scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A comprehensive dataset of therapeutic peptides on multi-function property and structure information

Baichuan Xiao[1,6], Yixiang Zhou[1,6], Long Zhao[2,6], Haodong Huang[1], Xuwen Fei[3] & Yong-Biao Zhang[1,4,5] ✉

This paper presents a comprehensive dataset comprising 58,583 experimentally validated therapeutic peptides with annotated structure information. These peptides are grouped into 47 categories based on their function or therapeutic property like antimicrobial or glucose-regulatory, of which 21,130 are multi-function peptides and 54,722 possess structural annotation information. We believe this dataset can be useful for the relevant research of therapeutic peptides, especially for computational tool developments in therapeutic peptide discovery and further exploration of the 'sequence-structure-function' relationship for therapeutic peptides.

Background & Summary

Peptide drugs have emerged as important therapeutic agents due to their unique advantages, such as high specificity, low immunogenicity, and high potency[1–3]. They have shown promising development trends with significant growth potential[2]. Fully understanding therapeutic peptides is a prerequisite for accelerating the development of peptide drugs[1,3]. As small proteins, peptides adhere to the 'sequence-structure-function' dogma and exhibit moonlighting characteristics, that is, multifunctionality[4,5]. Therefore, a comprehensive dataset of therapeutic peptides that includes sequence, structure, and functional information, especially multifunctional information, is crucial for fully understanding underlying principles of peptide drug discovery and design.

In recent years, numerous peptide databases have been established to integrate various peptides[6–10]. These include databases dedicated to specific functions such as antimicrobial peptides[9,11,12], antiviral peptides[10,13,14], glucose-regulating peptides[15], or peptide hormones[16,17], as well as comprehensive databases that collect various functional peptides, like SATPdb[18], EROP-Moscow[19], and BIOPEP-UWM[20]. However, these databases still lack sufficient attention to multifunctional peptides and structural information[18,21,22]. Among these databases, the database with the largest collection of multifunctional peptides contains only 9,986 entries[18]. Given the ubiquity of moonlighting, more complete and diverse functional information enable better elucidating the 'sequence-structure-function' relationship but also facilitates the repurposing of peptide drugs[23–25]. Further, these databases rarely comprise structural information, and structure data in some of these databases are usually obtained through traditional structure prediction tools[26,27]. Specifically, the best-performing databases/datasets current includes only 16,131 structural annotation entries for therapeutic peptides[18]. In recent years, structure prediction tools represented by AlphaFold2 have brought significant breakthroughs to the protein structure prediction field[28,29]. Acquiring more accurate structural information based on these latest prediction tools is undeniably important for understanding the 'sequence-structure-function' relationship of peptide[30].

[1]School of Engineering Medicine, Beihang University, Beijing, 100191, China. [2]School of Materials and Chemical Engineering, Beijing Institute of Petrochemical Technology, Beijing, 102627, China. [3]School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China. [4]Key Laboratory of Big Data-Based Precision Medicine and Key Laboratory of Innovation and Transformation of Advanced Medical Devices, Ministry of Industry and Information Technology; Key Laboratory of Biomechanics and Mechanobiology, Ministry of Education, Beijing, China. [5]National Medical Innovation Platform for Industry-Education Integration in Advanced Medical Devices (Interdiscipline of Medicine and Engineering), Beihang University, Beijing, 100083, China. [6]These authors contributed equally: Baichuan Xiao, Yixiang Zhou, Long Zhao. ✉e-mail: zhangyongbiao@buaa.edu.cn

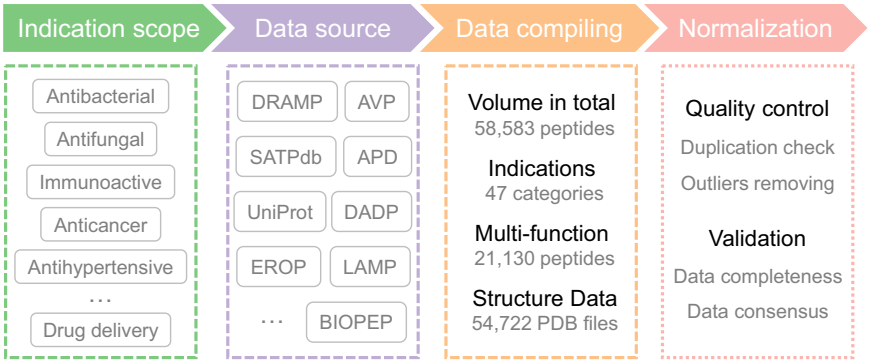

Fig. 1 Technical process for data compiling. The pipeline consists of four major steps (from left to right): (1) Indication scope (green box), describing the 47 specific function classes focused by this study; (2) Data source (purple box), listing the 33 databases or datasets used as data source of our dataset; (3) Data compiling (orange box), comprising 21,130 multifunctional peptides, 54,722 tertiary structure files; (4) Normalization (pink box), which describes quality measures like duplication checks or outlier removal, completeness validation, and consensus validation in this study.

Datasets / datasets	Volume in total	Number of categories	Number of multi-function peptides	Number of structure files
Ours	58,583	47	21,130	54,722
Other	30,260	22	9,986	16,131

Table 1. Comparison between our dataset and other databases/datasets. For each metric (each column of the table), we show only the best-performing databases/datasets and their values.

In this study, we constructed a dataset[31] comprising 58,583 experimentally validated therapeutic peptides (Fig. 1). These peptides are classified into 47 categories based on their functions or therapeutic properties, among which 21,130 peptides are multifunctional. We obtained experimentally determined structural information for 54,722 therapeutic peptides from the Protein Data Bank and predicted structural information of the other ones using computational tools like AlphaFold2. To our knowledge, this data descriptor[31] is the most comprehensive and largest dataset of therapeutic peptides currently available, especially in terms of multifunctional peptides and structural information (Table 1). In this dataset[31], therapeutic peptides possess information of, amino acid sequence, function, sequence modifications, origin, secondary structure, and tertiary structure. Based on this dataset[31], computational pipelines can be readily constructed for therapeutic peptide discovery, such as antimicrobial peptide prediction, multifunctional peptide prediction, and peptide drug repurposing. And, users can explore the 'sequence-structure-function' relationships of therapeutic peptides as well, which is crucial for understanding the underlying principles of peptide drug design. We believe this dataset[31] is helpful in the discovery and design of peptide drugs.

Methods

Data source. We collected 58,583 experimentally validated therapeutic peptides from Uniprot (https://www.uniprot.org/)[32] and 32 publicly accessible therapeutic peptide databases (Supplementary Table 1), including AntiAngioPred (https://webs.iiitd.edu.in/raghava/antiangiopred/index.html)[6], Hemolytik (https://webs.iiitd.edu.in/raghava/satpdb/catalogs/hemolytik)[33], DADP (http://split4.pmfst.hr/dadp/)[34], HIPdb (https://webs.iiitd.edu.in/raghava/satpdb/catalogs/hipdb)[14], AHTPDB (https://webs.iiitd.edu.in/raghava/ahtpdb)[7], Fermfoodb (https://webs.iiitd.edu.in/raghava/fermfoodb)[35], BaAMPs (http://www.baamps.it)[11], EROP-Moscow (http://erop.inbi.ras.ru)[19], LAMP (http://biotechlab.fudan.edu.cn/database/lamp/index.php)[36], DRAVP (http://dravp.cpu-bioinfor.org)[13], NeuroPedia (http://isyslab.info/NeuroPep)[37], Cancerppd (https://webs.iiitd.edu.in/raghava/cancerppd)[38], BioPEPDB (https://bis.zju.edu.cn/biopepdbr/)[39], HMRbase (https://webs.iiitd.edu.in/raghava/hmrbase2/)[16], HORDB (http://hordb.cpu-bioinfor.org/)[17], DRAMP 4.0 (http://dramp.cpu-bioinfor.org/)[40], BIOPEP-UMW (https://biochemia.uwm.edu.pl/biopep/peptide_data.php)[20], SATPdb (https://webs.iiitd.edu.in/raghava/vaxinpad/index1.html)[18], TumorHoPe (https://webs.iiitd.edu.in/raghava/tumorhope/index.php)[41], PlantPepDB (http://www.nipgr.ac.in/PlantPepDB/)[42], YADAMP (https://webs.iiitd.edu.in/raghava/satpdb/catalogs/yadamp/)[43], APD3 (http://aps.unmc.edu/AP/)[9], THPDB2 (https://webs.iiitd.edu.in/raghava/thpdb2/)[44], AVPdb (https://webs.iiitd.edu.in/raghava/satpdb/catalogs/avpdb/)[10], AntiTbPdb (https://webs.iiitd.edu.in/raghava/antitbpdb/)[8], DBAASP v3 (https://dbaasp.org/home)[45], StraPep (http://isyslab.info/StraPep/)[27], CPPsite (https://webs.iiitd.edu.in/raghava/cppsite1/)[26], BioDADpep (http://omicsbase.com/BioDADPep/)[15], CAMP R4 (https://camp.bicnirrh.res.in/index.php)[12], MBPDB (http://mbpdb.nws.oregonstate.edu)[46] and DFBP (http://www.cqudfbp.net/)[47]. Specifically, we first categorized all of the 120 approved peptide drugs recorded from PepTherDia (http://peptherdia.herokuapp.com/)[48] into 14 categories based on their therapeutic properties, including neuropeptide, antimicrobial, anticancer, antihypertensive, immunoactive, glucose regulatory, antiviral, osteogenic, analgesic, thrombolytic, growth regulatory, lipid metabolism, antioxidant, and skin regeneration. These categories of peptides address urgent clinical needs and are the primary focus of our dataset[31].

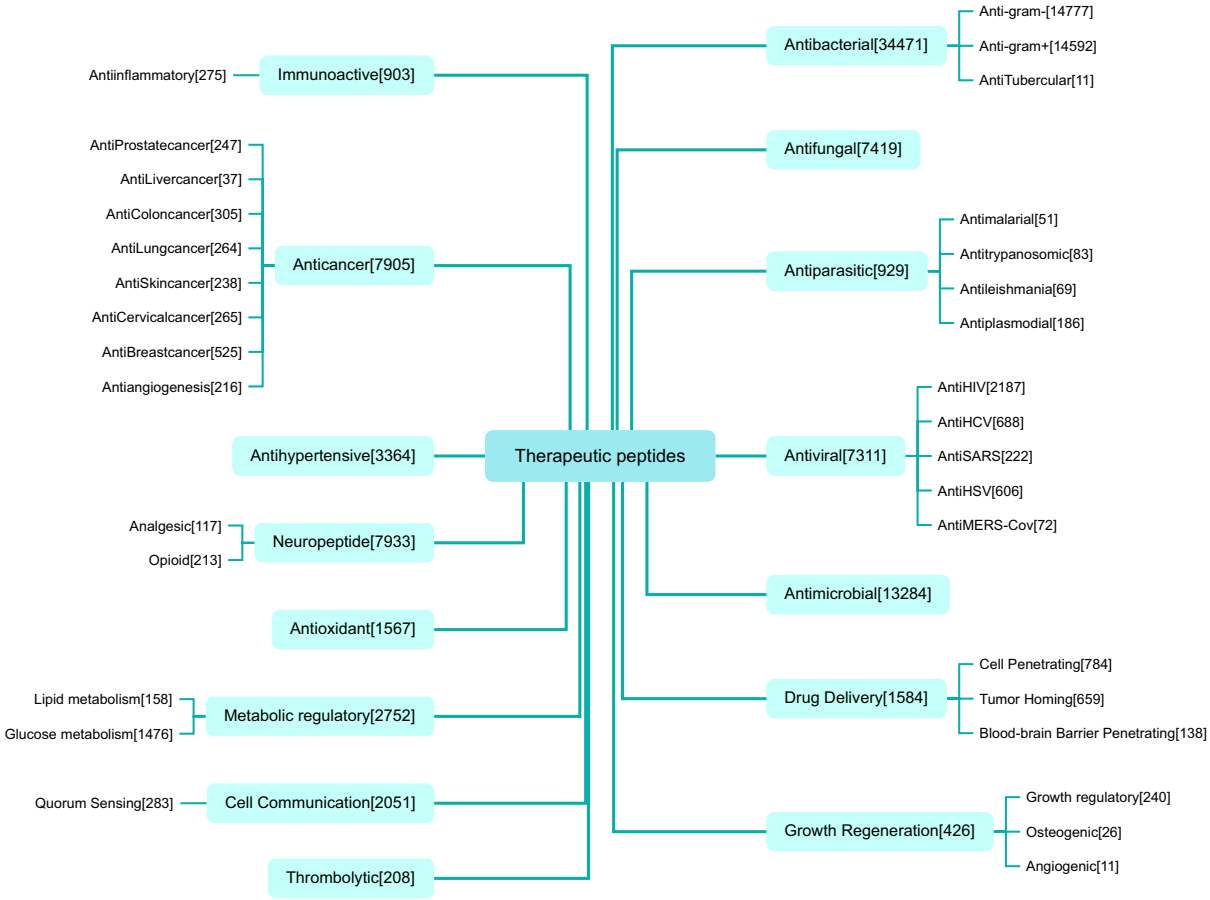

Fig. 2 Dataset comprising 15 major categories and 47 subcategories. Therapeutic peptides are classified into 15 major categories, each with specific subcategories.

We then used combinations of 'category keywords (listed above) AND peptide AND database/dataset' to search for relevant databases or datasets via Google Scholar. Further, we gathered commonly used peptide databases/datasets of research papers by searching highly cited research and review articles on computational modeling of therapeutic peptides through Google Scholar, with query terms of 'category keywords (listed above) AND peptide AND (prediction OR discovery OR mining)'. Ultimately, we obtained the 33 therapeutic peptide databases/datasets (Fig. 1).

Criteria for categories. As illustrated in Fig. 2, our data descriptor[31] categorizes peptides into 15 major categories and 47 subcategories based on their functions or therapeutic properties. The 15 major categories include: antibacterial, antifungal, antiparasitic, antiviral, antimicrobial, immunoactive, anticancer, antihypertensive, thrombolytic, neuropeptide, antioxidant, metabolic regulatory, growth/regeneration, cell communication, and drug delivery.

Notably, categories such as antibacterial, antifungal, antiparasitic, and antiviral theoretically fall under antimicrobial. However, considering that these categories have many subcategories, they are treated as independent major categories. Consequently, the antimicrobial category includes only those peptides that are not classified under antibacterial, antifungal, antiparasitic, or antiviral.

The 47 subcategories consist of 32 specific subcategories and 15 'unknown label' subcategories. Each major category contains one 'unknown label' subcategory, which is used to classify peptides that do not belong to any of the defined specific subcategories under that major category. For example, the major category metabolic regulatory includes two specific subcategories, lipid metabolism and glucose metabolism, as well as an "unknown label" subcategory (not shown in Fig. 2), which accommodates peptides that do not fall under either lipid metabolism or glucose metabolism.

The 32 specific subcategories include: anti-gram+, anti-gram-, anti-tubercular, anti-malarial, anti-trypanosomic, anti-leishmania, anti-plasmodial, anti-HIV, anti-HCV, anti-SARS, anti-HSV, anti-MERS-Cov, anti-inflammatory, anti-breast-cancer, anti-cervical-cancer, anti-colon-cancer, anti-lung-cancer, anti-prostate-cancer, anti-skin-cancer, anti-angiogenesis, anti-liver-cancer, analgesic, opioid, lipid metabolism, glucose metabolism, growth regulatory, angiogenic, osteogenic, quorum sensing, cell penetrating, blood-brain barrier penetrating, and tumor homing.

The 15 'unknown label' subcategories adopt the names of their corresponding major categories. For instance, the 'unknown label' subcategory under metabolic regulatory is named metabolic regulatory and includes peptides that cannot be classified under either lipid metabolism or glucose metabolism.

Definition of multi-function property. A therapeutic peptide is identified as multifunctional when it possesses at least two functional labels that do not have a subordinate relationship[4,18]. For example, a peptide with both antimicrobial and antibacterial labels is not considered multifunctional because antibacterial falls within the scope of antimicrobial. However, a peptide with both anti-gram+ and anti-angiogenesis labels is classified as multifunctional because these two functions do not have a subordinate relationship.

Structure information. To obtain structural information for these peptides, we first retrieved the experimentally determined tertiary structure files for 179 peptides from the PDB (Protein Data Bank) database (https://www.rcsb.org/)[49]. Then, we used AlphaFold2 and ESMFold to predict the spatial structures of natural peptides with sequence lengths over or less than 16 amino acids, respectively. Further, we utilized PEPstrMOD[50] to predict structure for the remaining peptides with simple modifications. Finally, we calculated the secondary structure information of natural peptides based on their tertiary structure files using DSSP[51].

Technically, we use AlphaFold2 for tertiary structure prediction of peptide sequences, paired with UniRef databases (UniRef30_2202/PDB70_220313) of UniProt. Considering the short length of peptides, the number of models is set to 1, and the number of recursive cycles is set to 1. Disable template structure guidance (use_templates = False) and do not enable the AMBER energy minimization post-processing step (num_delax = 0). When the average prediction reliability score (pLDDT) of the model is higher than 90 points, the subsequent model prediction process is terminated in advance to make a balance of time and accuracy (stop_ot_score = 90). As results, we obtained structural files of 24,746 peptide sequences.

We access the ESMFold online platform (https://esmatlas.com/resources?action=fold) and input peptide's single-letter amino acid sequence into the designated text box, using default parameters. Upon submission, the system automatically performs the structure prediction, and then a corresponding PDB file is provided for download. As results, we obtained structural files of 29,162 peptide sequences.

For PEPstrMOD (https://webs.iiitd.edu.in/raghava/pepstrmod/), we access its online webserver and utilize the 'Advance Modification/Beginner' module. After selecting amino acid and its modification type for each position of peptide sequence, we run the prediction with default parameters and then download the predicted PDB file. A total of 635 peptide structural files were obtained using the PEPstrMOD approach.

Data compilation and processing. We initially compiled peptide data from 33 databases and datasets (Fig. 1), unifying all peptide sequences using the single-letter amino acid abbreviation standard. During the data compilation process, sequence cleaning and sequence normalization were applied to ensure data quality and consistency.

Sequence cleaning was primarily aimed at addressing two types of issues: (1) peptide sequences containing unknown amino acids or ambiguous characters without clear annotations; (2) peptides that primarily served auxiliary roles in the labeled functions (e.g., peptides functioning as drug delivery carriers rather than directly contributing to the anticancer therapeutic effect).

Sequence normalization involved the use of HELM (Hierarchical Editing Language for Macromolecules)[52] notation to encode complex structural information, ensuring a standardized representation of peptide structures.

To further refine the dataset, we retained only peptides with lengths between 2 and 50 amino acids, grouping them according to the functional labels provided in the original datasets. For peptides with unclear functional labels, their functional categories were determined through a manual review of the associated references.

As a final step, we performed sequence deduplication within each functional category. After deduplication, only functional categories containing more than 50 unique peptide sequences were retained, with the exception of 14 clinically urgent categories, which were included regardless of the sequence count.

All data processing steps, including cleaning, normalization, and deduplication, were implemented using custom Python scripts without the use of external tools.

Data Records

The dataset[31] is available at FigShare (https://doi.org/10.6084/m9.figshare.28691885), with this section being the primary source of information on the availability and content of the data being described. The dataset[31] includes a worksheet named 'main.xlsx' and PDB format files of peptide tertiary structure. The main.xlsx file records all information except for tertiary structures, of which each data record includes items of ID, amino acid sequence, function, modifications, origin, is_natural_peptide, HELM notation, label encoding, secondary structure, tertiary structure (available for 93% records) and references. Specifically, modifications here include N-ter modifications like acetylation, C-ter modifications like amidation, and post-translational modifications like methylation. Origin describes species or substances the peptide may originate from, such as human or milks. L-type and D-type amino acids in peptide sequence are represented by uppercase and lowercase letters, respectively.

To access the dataset[31], users can visit the FigShare link provided above. The main.xlsx is located in the root directory of the dataset, and the PDB format files of tertiary structure are named by their uniquely peptide ID and deposited in a subfolder named 'TertiaryStructure' in the same root directory as the main.xlsx. User can easily access and download the dataset[31] directly from the FigShare page.

Technical Validation

Data completeness. We analyzed the data completeness in terms of peptide sequence, structure, function, and origin information. As results, all peptides have amino acid sequence and function information. 93% of the peptides have structural information, with the remainder primarily lacking this due to complex modification information, making tertiary structure prediction difficult. 75% of the peptides have origin information.

We further analyzed the distribution of peptide lengths, origins, and functional categories (Fig. 3). The length distribution covers every unit from 2 to 50. The origin distribution spans over 4,500 species, and 42 out of the

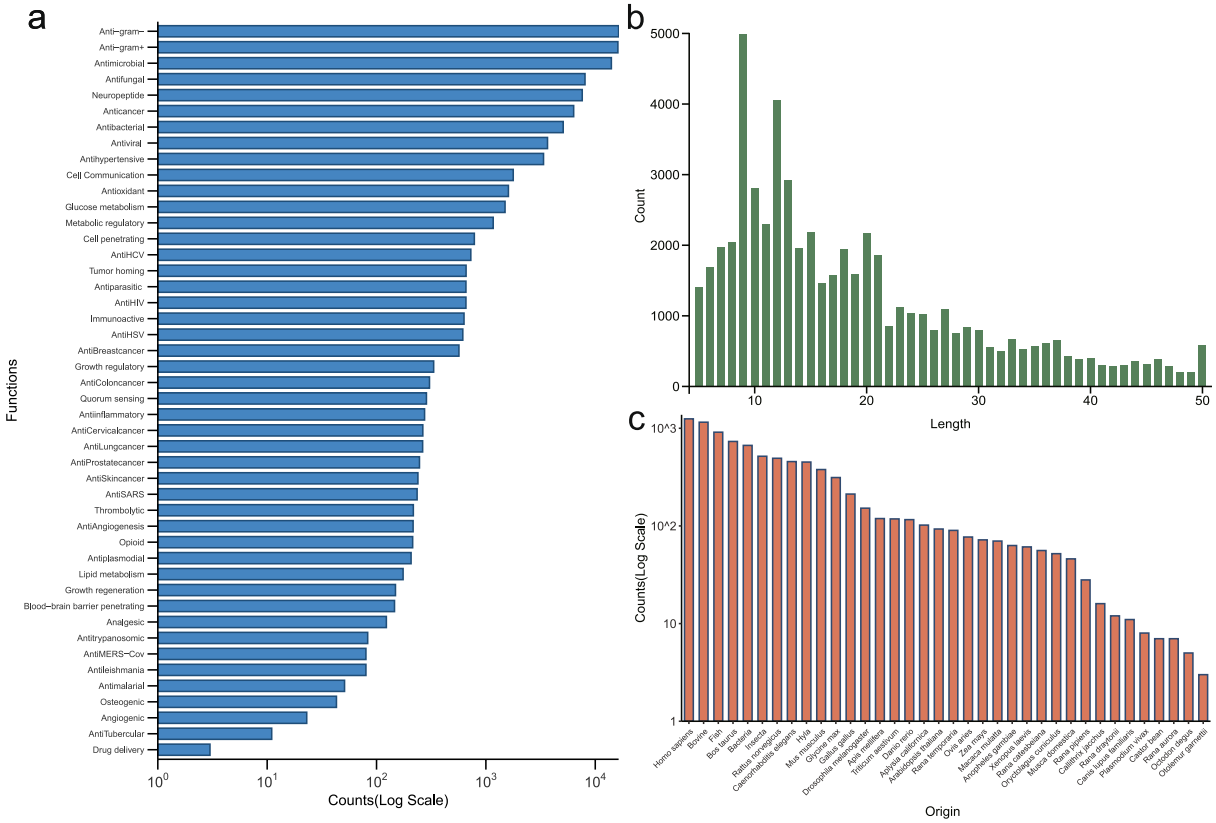

Fig. 3 Distribution of peptide lengths, origins, and function categories. (**a**–**c**) Distribution of functional categories, sequence length and origin of therapeutic peptides.

47 functional categories have at least 50 peptides, with 38 having at least 100 peptides. These analyses indicate a high level of data completeness in our dataset.

Data consensus. Since our data is integrated from 33 different databases, we validated the information consistency of various databases. Given the varied ways of expressing information across different databases, we used sequences as the unique identifier for each peptide. Therefore, for peptides present in various databases, we first verified sequence consistency. For entries from the same reference but with inconsistent sequences, we manually verified to determine a unique sequence.

Further, we checked the consistency of origin and function information, mainly addressing issues like spelling errors, case differences (e.g., Antiinflammatory and anti-inflammation), or synonymous expressions (e.g., alphabet cone snail and alphabet cone; anti-cancer and anti-tumor). Typically, annotations across different databases describe same origin or function information with slightly different terms. To address this, we first compile all possible labels under the same functional or origin category, and then determine the final retained label based on its occurrence frequency across the databases. If two labels have the same frequency, we prioritize the label that appears more frequently in databases published in high-impact journals. This process ensures the use of a consistent and widely accepted terminology for each origin or function information.

Finally, for origin and function information with different annotation levels in different databases, we generally chose the result with a higher annotation level unless the peptide's reference information supports the other one. For example, between bacteria and Bacillus pumilus, we chose Bacillus pumilus in principle. Between antimicrobial and antibacterial, we usually chose antibacterial. All steps were implemented using custom Python code without the use of other tools.

Usage Notes

This section primarily discusses the use of this dataset[31] in the programming languages Python (v3.8) and R (v4.4), as downstream analyses based on this dataset—whether constructing computational pipelines for discovering therapeutic peptides or exploring the 'sequence-structure-function' relationship of therapeutic peptides—are mainly based on these two languages.

In Python, we can load the dataset's[31] worksheets using common data processing modules such as pandas[53], or numpy[54]. For the loaded dataframe, users can implement feature engineering by common peptide feature encoding modules like AAindex[55], modlamp[56], or common small molecule feature encoding module like Mordred[57] (first converting amino acid sequences to molecular format using 'Chem.MolFromFASTA' function in rdkit[58] module), and then conduct their interested downstream analysis based on feature encoding results. Similarly, in

R, after loading dataframe with 'read.csv' function, users can obtain various feature information of therapeutic peptides using peptide feature encoding package like AAindex, and then perform data analyses of interest.

Code availability

There was no custom code to generate the data. Code for the relevant data preprocessing and analysis in this study is available in FigShare (https://doi.org/10.6084/m9.figshare.28691885).

Received: 1 April 2025; Accepted: 2 July 2025;
Published online: 14 July 2025

References

1. Basith, S., Manavalan, B., Hwan Shin, T. & Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.* **40**, 1276–1314 (2020).
2. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discov.* **20**, 309–325 (2021).
3. Chen, Z., Wang, R., Guo, J. & Wang, X. The role and future prospects of artificial intelligence algorithms in peptide drug development. *Biomed. Pharmacother.* **175**, 116709 (2024).
4. Zanzoni, A., Ribeiro, D. M. & Brun, C. Understanding protein multifunctionality: from short linear motifs to cellular functions. *Cell. Mol. Life Sci.* **76**, 4407–4412 (2019).
5. Kustatscher, G. *et al*. Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods.* **19**, 774–779 (2022).
6. Ettayapuram Ramaprasad, A. S., Singh, S., Gajendra, P. S. R. & Venkatesan, S. AntiAngioPred: a server for prediction of anti-angiogenic peptides. *PLoS One.* **10**, e0136990 (2015).
7. Kumar, R. *et al*. AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic. Acids. Res.* **43**, D956–D962 (2015).
8. Usmani, S. S., Kumar, R., Kumar, V., Singh, S. & Raghava, G. P. AntiTbPdb: a knowledgebase of anti-tubercular peptides. *Database* **2018**, bay025 (2018).
9. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic. Acids. Res.* **44**, D1087–D1093 (2016).
10. Qureshi, A., Thakur, N., Tandon, H. & Kumar, M. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic. Acids. Res.* **42**, D1147–D1153 (2014).
11. Di Luca, M., Maccari, G., Maisetta, G. & Batoni, G. BaAMPs: the database of biofilm-active antimicrobial peptides. *Biofouling* **31**, 193–199 (2015).
12. Gawde, U. *et al*. CAMPR4: a database of natural and synthetic antimicrobial peptides. *Nucleic. Acids. Res.* **51**, D377–D383 (2023).
13. Liu, Y. *et al*. DRAVP: a comprehensive database of antiviral peptides and proteins. *Viruses* **15**, 820 (2023).
14. Qureshi, A., Thakur, N. & Kumar, M. HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One* **8**, e54908 (2013).
15. Roy, S. & Teron, R. BioDADPep: a bioinformatics database for anti diabetic peptides. *Bioinformation* **15**, 780 (2019).
16. Rashid, M., Singla, D., Sharma, A., Kumar, M. & Raghava, G. P. Hmrbase: a database of hormones and their receptors. *BMC Genomics* **10**, 1–10 (2009).
17. Zhu, N., Dong, F., Shi, G., Lao, X. & Zheng, H. HORDB a comprehensive database of peptide hormones. *Sci. Data.* **9**, 187 (2022).
18. Singh, S. *et al*. SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic. Acids. Res.* **44**, D1119–D1126 (2016).
19. Zamyatnin, A. A., Borchikov, A. S., Vladimirov, M. G. & Voronina, O. L. The EROP-Moscow oligopeptide database. *Nucleic. Acids. Res.* **34**, D261–D266 (2006).
20. Minkiewicz, P., Iwaniak, A. & Darewicz, M. BIOPEP-UWM database of bioactive peptides: Current opportunities. *International Journal of Molecular Sciences* **20**, 5978 (2019).
21. Apostolopoulos, V. *et al*. A global review on short peptides: frontiers and perspectives. *Molecules* **26**, 430 (2021).
22. Wang, L. *et al*. Therapeutic peptides: current applications and future directions. *Signal Transduct. Target. Ther.* **7**, 48 (2022).
23. Koehler Leman, J. *et al*. Sequence-structure-function relationships in the microbial protein universe. *Nat. Commun.* **14**, 2351 (2023).
24. Randall, J. R., Vieira, L. C., Wilke, C. O. & Davies, B. W. Deep mutational scanning and machine learning for the analysis of antimicrobial-peptide features driving membrane selectivity. *Nat. Biomed. Eng.* **8**, 842–853 (2024).
25. Roy, S., Dhaneshwar, S. & Bhasin, B. Drug repurposing: an emerging tool for drug reuse, recycling and discovery. *Current Drug Research Reviews Formerly: Current Drug Abuse Reviews* **13**, 101–119 (2021).
26. Kardani, K. & Bolhassani, A. Cppsite 2.0: an available database of experimentally validated cell-penetrating peptides predicting their secondary and tertiary structures. *J. Mol. Biol.* **433**, 166703 (2021).
27. Wang, J. *et al*. StraPep: a structure database of bioactive peptides. *Database* **2018**, bay038 (2018).
28. Jumper, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
29. Lin, Z. *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
30. Kumar, H. & Kim, P. Artificial intelligence in fusion protein three-dimensional structure prediction: Review and perspective. *Clin. Transl. Med.* **14**, e1789 (2024).
31. Xiao, B.-C., Zhou, Y.-Y. & Zhao, L. Dataset of therapeutic peptides on multi-function property and structure information. *figshare* https://doi.org/10.6084/m9.figshare.28691885 (2025).
32. UniProt: the Universal protein knowledgebase in 2025. *Nucleic. Acids. Res*., gkae1010 (2024).
33. Gautam, A. *et al*. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic. Acids. Res.* **42**, D444–D449 (2014).
34. Novković, M., Simunić, J., Bojović, V., Tossi, A. & Juretić, D. DADP: the database of anuran defense peptides. *Bioinformatics* **28**, 1406–1407 (2012).
35. Chaudhary, A., Bhalla, S., Patiyal, S., Raghava, G. P. & Sahni, G. FermFooDb: A database of bioactive peptides derived from fermented foods. *Heliyon* **7**, e06668 (2021).
36. Ye, G. *et al*. LAMP2: a major update of the database linking antimicrobial peptides. *Database*. **2020** (2020).
37. Kim, Y., Bark, S., Hook, V. & Bandeira, N. NeuroPedia: neuropeptide database and spectral library. *Bioinformatics* **27**, 2772–2773 (2011).
38. Tyagi, A. *et al*. CancerPPD: a database of anticancer peptides and proteins. *Nucleic. Acids. Res.* **43**, D837–D843 (2015).
39. Li, Q. *et al*. BioPepDB: an integrated data platform for food-derived bioactive peptides. *Int. J. Food Sci. Nutr.* **69**, 963–968 (2018).
40. Ma, T. *et al*. DRAMP 4.0: an open-access data repository dedicated to the clinical translation of antimicrobial peptides. *Nucleic. Acids. Res.* **53**, D403–D410 (2025).
41. Kapoor, P. *et al*. TumorHoPe: a database of tumor homing peptides. *PLoS One* **7**, e35187 (2012).

42. Das, D., Jaiswal, M., Khan, F. N., Ahamad, S. & Kumar, S. PlantPepDB: A manually curated plant peptide database. *Sci. Rep.* **10**, 2194 (2020).
43. Piotto, S. P., Sessa, L., Concilio, S. & Iannelli, P. YADAMP: yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents.* **39**, 346–351 (2012).
44. Jain, S., Gupta, S., Patiyal, S. & Raghava, G. P. THPdb2: compilation of FDA approved therapeutic peptides and proteins. *Drug Discov. Today*, 104047 (2024).
45. Pirtskhalava, M. *et al*. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic. Acids. Res.* **49**, D288–D297 (2021).
46. Nielsen, S. D., Beverly, R. L., Qu, Y. & Dallas, D. C. Milk bioactive peptide database: A comprehensive database of milk protein-derived bioactive peptides and novel visualization. *Food Chem* **232**, 673–682 (2017).
47. Qin, D. *et al*. DFBP: a comprehensive database of food-derived bioactive peptides for peptidomics research. *Bioinformatics* **38**, 3275–3280 (2022).
48. D'Aloisio, V., Dognini, P., Hutcheon, G. A. & Coxon, C. R. PepTherDia: Database and structural composition analysis of approved peptide therapeutics and diagnostics. *Drug Discov. Today.* **26**, 1409–1419 (2021).
49. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).
50. Singh, S. *et al*. PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues. *Biol. Direct.* **10**, 1–19 (2015).
51. Hekkelman, M. L., Alvarez Salmoral, D., Perrakis, A. & Joosten, R. P. DSSP 4: FAIR annotation of protein secondary structure. *BioRxiv*, 2024-2025 (2025).
52. Zhang, T., Li, H., Xi, H., Stanton, R. V. & Rotstein, S. H. HELM: a hierarchical notation language for complex biomolecule structure representation.: ACS Publications; 2012.
53. McKinney, W. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* **14**, 1–9 (2011).
54. Harris *et al*. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
55. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic. Acids. Res.* **28**, 374 (2000).
56. Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **33**, 2753–2755 (2017).
57. Moriwaki, H., Tian, Y., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics.* **10**, 4 (2018).
58. Bento, A. P. *et al*. An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics.* **12**, 1–16 (2020).

Acknowledgements

The authors thank all those who contributed to this study. This work was supported by grants from the National Natural Science Foundation of China (32470644, 82171844 to Y.-B.Z.).

Author contributions

Y.-B.Z. and B.X. conceptualized and designed the overall project. B.X., Y.Z., L.Z., H.H., X.F. and Y.-B.Z. implemented data collection, preprocessing and analysis, and also responsible for writing and revision of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05528-1.

Correspondence and requests for materials should be addressed to Y.-B.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.