toxin research across them. Furthermore, it is advisable to use existing ontologies and incorporate controlled terms already in use or map redundant terms among them. This can be facilitated by searching existing terms in semantic resource catalogs such as Ontobee, the Ontology Lookup Service (OLS), or BioPortal [152, 169]. This practice prevents unnecessary duplications, reduces redundancy, and enhances data reusability and interoperability, which is particularly relevant to a high multidisciplinary field like venom research.

Another issue raised by the venom research community is the absence of a repository for protocols and methods for recombinantly producing or chemically synthesizing venom peptides, which would benefit researchers by preventing redundant protocol optimization efforts, especially in the case of toxins difficult to refold. Additionally, there is a need for a centralized, nonprofit database of biological materials related to venoms and natural or engineered toxins stored or generated in research institutes, similar to plasmid repositories or even catalogs for museum specimens, to aid researchers in accessing preexisting materials for their own studies.

Ensuring data and information accessibility and standardization to the research and clinician communities and the public remains crucial, as discussed in previous sections. The importance of making these data publicly available is further emphasized by the FAIR principles [159] and the recent European Open Access policies [170], which advocate for open access not only to publications but also to all underlying data. Addressing these needs and challenges will require collaboration and concerted efforts from

researchers, clinicians, and organizations to advance venom research and its applications.

Perspectives on a unified venom web resource

Steps toward satisfying the needs of the venom research community include the creation of a venom-specific resource containing detailed information on venomous species and their venoms and toxins. This database could encompass genome architecture and function of venomous species, venom gland transcriptomes, toxin genes and their translated amino acid sequences, PTMs, 3D structures, pharmacological activities and toxicity levels, molecular and cellular targets, and mechanisms of action, coupled with ecological and evolutionary information of the corresponding species (e.g., diet and geographical distribution). By consolidating such diverse information into a single resource or interface uniting a range of resources, scientists working in the interdisciplinary field of animal venoms and toxins would have a valuable tool at their disposal. It would enable them to access both general and specific information on a vast number of venomous species and toxins and would decrease the time spent on extensive literature searches.

Such a resource could also significantly contribute to venom research by facilitating the classification of venom proteins, aiding in the design of peptides with desired pharmacological properties, and identifying potential interactions. However, the creation and maintenance of such a platform would present considerable challenges, requiring a substantial workforce, financial resources, and international interdisciplinary collaborations to ensure its continual updates and accuracy.

An existing resource like VenomZone could serve as a starting point toward realizing this unified resource. However, significant expansions would be necessary to incorporate the additional data proposed. A promising initiative is the interactive table that we have compiled within the framework of this work and made available on the VenomZone website [15] (Supplementary Information Fig. S4). It includes current web resources relevant to venom research in an interactive way. It therefore represents a positive step toward creating a comprehensive and accessible resource for the entire venom research community.

Conclusions

- Modern venom research is a multidisciplinary field resulting in the generation and analysis of highly diverse datasets.
- Currently, information on venom and toxin data is scattered across different resources, ranging from generalist to specialized platforms.
- Most multiomics analyses are performed using software and command-line tools that require advanced computational and command-line skills, while most available web resources mainly offer downstream analyses.
- One of the core challenges is accessing and providing information across the different databases. There is an urgent need to establish standards to facilitate interoperability and allow seamless querying of animal venom and toxin research across platforms.
- Progress toward meeting the needs of the venom research community requires the establishment of a dedicated venom-specific resource. VenomZone, together with our newly curated site on demanded tools and resources, represents an important first step towards this goal.

Additional Files

Additional file 1.pdf: Survey on web resources in venom research. Questions included in the survey sent to the members of the EUVEN COST Action and the participants of the First International EUVEN Congress in 2021.

Additional file 2.csv: Answers to the survey. Anonymized answers to the survey.

Additional file 3.tsv: Summary of venom research areas. Contingency table of the research areas represented by the respondents of the survey used to create [Supplementary Information Fig. S1](#).

Additional file 4.tsv: Summary of organisms studied in venom research. Contingency table of the organisms studied by the respondents of the survey used to create [Supplementary Information Fig. S2](#).

Additional file 5.pdf: Supplementary Information. Overview of the results from the survey and the compiled list of web resources, databases, and online tools utilised in venom research ([Supplementary Information Table S1](#)) and available on the VenomZone website.

Abbreviations

3D: 3-dimensional; AI: artificial intelligence; API: application program interface; AZA: Association of Zoos and Aquariums; BLAST: Basic Local Alignment Search Tool; BMRB: Biological Magnetic Resonance Data Bank; cryo-EM: cryo-electron microscopy; DDA: data-dependent acquisition; DDBJ: DNA Data Bank of Japan; DIA: data-independent acquisition; EMDB: Electron Microscopy Data Bank; ENA: European Nucleotide Archive; EUVEN: European Venom Network; FRIOC: fragment ion calculator; GEO: Gene Expression Omnibus; GNPS: Global Natural Products Social Molecular Networking; GO: Gene Ontology; GUI: graphical user interface; HMDB: Human Metabolome Database; KEGG: Kyoto Encyclopedia of Genes and Genomes; MASSIVE: Mass Spectrometry Interactive Virtual Environment; MAVIN: Munich AntiVenom INDEX; MS: mass spectrometry; NCBI: National Center for Biotechnology Information; NIH: National Institutes of Health; NMR: nuclear magnetic resonance; PDB: Protein Data Bank; PRIDE: Proteomics IDENTification; PTM: posttranslational modification; QSAR: quantitative structure–activity relationship; QSPR: quantitative structure–property relationship; RDF: Resource Description Framework; SW: Semantic Web; T3DB: Toxin and Toxin Target Database; TSA: Transcriptome Shotgun Assembly; WHO: World Health Organization.

Acknowledgments

The authors thank Ronald A. Jenner for his valuable comments on an earlier version of the manuscript, as well as Marc Robinson-Rechavi, Sébastien Moretti, and Valentine Rech De Laval for their feedback on additional useful web resources.

Author Contributions

Major conceptualization by M.V.M., G.A., G.Z., A.A., and B.M.v.R. G.Z. and F.J. analyzed the survey data. M.L.C., F.J., P.M., and B.M.v.R. conceptualized the graphics, and B.M.v.R. created the figures. G.Z. led the writing of the manuscript. All the authors contributed to the main text. All the authors have read and agreed to the published version of the manuscript.

Funding

This work is funded by the European Cooperation in Science and Technology (COST, www.cost.eu) to M.V.M. and based on work from the COST Action CA19144 European Venom Network (EUVEN, <https://euven-network.eu/>). This review is an outcome of EUVEN Working Group 4 (“Web resources”) led by A.A. and G.Z. G.Z. was supported by the European Union’s Horizon 2020 Research and Innovation program through a Marie Skłodowska-Curie Individual Fellowship (grant agreement No. 845674). B.M.v.R. acknowledges funding from the German Science Foundation (DFG RE3454/6–1). M.P.I. was supported by the TALENTO Program by the Regional Madrid Government (#2022-5A/BIO-24228) and the grant (#PID2021-126691OB-I00) funded by MICIU/AEI/10.13039/50110001100011033 and by the European Union. R.V. acknowledges the Portuguese Foundation for Science and Technology (FCT), QREN, FEDER, and COMPETE for funding to the Institute of Biomedicine (iBiMED) (UIDB/04501/2020, PO-CI-01-0145-FEDER-007628). AA was partially supported by FCT, ERDF, ESIF, and COMPETE Strategic Funding UIDB/04423/2020 and UIDP/04423/2020.

Data Availability

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

References

- Schendel V, Rash LD, Jenner RA, et al. The diversity of venom: the importance of behavior and venom system morphology in understanding its ecology and evolution. *Toxins* 2019;11:666. <https://doi.org/10.3390/toxins11110666>.
- Lewis RJ, Garcia ML. Therapeutic potential of venom peptides. *Nat Rev Drug Discov* 2003;2:790–802. <https://doi.org/10.1038/nrd1197>.
- Holford M, Daly M, King GF, et al. Venoms to the rescue. *Science* 2018;361:842–44. <https://doi.org/10.1126/science.aau7761>.
- Herzig V, Cristofori-Armstrong B, Israel MR, et al. Animal toxins—nature’s evolutionary-refined toolkit for basic research and drug discovery. *Biochem Pharmacol* 2020;181:114096. <https://doi.org/10.1016/j.bcp.2020.114096>.
- Waheed H, Moin SF, Choudhary MI. Snake venom: from deadly toxins to life-saving therapeutics. *Curr Med Chem* 2017;24:1874–91. <https://doi.org/10.2174/0929867324666170605091546>.
- Talukdar A, Maddhesiya P, Namsa ND, et al. Snake venom toxins targeting the central nervous system. *Toxin Rev* 2023;42:382–406. <https://doi.org/10.1080/15569543.2022.2084418>.
- Oliveira AL, Viegas MF, da Silva SL, et al. The chemistry of snake venom and its medicinal potential. *Nat Rev Chem* 2022;6:451–69. <https://doi.org/10.1038/s41570-022-00393-7>.
- Marsh NA. Diagnostic uses of snake venom. *Pathophysiol Haemos Thromb* 2001;31:211–17. <https://doi.org/10.1159/000048065>.
- Estevão-Costa M-I, Sanz-Soler R, Johanningmeier B, et al. Snake venom components in medicine: from the symbolic rod of Asclepius to tangible medical research and application. *Int J*

- Biochem Cell Biol 2018;104:94–113. <https://doi.org/10.1016/j.bio.2018.09.011>.
10. Windley MJ, Herzig V, Dziemborowicz SA, et al. Spider-venom peptides as bioinsecticides. *Toxins* 2012;4:191–227. <https://doi.org/10.3390/toxins4030191>.
 11. King GF, Hardy MC. Spider-venom peptides: structure, pharmacology, and potential for control of insect pests. *Annu Rev Entomol* 2013;58:475–96. <https://doi.org/10.1146/annurev-ento-120811-153650>.
 12. Modahl CM, Brahma RK, Koh CY, et al. Omics technologies for profiling toxin diversity and evolution in snake venom: impacts on the discovery of therapeutic and diagnostic agents. *Annu Rev Anim Biosci* 2020;8:91–116. <https://doi.org/10.1146/annurev-animal-021419-083626>.
 13. Dutertre S, Lewis RJ. Use of venom peptides to probe ion channel structure and function. *J Biol Chem* 2010;285:13315–20. <https://doi.org/10.1074/jbc.R109.076596>.
 14. von Reumont BM, Anderlüh G, Antunes A, et al. Modern venomomics—current insights, novel methods, and future perspectives in biological and applied animal venom research. *Gigascience* 2022;11:giac048. <https://doi.org/10.1093/gigascience/e/giac048>.
 15. VenomZone Web Resources. <https://venomzone.expasy.org/10897>. Accessed 9 July 2024.
 16. Modica MV, Ahmad R, Ainsworth S, et al. The new COST Action European Venom Network (EUVEN)—synergy and future perspectives of modern venomomics. *Gigascience* 2021;10:giab019. <https://doi.org/10.1093/gigascience/giab019>.
 17. Perez-Riverol Y, Bai J, Bandla C, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;50:D543–52. <https://doi.org/10.1093/nar/gkab1038>.
 18. Pineda SS, Chaumeil P-A, Kunert A, et al. ArachnoServer 3.0: an online resource for automated discovery, analysis and annotation of spider toxins. *Bioinformatics* 2018;34:1074–76. <https://doi.org/10.1093/bioinformatics/btx661>.
 19. Kaas Q, Yu R, Jin A-H, et al. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 2012;40:D325–30. <https://doi.org/10.1093/nar/gkr886>.
 20. Jungo F, Bougueleret L, Xenarios I, et al. The UniProtKB/Swiss-Prot Tox-Prog program: a central hub of integrated venom protein data. *Toxicon* 2012;60:551–57. <https://doi.org/10.1016/j.toxicon.2012.03.010>.
 21. Duvaud S, Gabella C, Lisacek F, et al. ExPasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res* 2021;49:W216–27. <https://doi.org/10.1093/nar/gkab225>.
 22. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;47:D520–28. <https://doi.org/10.1093/nar/gky949>.
 23. Romero PR, Kobayashi N, Wedell JR, et al. BioMagResBank (BMRB) as a resource for structural biology. In: Gáspári Z, ed. *Structural Bioinformatics: Methods and Protocols*. New York: Springer US; 2020:187–218. https://doi.org/10.1007/978-1-0716-0270-6_14.
 24. The wwPDB Consortium. EMDb—The Electron Microscopy Data Bank. *Nucleic Acids Res* 2024;52:D456–65. <https://doi.org/10.1093/nar/gkad1019>.
 25. Haji-Ghassemi O, Chen YS, Woll K, et al. Cryo-EM analysis of scorpion toxin binding to ryanodine receptors reveals subconductance that is abolished by PKA phosphorylation. *Sci Adv* 2023;9:eadf4936. <https://doi.org/10.1126/sciadv.adf4936>.
 26. Nys M, Zarkadas E, Brams M, et al. The molecular mechanism of snake short-chain α -neurotoxin binding to muscle-type nicotinic acetylcholine receptors. *Nat Commun* 2022;13:4543. <https://doi.org/10.1038/s41467-022-32174-7>.
 27. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;50:D439–44. <https://doi.org/10.1093/nar/gkab1061>.
 28. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–89. <https://doi.org/10.1038/s41586-021-03819-2>.
 29. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;28:27–30. <https://doi.org/10.1093/nar/28.1.27>.
 30. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;25:25–29. <https://doi.org/10.1038/75556>.
 31. Kim S, Chen J, Cheng T, et al. PubChem 2023 update. *Nucleic Acids Res* 2023;51:D1373–80. <https://doi.org/10.1093/nar/gkac956>.
 32. Krylov NA, Tabakmakher VM, Yureva DA, et al. Kalium 3.0 is a comprehensive depository of natural, artificial, and labeled polypeptides acting on potassium channels. *Protein Sci* 2023;32:e4776. <https://doi.org/10.1002/pro.4776>.
 33. Postic G, Gracy J, Périn C, et al. KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Res* 2018;46:D454–58. <https://doi.org/10.1093/nar/gkx1084>.
 34. Liu J, Maxwell M, Cuddihy T, et al. SrepeYard: an online resource for disulfide-stabilized tandem repeat peptides. *Protein Sci* 2023;32:e4566. <https://doi.org/10.1002/pro.4566>.
 35. Jungo F, Estreicher A, Bairoch A, et al. Animal toxins: how is complexity represented in databases? *Toxins* 2010;2:262–82. <https://doi.org/10.3390/toxins2020261>.
 36. Tan PTJ, Veeramani A, Srinivasan KN, et al. SCORPION2: a database for structure-function analysis of scorpion toxins. *Toxicon* 2006;47:356–63. <https://doi.org/10.1016/j.toxicon.2005.12.001>.
 37. He Q-Y, He Q-Z, Deng X-C, et al. ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res* 2008;36:D293–97. <https://doi.org/10.1093/nar/gkm832>.
 38. Forster YM, Reusser S, Forster F, et al. Venoms—a website for the low molecular mass compounds in spider venoms. *Metabolites* 2020;10:327. <https://doi.org/10.3390/metabo10080327>.
 39. Wishart D, Arndt D, Pon A, et al. T3DB: the toxic exposome database. *Nucleic Acids Res* 2015;43:D928–34. <https://doi.org/10.1093/nar/gku1004>.
 40. VenomZone. <https://venomzone.expasy.org/>. Accessed 9 July 2024.
 41. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 2022;50:W345–51. <https://doi.org/10.1093/nar/gkac247>.
 42. Perry BW, Gopalan SS, Pasquesi GIM, et al. Snake venom gene expression is coordinated by novel regulatory architecture and the integration of multiple co-opted vertebrate pathways. *Genome Res* 2022;32:1058–73. <https://doi.org/10.1101/gr.27625.1.121>.
 43. Dowell NL, Giorgianni MW, Kassner VA, et al. The deep origin and recent loss of venom toxin genes in rattlesnakes. *Curr Biol* 2016;26:2434–45. <https://doi.org/10.1016/j.cub.2016.07.038>.

44. Vonk FJ, Casewell NR, Henkel CV, et al. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc Natl Acad Sci USA* 2013;110:20651–56. <https://doi.org/10.1073/pnas.1314702110>.
45. Schield DR, Card DC, Hales NR, et al. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* 2019;29:590–601. <https://doi.org/10.1101/gr.240952.118>.
46. Drukewitz SH, von Reumont BM. The significance of comparative genomics in modern evolutionary venomomics. *Front Ecol Evol* 2019;7:163. <https://www.frontiersin.org/articles/10.3389/feco.2019.00163>.
47. Grabherr MG, Haas BJ, Yassour M, et al. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat Biotechnol* 2011;29:644–52. <https://doi.org/10.1038/nbt.1883>.
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–59. <https://doi.org/10.1038/nmeth.1923>.
49. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37:907–15. <https://doi.org/10.1038/s41587-019-0201-4>.
50. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
51. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptsomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;14:R36. <https://doi.org/10.1186/gb-2013-14-4-r36>.
52. Musich R, Cadle-Davidson L, Osier MV. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Front Plant Sci* 2021;12:657240. <https://doi.org/10.3389/fpls.2021.657240>.
53. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;21:30. <https://doi.org/10.1186/s13059-020-1935-5>.
54. Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;39:1348–65. <https://doi.org/10.1038/s41587-021-01108-x>.
55. Larivière D, Abueg L, Brajuka N, et al. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat Biotechnol* 2024;42:367–70. <https://doi.org/10.1038/s41587-023-02100-3>.
56. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res* 2022;50:D988–95. <https://doi.org/10.1093/nar/gkab1049>.
57. Rhie A, McCarthy SA, Fedrigo O, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021;592:737–46. <https://doi.org/10.1038/s41586-021-03451-0>.
58. Koepfli K-P, Paten B, Genome 10 K Community of Scientists, et al. The Genome 10 K Project: a way forward. *Annu Rev Anim Biosci* 2015;3:57–111. <https://doi.org/10.1146/annurev-animal-090414-014900>.
59. Woolstra CR, Woerheide G, Lopez JV, et al. Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). *Invert Systematics* 2017;31:1–7. <https://doi.org/10.1071/IS16059>.
60. Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci USA* 2018;115:4325–33. <https://doi.org/10.1073/pnas.1720115115>.
61. Formenti G, Theissinger K, Fernandes C, et al. The era of reference genomes in conservation genomics. *Trends Ecol Evol* 2022;37:197–202. <https://doi.org/10.1016/j.tree.2021.11.008>.
62. The Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc Natl Acad Sci USA* 2022;119:e2115642118. <https://doi.org/10.1073/pnas.2115642118>.
63. Zhang G, Li C, Li Q, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 2014;346:1311–20. <https://doi.org/10.1126/science.1251385>.
64. Zimmermann B, Montenegro JD, Robb SMC, et al. Topological structures and syntenic conservation in sea anemone genomes. *Nat Commun* 2023;14:8270. <https://doi.org/10.1038/s41467-023-44080-7>.
65. Kon-Nanjo K, Kon T, Horkan HR, et al. Chromosome-level genome assembly of hydractinia symbiolongicarpus. *G3 (Bethesda)* 2023;13:jkad107. <https://doi.org/10.1093/g3journal/jkad107>.
66. Chapman JA, Kirkness EF, Simakov O, et al. The dynamic genome of Hydra. *Nature* 2010;464:592–96. <https://doi.org/10.1038/nature08830>.
67. Putnam NH, Srivastava M, Hellsten U, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 2007;317:86–94. <https://doi.org/10.1126/science.1139158>.
68. Shinzato C, Shoguchi E, Kawashima T, et al. Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* 2011;476:320–23. <https://doi.org/10.1038/nature10249>.
69. Baumgarten S, Simakov O, Esherick LY, et al. The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proc Natl Acad Sci USA* 2015;112:11893–98. <https://doi.org/10.1073/pnas.1513318112>.
70. Bhattacharya D, Agrawal S, Aranda M, et al. Comparative genomics explains the evolutionary success of reef-forming corals. *eLife* 2016;5:e13288. <https://doi.org/10.7554/eLife.13288>.
71. Grimson A, Srivastava M, Fahey B, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 2008;455:1193–97. <https://doi.org/10.1038/nature07415>.
72. Moran Y, Fredman D, Praher D, et al. Cnidarian microRNAs frequently regulate targets by cleavage. *Genome Res* 2014;24:651–63. <https://doi.org/10.1101/gr.162503.113>.
73. Schwaiger M, Schönauer A, Rendeiro AF, et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res* 2014;24:639–50. <https://doi.org/10.1101/gr.162529.113>.
74. Cazet JF, Siebert S, Little HM, et al. A chromosome-scale epigenetic map of the Hydra genome reveals conserved regulators of cell state. *Genome Res* 2023;33:283–98. <https://doi.org/10.1101/gr.277040.122>.
75. Jackson TNW, Koludarov I. How the toxin got its toxicity. *Front Pharmacol* 2020;11:574925. <https://doi.org/10.3389/fphar.2020.574925>.
76. Koludarov I, Velasque M, Senoner T, et al. Prevalent bee venom genes evolved before the aculeate stinger and eusociality. *BMC Biol* 2023;21:229. <https://doi.org/10.1186/s12915-023-01656-5>.
77. Koludarov I, Senoner T, Jackson TNW, et al. Domain loss enabled evolution of novel functions in the snake three-finger toxin gene superfamily. *Nat Commun* 2023;14:4861. <https://doi.org/10.1038/s41467-023-40550-0>.
78. Fuentes D, Molina M, Chorostecki U, et al. PhylomeDB V5: an expanding repository for genome-wide catalogues of

- annotated gene phylogenies. *Nucleic Acids Res* 2022;50:D1062–68. <https://doi.org/10.1093/nar/gkab966>.
79. Nachtigall PG, Durham AM, Rokyta DR, et al. ToxCodAn-genome: an automated pipeline for toxin-gene annotation in genome assembly of venomous lineages. *Gigascience* 2024;13:giad116. <https://doi.org/10.1093/gigascience/giad116>.
 80. Hogan MP, Holding ML, Nystrom GS, et al. The genetic regulatory architecture and epigenomic basis for age-related changes in rattlesnake venom. *Proc Natl Acad Sci USA* 2024;121:e2313440121. <https://doi.org/10.1073/pnas.2313440121>.
 81. Li D, Purushotham D, Harrison JK, et al. WashU Epigenome Browser update 2022. *Nucleic Acids Res* 2022;50:W774–81. <https://doi.org/10.1093/nar/gkac238>.
 82. Bryant DM, Johnson K, DiTommaso T, et al. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep* 2017;18:762–76. <https://doi.org/10.1016/j.celrep.2016.12.063>.
 83. MacManes MD. The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ* 2018;6:e5428. <https://doi.org/10.7717/peerj.5428>.
 84. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;39:W29–W37. <https://doi.org/10.1093/nar/gkr367>.
 85. Blum M, Chang H-Y, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021;49:D344–54. <https://doi.org/10.1093/nar/gkaa977>.
 86. Mistry J, Chuguransky S, Williams L, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;49:D412–19. <https://doi.org/10.1093/nar/gkaa913>.
 87. Wang J, Chitsaz F, Derbyshire MK, et al. The conserved domain database in 2023. *Nucleic Acids Res* 2023;51:D384–88. <https://doi.org/10.1093/nar/gkac1096>.
 88. Agüero-Chapin G, Domínguez-Pérez D, Marrero-Ponce Y, et al. Unveiling encrypted antimicrobial peptides from cephalopods' salivary glands: a proteolysis-driven virtual approach. *ACS Omega* 2024. <https://doi.org/10.1021/acsomega.4c01959>.
 89. Macrander J, Panda J, Janies D, et al. Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. *PeerJ* 2018;6:e5361. <https://doi.org/10.7717/peerj.5361>.
 90. Gacesa R, Barlow DJ, Long PF. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *Peer J Comput Sci* 2016;2:e90. <https://doi.org/10.7717/peerj-cs.90>.
 91. Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ* 2019;7:e7200. <https://doi.org/10.7717/peerj.7200>.
 92. Ringeval A, Farhat S, Fedosov A, et al. DeTox: a pipeline for the detection of toxins in venomous organisms. *Briefings Bioinform* 2024;25:bbae094. <https://doi.org/10.1093/bib/bbae094>.
 93. Pan X, Zuallaert J, Wang X, et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* 2021;36:5159–68. <https://doi.org/10.1093/bioinformatics/btaa656>.
 94. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res* 2018;46:W84–88. <https://doi.org/10.1093/nar/gky350>.
 95. Zancolli G, Reijnders M, Waterhouse RM, et al. Convergent evolution of venom gland transcriptomes across Metazoa. *Proc Natl Acad Sci USA* 2022;119:e2111392119. <https://doi.org/10.1073/pnas.2111392119>.
 96. Perry BW, Schield DR, Westfall AK, et al. Physiological demands and signaling associated with snake venom production and storage illustrated by transcriptional analyses of venom glands. *Sci Rep* 2020;10:18083. <https://doi.org/10.1038/s41598-020-75048-y>.
 97. Haney RA, Ayoub NA, Clarke TH, et al. Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. *BMC Genomics* 2014;15:366. <https://doi.org/10.1186/1471-2164-15-366>.
 98. Bray NL, Pimentel H, Melsted P, et al. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–27. <https://doi.org/10.1038/nbt.3519>.
 99. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
 100. Tan CH, Tan KY, Tan NH. De novo assembly of venom gland transcriptome of *Tropidolaemus wagleri* (Temple pit viper, Malaysia) and insights into the origin of its major toxin, wagnerin. *Toxins* 2023;15:585. <https://doi.org/10.3390/toxins15090585>.
 101. So WL, Leung TCN, Nong W, et al. Transcriptomic and proteomic analyses of venom glands from scorpions *Liocheles australasiae*, *Mesobuthus martensii*, and *Scorpio maurus palmatus*. *Peptides* 2021;146:170643. <https://doi.org/10.1016/j.peptides.2021.170643>.
 102. Menk JJ, Matuhara YE, Sebestyen-França H, et al. Antimicrobial peptide arsenal predicted from the venom gland transcriptome of the tropical trap-jaw ant *Odontomachus chelifer*. *Toxins* 2023;15:345. <https://doi.org/10.3390/toxins15050345>.
 103. Xie B, Yu H, Kerkkamp H, et al. Comparative transcriptome analyses of venom glands from three scorpionfishes. *Genomics* 2019;111:231–41. <https://doi.org/10.1016/j.ygeno.2018.11.012>.
 104. Ramírez DS, Alzate JF, Simone Y, et al. Intersexual differences in the gene expression of *Phoneutria depilata* (Araneae, Ctenidae) toxins revealed by venom gland transcriptome analyses. *Toxins* 2023;15:429. <https://doi.org/10.3390/toxins15070429>.
 105. Chen C, Hou J, Tanner JJ, et al. Bioinformatics methods for mass spectrometry-based proteomics data analysis. *Int J Mol Sci* 2020;21:2873. <https://doi.org/10.3390/ijms21082873>.
 106. Perkins DN, Pappin DJC, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–67. [https://doi.org/10.1002/\(SICI\)1522-2683\(19991201\)20:18%3C3551::AID-ELPS3551%3E3.0.CO;2-2](https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18%3C3551::AID-ELPS3551%3E3.0.CO;2-2).
 107. Zhang J, Xin L, Shan B, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 2012;11:M111.010587. <https://doi.org/10.1074/mcp.M111.010587>.
 108. Orsburn BC. Proteome discoverer—a community enhanced data processing suite for protein informatics. *Proteomes* 2021;9:15. <https://doi.org/10.3390/proteomes9010015>.
 109. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–89. [https://doi.org/10.1016/1044-0305\(94\)80016-2](https://doi.org/10.1016/1044-0305(94)80016-2).
 110. Bern M, Kil YJ, Byonic BC. Advanced peptide and protein identification software. *Curr Protoc Bioinformatics* 2012;40:13.20.1–13.20.14. <https://doi.org/10.1002/0471250953.bi1320s40>.
 111. Chi H, Liu C, Yang H, et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat Biotechnol* 2018;36:1059–61. <https://doi.org/10.1038/nbt.4236>.

112. Kong AT, Leprevost FV, Avtonomov DM, et al. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 2017;14:513–20. <https://doi.org/10.1038/nmeth.4256>.
113. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 2015;33:22–24. <https://doi.org/10.1038/nbt.3109>.
114. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;26:1367–72. <https://doi.org/10.1038/nbt.1511>.
115. MacLean B, Tomazela DM, Shulman N, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010;26:966–68. <https://doi.org/10.1093/bioinformatics/btq054>.
116. Demichev V, Messner CB, Vernardis SI, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 2020;17:41–44. <https://doi.org/10.1038/s41592-019-0638-x>.
117. Sinitcyn P, Hamzeiy H, Salinas Soto F, et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat Biotechnol* 2021;39:1563–73. <https://doi.org/10.1038/s41587-021-00968-7>.
118. Doerr A. DIA mass spectrometry. *Nat Methods* 2015;12:35. <https://doi.org/10.1038/nmeth.3234>.
119. Ma B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* 2015;26:1885–94. <https://doi.org/10.1007/s13361-015-1204-0>.
120. Yang H, Chi H, Zeng W-F, et al. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* 2019;35:i183–90. <https://doi.org/10.1093/bioinformatics/btz366>.
121. Melani RD, Nogueira FCS, Domont GB. It is time for top-down venomomics. *J Venom Anim Toxins Incl Trop Dis* 2017;23:44. <https://doi.org/10.1186/s40409-017-0135-6>.
122. Röst HL, Sachsenberg T, Aiche S, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* 2016;13:741–48. <https://doi.org/10.1038/nmeth.3959>.
123. Schmid R, Heuckeroth S, Korf A, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* 2023;41:447–49. <https://doi.org/10.1038/s41587-023-01690-2>.
124. Liu X, Inbar Y, Dorrestein PC, et al. Deconvolution and database search of complex tandem mass spectra of intact proteins. *Mol Cell Proteomics* 2010;9:2772–82. <https://doi.org/10.1074/mcp.M110.002766>.
125. Adusumilli R, Mallick P. Data conversion with ProteoWizard msConvert. In: Comai L, Katz JE, Mallick P, eds. *Proteomics: Methods and Protocols*. New York: Springer US; 2017:339–68. https://doi.org/10.1007/978-1-4939-6747-6_23.
126. Liu X, Sirotkin Y, Shen Y, et al. Protein identification using top-down spectra. *Mol Cell Proteomics* 2012;11:M111.008524. <https://doi.org/10.1074/mcp.M111.008524>.
127. Guner H, Close PL, Cai W, et al. MASH Suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. *J Am Soc Mass Spectrom* 2014;25:464–70. <https://doi.org/10.1007/s13361-013-0789-4>.
128. Sun R-X, Luo L, Wu L, et al. pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. *Anal Chem* 2016;88:3082–90. <https://doi.org/10.1021/acs.analchem.5b03963>.
129. Kou Q, Wu S, Tolić N, et al. A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. *Bioinformatics* 2017;33:1309–16. <https://doi.org/10.1093/bioinformatics/btw806>.
130. Kou Q, Xun L, Liu X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 2016;32:3495–97. <https://doi.org/10.1093/bioinformatics/btw398>.
131. Park J, Piehowski PD, Wilkins C, et al. Informed-Proteomics: open-source software package for top-down proteomics. *Nat Methods* 2017;14:909–14. <https://doi.org/10.1038/nmeth.4388>.
132. Gupta S, Kapoor P, Chaudhary K, et al. In silico approach for predicting toxicity of peptides and proteins. *PLoS One* 2013;8:e73957. <https://doi.org/10.1371/journal.pone.0073957>.
133. Sharma N, Naorem LD, Jain S, et al. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief Bioinform* 2022;23:bbac174. <https://doi.org/10.1093/bib/bbac174>.
134. Mooney C, Haslam NJ, Pollastri G, et al. Towards the improved discovery and design of functional peptides: common features of diverse classes permit generalized prediction of bioactivity. *PLoS One* 2012;7:e45012. <https://doi.org/10.1371/journal.pone.0045012>.
135. Lamiable A, Thévenet P, Rey J, et al. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res* 2016;44:W449–54. <https://doi.org/10.1093/nar/gkw329>.
136. Rey J, Murail S, de Vries S, et al. PEP-FOLD4: a pH-dependent force field for peptide structure prediction in aqueous solution. *Nucleic Acids Res* 2023;51:W432–37. <https://doi.org/10.1093/nar/gkad376>.
137. Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. *Nat Methods* 2022;19:679–82. <https://doi.org/10.1038/s41592-022-01488-1>.
138. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;373:871–76. <https://doi.org/10.1126/science.abj8754>.
139. Källberg M, Wang H, Wang S, et al. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 2012;7:1511–22. <https://doi.org/10.1038/nprot.2012.085>.
140. Choi M, Carver J, Chiva C, et al. MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat Methods* 2020;17:981–84. <https://doi.org/10.1038/s41592-020-0955-0>.
141. Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res* 2007;35:D521–26. <https://doi.org/10.1093/nar/gkl923>.
142. Alonso LL, Slagboom J, Casewell NR, et al. Metabolome-based classification of snake venoms by bioinformatic tools. *Toxins* 2023;15:161. <https://doi.org/10.3390/toxins15020161>.
143. Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 2016;34:828–37. <https://doi.org/10.1038/nbt.3597>.
144. Fischer T, Riedl R. Paracelsus' legacy in the faunal realm: drugs deriving from animal toxins. *Drug Discov Today* 2022;27:567–75. <https://doi.org/10.1016/j.drudis.2021.10.003>.
145. Crnković A, Srnko M, Anderluh G. Biological nanopores: engineering on demand. *Life* 2021;11:27. <https://doi.org/10.3390/li11010027>.
146. Tyagi A, Tuknait A, Anand P, et al. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015;43:D837–43. <https://doi.org/10.1093/nar/gku892>.

147. di Micco P, Antolin AA, Mitsopoulos C, et al. canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res* 2023;51:D1212–19. <https://doi.org/10.1093/nar/gkac1004>.
148. Faraji N, Arab SS, Doustmohammadi A, et al. ApInAPDB: a database of apoptosis-inducing anticancer peptides. *Sci Rep* 2022;12:21341. <https://doi.org/10.1038/s41598-022-25530-6>.
149. Cadow J, Born J, Manica M, et al. PaccMann: a web service for interpretable anticancer compound sensitivity prediction. *Nucleic Acids Res* 2020;48:W502–8. <https://doi.org/10.1093/nar/gkaa327>.
150. Petrov I, Alexeyenko A. EviCor: interactive web platform for exploration of molecular features and response to anti-cancer drugs. *J Mol Biol* 2022;434:167528. <https://doi.org/10.1016/j.jmb.2022.167528>.
151. Romano JD, Tatonetti NP. VenomKB, a new knowledge base for facilitating the validation of putative venom therapies. *Sci Data* 2015;2:150065. <https://doi.org/10.1038/sdata.2015.65>.
152. SIB Swiss Institute of Bioinformatics RDF Group Members. The SIB Swiss Institute of Bioinformatics Semantic Web of data. *Nucleic Acids Res* 2023;52:D44–51. <https://doi.org/10.1093/nar/gkad902>.
153. Sima AC, Mendes de Farias T, Zbinden E, et al. Enabling semantic queries across federated bioinformatics databases. *Database* 2019;2019:baz106. <https://doi.org/10.1093/database/baz106>.
154. Galgonek J, Vondrášek J. IDSMS ChemWebRDF: sPARQLing small-molecule datasets. *J Cheminform* 2021;13:38. <https://doi.org/10.1186/s13321-021-00515-1>.
155. Groza T, Gomez FL, Mashhadi HH, et al. The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res* 2023;51:D1038–45. <https://doi.org/10.1093/nar/gkac972>.
156. Howe DG, Bradford YM, Eagle A, et al. The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. *Nucleic Acids Res* 2017;45:D758–68. <https://doi.org/10.1093/nar/gkw1116>.
157. Gramates LS, Agapite J, Attrill H, et al. FlyBase: a guided tour of highlighted features. *Genetics* 2022;220:iyac035. <https://doi.org/10.1093/genetics/iyac035>.
158. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* 2011;39:W270–77. <https://doi.org/10.1093/nar/gkr366>.
159. Romero-Molina S, Ruiz-Blanco YB, Mieres-Perez J, et al. PPI-affinity: a web tool for the prediction and optimization of protein-peptide and protein-protein binding affinity. *J Proteome Res* 2022;21:1829–41. <https://doi.org/10.1021/acs.jproteome.2c00020>.
160. Lei Y, Li S, Liu Z, et al. A deep-learning framework for multi-level peptide-protein interaction prediction. *Nat Commun* 2021;12:5465. <https://doi.org/10.1038/s41467-021-25772-4>.
161. Vidal-Limon A, Aguilar-Toalá JE, Liceaga AM. Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides. *J Agric Food Chem* 2022;70:934–43. <https://doi.org/10.1021/acs.jafc.1c06110>.
162. Almeida JR, Palacios ALV, Patiño RSP, et al. Harnessing snake venom phospholipases A2 to novel approaches for overcoming antibiotic resistance. *Drug Dev Res* 2019;80:68–85. <https://doi.org/10.1002/ddr.21456>.
163. Clark GC, Casewell NR, Elliott CT, et al. Friends or foes? Emerging impacts of biological toxins. *Trends Biochem Sci* 2019;44:365–79. <https://doi.org/10.1016/j.tibs.2018.12.004>.
164. Holmes DE. The data explosion. In: Holmes DE, ed. *Big Data: A Very Short Introduction*. Oxford, UK: Oxford University Press; 2017. <https://doi.org/10.1093/actrade/9780198779575.003.0001>.
165. Di Muri C, Pulieri M, Raho D, et al. Assessing semantic interoperability in environmental 1 sciences: variety of approaches and semantic artefacts. *Scientific Data* 2024; <https://doi.org/10.1038/s41597-024-03669-3>
166. Ong E, Xiang Z, Zhao B, et al. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res* 2017;45:D347–52. <https://doi.org/10.1093/nar/gkw918>.
167. Emam M, Tarek A, Soudy M, et al. Comparative evaluation of multiomics integration tools for the study of prediabetes: insights into the earliest stages of type 2 diabetes mellitus. *Netw Model Anal Health Inform Bioinforma* 2024;13:8. <https://doi.org/10.1007/s13721-024-00442-9>.
168. Wilkinson MD, Dumontier M, IJJ A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>.
169. Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011;39:W541–45. <https://doi.org/10.1093/nar/gkr469>.
170. European Commission. Directorate-General for Research and Innovation. Turning FAIR into reality—final report and action plan from the European Commission expert group on FAIR data. Brussels: Publications Office; 2018. <https://dx.doi.org/10.2777/54599>.