Computational and Structural Biotechnology Journal 20 (2022) 6526–6533

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

The dynamic landscape of peptide activity prediction

Oriol Bárcenas [a,1], Carlos Pintado-Grima [a,1], Katarzyna Sidorczuk [b], Felix Teufel [c,d], Henrik Nielsen [e], Salvador Ventura [a,*], Michał Burdukiewicz [a,f,*]

[a] Autonomous University of Barcelona, Institute of Biotechnology and Biomedicine, Spain
[b] University of Wrocław, Faculty of Biotechnology, Poland
[c] University of Copenhagen, Copenhagen, Denmark
[d] Novo Nordisk A/S, Digital Science and Innovation, Denmark
[e] Technical University of Denmark, Denmark
[f] Medical University of Białystok, Clinical Research Centre, Poland

ARTICLE INFO

Article history:
Received 19 October 2022
Received in revised form 21 November 2022
Accepted 21 November 2022
Available online 24 November 2022

Keywords:
Peptides
Activity
Prediction
Functional peptides
Machine learning
Deep learning
Reproducibility

ABSTRACT

Peptides are known to possess a plethora of beneficial properties and activities: antimicrobial, anticancer, anti-inflammatory or the ability to cross the blood–brain barrier are only a few examples of their functional diversity. For this reason, bioinformaticians are constantly developing and upgrading models to predict their activity *in silico*, generating a steadily increasing number of available tools. Although these efforts have provided fruitful outcomes in the field, the vast and diverse amount of resources for peptide prediction can turn a simple prediction into an overwhelming searching process to find the optimal tool. This minireview aims at providing a systematic and accessible analysis of the complex ecosystem of peptide activity prediction, showcasing the variability of existing models for peptide assessment, their domain specialization and popularity. Moreover, we also assess the reproducibility of such bioinformatics tools and describe tendencies observed in their development. The list of tools is available under https://biogenies.info/peptide-prediction-list/.

Contents

[*] Corresponding authors at: Autonomous University of Barcelona, Institute of Biotechnology and Biomedicine, Spain.
E-mail addresses: salvador.ventura@uab.cat (S. Ventura), michalburdukiewicz@gmail.com (M. Burdukiewicz).
[1] Oriol Bárcenas and Carlos Pintado-Grima are joint first authors.

https://doi.org/10.1016/j.csbj.2022.11.043
2001-0370/© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

1. Introduction

In living organisms, peptides are short amino acid sequences that accomplish a wide variety of biological functions [1–4]. Peptides are expressed either as mature products [5] or cleaved from natural precursor proteins. Considering their potential, researchers have been long interested in the identification and functional characterization of peptides with relevant activities.

Several machine learning (ML) models have been developed to this end, providing new and relevant insights into the field [6]. However, considerable discrepancies in theoretical assumptions can arise. For example, peptides are generally defined as short amino acidic segments, but their minimum and maximum lengths is a matter of discussion. While some researchers have provided evidence of dipeptide self-assembly into higher-order structures [7,8], certain activities require longer peptide lengths [9]. A similar case applies to maximum lengths, with some studies considering peptides above 100 residues [10,11]. Beyond theoretical assumptions, the goal behind model generation can also create significant differences. While some tools specialize in predicting single-activity peptides such as antimicrobial [12–17], anticancer [18–21], antiparasitic [22] or antiviral [23], others intend to predict multiple functions overlapping the same peptide [24,25]. In some cases, these multiple classifications are faded by blurry definitions such as "antimicrobial peptides" (AMPs). This category is often found in the description of these tools and is sometimes inappropriately used to refer exclusively to antibacterial peptides instead of more broad AMPs (which include e.g. antifungal and antiviral peptides), a trend that can lead to unclear results [12,26].

Beyond traditional peptide functionalities, new activities are arising and expanding this field of study. The characterization and identification of peptides for therapeutic use has been a topic of interest for many researchers in the last years. In this sense, anti-aggregating peptides have emerged as promising candidates for clinical therapies. For example, amphipathic and cationic alpha-helical peptidic scaffolds have been described to bind alpha-synuclein toxic species found in Parkinson's Disease (PD) with nanomolar affinity [27,28]. Strikingly, peptides with these features often exhibit a crosstalk between anti-amyloid, antimicrobial and antibiofilm activities [29]. This opens a new window in the search for therapeutic strategies to treat and diagnose PD and other related diseases at the molecular level. The emergence of new peptide entities and applications points at developing novel tools for predicting peptide activities with high confidence.

ML models are algorithms that can find patterns in provided training data and make predictions on new and unseen datasets. Therefore, the original dataset used for training such models is of tremendous importance, as it determines the quality of the obtained results. The blurry definitions of peptides with some activities make the acquisition of high-quality positive data difficult. However, creating negative data is even more challenging as there are essentially no experimentally verified negative examples. The most common practice is generating negative data by sampling a large database for peptides that most likely do not possess the intended activities, e.g., by searching UniProt for entries not annotated with specific keywords and restricting sequence length. However, there are many ways to perform data sampling, and authors usually define their own sampling method. It leads to models that are meant to predict the same activity but possess different areas of competence. Moreover, it results in biased comparisons of the performance of such models [30].

Different strategies can be used to extract the information from unstructured data, such as peptide primary sequences. The usage of shallow models requires the conversion of unstructured sequential data into the structured, tabular format by employing some heuristic algorithm for feature engineering. Thanks to their architectural flexibility, deep models can learn from structured and unstructured data [31]. Irrespective of the method, predictive models are created to provide useful tools to the scientific community to solve specific problems. In this sense, users are interested in obtaining a wide range of high-quality predictions as quickly and efficiently as possible, providing valuable insight into their data. Furthermore, the method should be robust and stable so that users can accurately re-run computations. Web servers are usually the preferable option to fulfill these requirements, but researchers and private companies with privacy concerns may consider them unsuitable when dealing with sensitive data. Consequently, developers are also encouraged to provide standalone applications in code repositories that can be employed independently by users in local environments. Additionally, standalone tools can be easily integrated with the local analytical pipelines and thus be more appropriate for high-throughput studies.

Considering all the benefits and general appeal of predictive models, an overwhelming number of bioinformatics tools have been developed in the last years to systematically predict and classify peptides. Such tools are based on different features and ML algorithms that aim at identifying specific activities. This seemingly positive fact sometimes turns against users as they may struggle to find the best suited predictor for their precise needs. The problem is exacerbated when functional published tools drown among broken web servers that are no longer available [32] or unreported code repositories [33] that lead to serious reproducibility issues.

Although some overviews of these ML models have been published [34], there is still a need for a systematic review of the current state of the peptide prediction field. In this work, we inspect the characteristics of 140 existing ML tools for peptide activity identification and describe observed associations and tendencies. With this minireview, we expect readers to find an exhaustive and integrated resource for conducting their peptide activity prediction tasks successfully, as well as raising awareness about the reproducibility crisis of many ML models for the scientific community.

2. Methods for data acquisition

Current software tools appearing in the tool list of the online supplementary website were screened using PubMed's API and Google Scholar looking for the keywords *antimicrobial, anticancer, antifungal, antiviral, antiparasitic, antitoxin, antiangiogenic, antibiofilm, antihypertensive, antiinflammatory, cell-penetrating peptides, blood–brain barrier peptides, chemotactic, quorum sensing, surface-binding* or *neuropeptides* and *peptide prediction* in the title or the abstract of the publication until July 1st, 2022. The code provided in repositories was not tested, as it is considered to be the same that was provided during the review process. However, we did test web servers both for the availability of the hosting site and the functionality of the tool on October 14th, 2022. If web server links are no longer working, they are considered non-active. On the other hand, tools that provide an active web server but do not provide results when a model input is given are deemed as non-functional. It is important to note that this assessment of functionality of the web server depends very highly on the date and hour of the accession and it could change dynamically.

To measure the popularity of tools, we obtained their citations from CrossRef on October 14th, 2022. Due to their nature, citations are noisier measures of popularity for more recent tools. However, they still at least partially reflect the general interest of the scientific community.

The list of tools was manually curated from original publications to assess their reproducibility into golden, silver, or bronze categories as proposed by Heil et al. in 2021 [35]. We included an additional category ("below bronze") for tools that did not fulfill the minimum information required for bronze reproducibility: data, models, and source code published and downloadable. Besides these minimum requirements, if the model repository is well documented, key analysis details are suitably recorded and dependencies can be set up in a single command (e.g. via docker, conda, snakemake or renv) the tool is granted a silver category. The gold category is only given for those methods that allow the reproduction of the entire analysis with a single command (e.g. via makefile, snakemake, drake or targets). This workflow is summarized in the Supplementary Fig. 1.

3. List of existing models

The PubMed and Google Scholar screening rendered a total of 140 tools for predicting different peptide functions published from 2009 to July 2022, which are freely available for all users as an online supplementary website at https://biogenies.info/peptide-prediction-list/.

In this peptide prediction list, tools can be easily sorted according to their functional prediction as detailed in the previous section by keyword search. The overall availability of each method can be assessed by the presence of an active and functional web server, code and training repositories and general reproducibility. Therefore, each tool can be searched by name and publication DOI so that users can identify all available bioinformatics resources linked to the tool according to their needs.

To increase the accessibility of the online supplementary website, tools with available web servers or repositories can be easily accessed through the link provided in the corresponding columns. Additionally, all the information can be copied or downloaded as csv, Excel or PDF files.

4. Trends and associations in ML-based peptide predictors over time

4.1. The vast diversity of functional activities derived from peptide predictors

One of the main questions that developers ask themselves when developing a new predictor is what are the key points that differentiate average from successful and well-cited tools, beyond the journal of publication. This is important, as bioinformaticians are constantly aiming to provide the best possible resources that would help researchers in their respective work fields, offering solutions to major concerns. For example, pathologies affecting the nervous system, such as Parkinson's or Alzheimer's disease, require dedicated software to screen for peptides with the ability to cross the blood–brain barrier. Other cases include AMPs or anticancer peptides, which have been widely studied for their implication for human health and disease and the development of novel therapeutics in the last decade [36,37]. This is exemplified by the high number of tools predicting these activities (Fig. 1A). Anticancer predictors are found to be the most prevalent with 42 tools, closely followed by the antimicrobial activity with 40, which accounts for 20.5 % and 19.5 % of all annotated activities (205), respectively. Nonetheless, activities with a lot of available tools do not have the highest median number of citations per year (Fig. 1B). This could be a consequence of the high level of competition between predictors in the most popular activities. In this sense, anticancer and AMPs have a median of around 5.5 citations per year, whereas, for example, toxic, antifungal or antiviral peptides surpass them despite having much less available tools (3, 10 and 20, respectively). Thus, the available activities highlight the variability of the functional landscape of peptide prediction, which is experiencing an evident rise in popularity. It is translated into a sixfold increase of available models developed in the last five years, from 34 in 2017 to 205 in 2022 (Fig. 1C). This year-to-year model increase is also appreciated when considering non-cumulative values of tools deployed per year (Supplementary Fig. 2). The overwhelming amount of peptide tools is becoming a source of confusion and misunderstanding for many researchers, as they struggle to find the optimal resource for their specific needs.

4.2. Emergence of deep architectures

Given the revolution of deep learning (DL) in the proteomics field, it is also relevant to consider the differences between deep and shallow ML algorithms in the peptide activity prediction area. The benefit of the deep frameworks in the field of peptide property predictions is not obvious as the peptide datasets are usually more limited than protein ones, especially considering the number of available sequences [38]. Still, DL models are rapidly gaining popularity, accounting for almost half of the published models in 2021, whereas no deep models were published before 2018 (Fig. 2A). When considering the number of citations of these models, some differences in the relative amount of citations for deep models in comparison with non-deep models can be observed (Fig. 2B). To quantify this observation, we introduce the citation score (C score) and plot it against the year of publication (Fig. 2C). The C score provides an overview of the appeal of deep models, representing the relative number of citations of deep models vs non-deep models. That is, a C score of 1 means that deep models have the same relative amount of citations as non-deep models. As an example, if this value were 1.2, it would indicate that deep models are 20 % more cited than shallow models. With this score, it is clear that, except for those tools published in 2021, deep models are more appealing than non-deep models according to their relative number of citations. It highlights the popularity of deep models and provides a plausible explanation for the explosion of deep models created in 2021.

4.3. ML algorithm: what is the most common choice?

A key question when developing an ML algorithm relies on the final classifier that will be selected for the prediction. Among all ML classifiers used by the tools, shallow models represented by support vector machines (SVM) and random forests (RF) are the most frequent, with 64 and 59 independent architectures, respectively (Supplementary Fig. 3). They are closely followed by DL algorithms (51) that have emerged as a common method to build ML models either alone or in a combination of an ensemble. For the sake of completeness, all independent ML classifiers from ensembles were included. Both RF and SVM, like all other shallow ML algorithms, work with structured data. They are also quite popular in the ML field for being easy to train and tune, but they require converting unstructured sequential data to a tabular format using feature engineering. SVMs can perform both linear and non-linear classification and regression but are often difficult to scale to large datasets. In contrast, RF methods are less appropriate for regression but are able to learn how important each feature is to the prediction [39].

Although individual ML algorithms have been described to be highly accurate and precise in their predictions, some tools use the combination of different models to create an ensemble that can potentially improve the prediction capabilities of individual models. This includes deep and non-deep combinations that claim

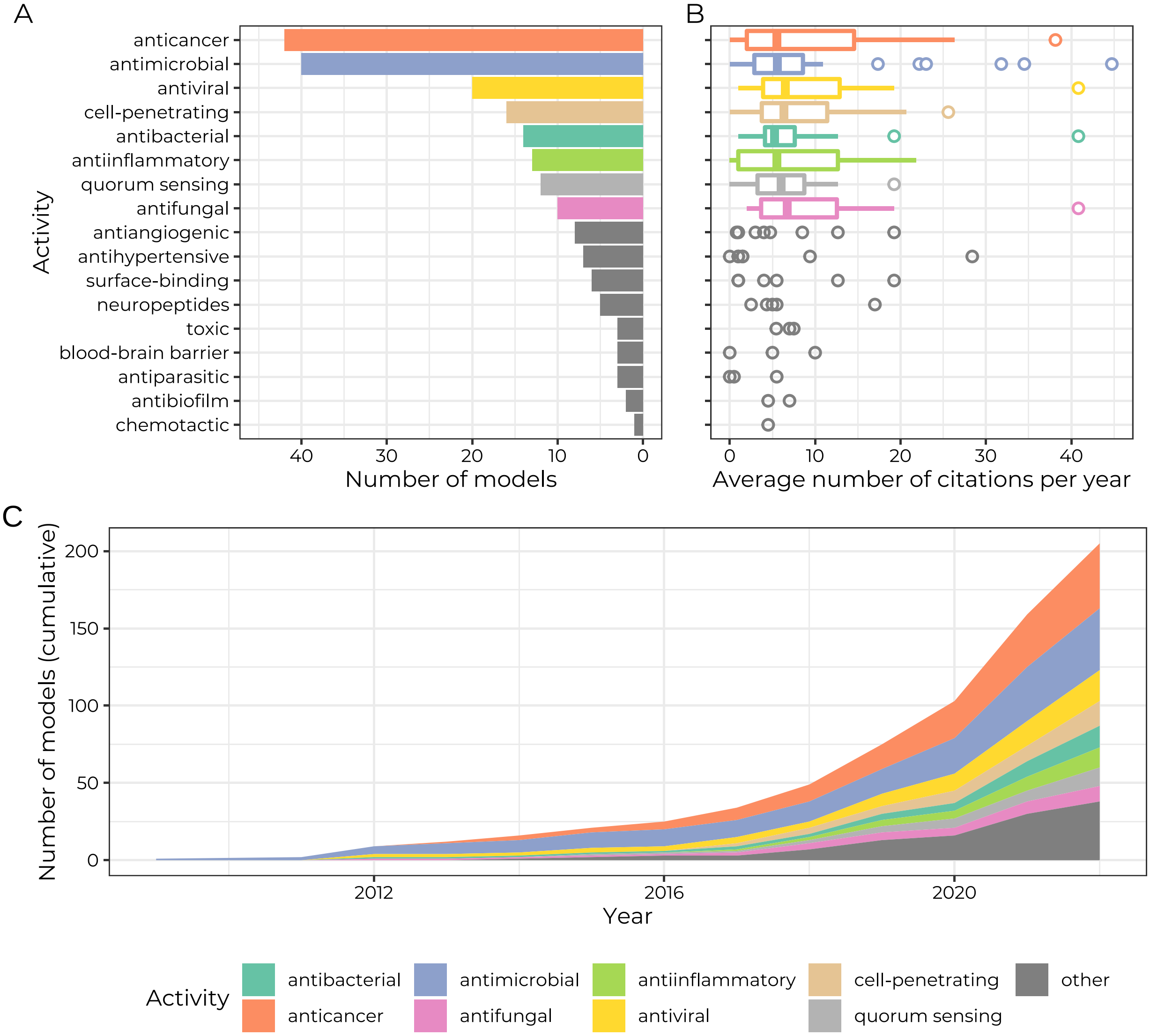

Fig. 1. Peptide prediction landscape by predictive activity, according to the number of publications per activity (A), the average number of citations per year (B) and the cumulative number of tools (C). In this plot, activities are counted individually. That is, if a tool has more than one predictive activity, each one is counted independently. Those activities with 10 or more published tools are presented in color (antibacterial, antimicrobial, antiinflammatory, cell-penetrating, anticancer, antifungal and antiviral). In these cases, the average number of citations is displayed as a boxplot to better show the distribution of the average citation differences between activities. These results indicate that the activities with the highest amount of models do not necessarily have the highest average citations.

to be a viable strategy for obtaining more reliable classifications, for example by using features obtained by a DL algorithm to train a shallow model [40].

4.4. Bioinformatics resources: web servers, code repositories and reproducibility issues

Once the purpose of the tool is defined and the corresponding method has been validated, developers need to think of suitable platforms to allow users to run predictions. Web servers are usually the priority option because they provide an easy framework to run calculations without the necessity of technical expertise in programming or command-line management.

Aside from practical usability aspects, the tool should be replicable given the source code and the data are provided. Interestingly, it is only recently that the authors have put more emphasis on the reproducibility of their tools. Indeed, the first tool with at least a bronze level of reproducibility was published in 2018 (Supplementary Fig. 4). In an effort to balance this bias, which might be driven by the recent assessment of reproducibility standards for machine learning in the life sciences [35], here we only analyze the reproducibility status of published tools from 2018 onwards (111 out of the 140 tools). We observed that tools that do not incorporate open web servers (57 out of 111) tend to have a lower number of citations than those that include them (Fig. 3A). Nonetheless, having an active web server at the time of publication is not enough. Published web servers tend to go offline due to the associated maintenance costs and technical management required. Among all 57 peptide tools with associated web servers, 24 were no longer accessible or functional (42.1 %) on October 15th 2022. It means that these tools have lost potential interactions with users that could be interested in running predictions using their algorithms. This is also reflected in the total number of citations, where tools with non-active web servers obtain fewer citations than active predictors.

The most concerning observation, however, is related to the reproducibility (or lack thereof) of ML methods (Fig. 3B). Of the 111 tools, only 38 (34.2 %) contained the minimum information

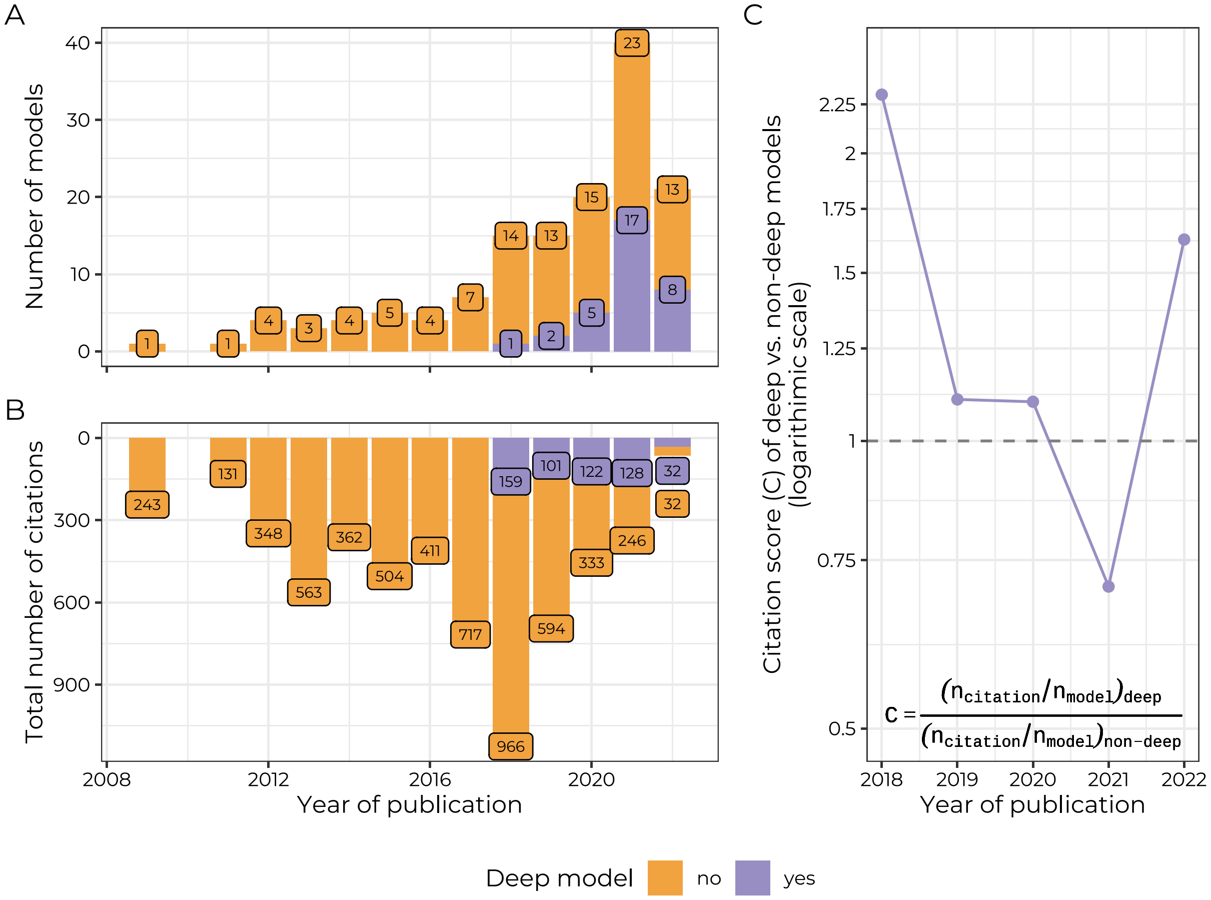

Fig. 2. Comparison between deep and non-deep models, according to the number of published models (A) and the number of citations (B). The citation score for the deep models (C) indicates for which years deep models were more cited than non-deep models (scores over 1). This score provides an accurate perspective on the attractiveness of DL against shallow tools.

required to reproduce the ML algorithm independently: source code, data, and models published and downloadable. Among them, nine gold and four silver tools presented the highest possible standard of reproducibility (Table 1). The rest fell into the "below bronze" category where, in the best-case scenario, only web servers were provided to allow users to run predictions. Something remarkable, however, is that web server availability is a much better indicator of the average number of citations per year than the reproducibility standard of these tools (Fig. 3C). This situation may be the root cause of the ML reproducibility crisis, as making a reproducible model is time-consuming and does not render an apparent citation-wise benefit. Given the existing lag between the date of publication and when tools start receiving their first citations, it is also possible that the average number of citations per year for very recent tools is underestimated compared to longer-lived tools.

5. Discussion

Peptide property prediction has positioned itself as a key field of activity within bioinformatics in recent years. The ability of peptides to mediate different types of functional activities and interactions with other biomolecules offers innovative candidates to target specific pathologies. At the molecular level, peptide therapy emerges as a promising strategy to specifically target the pathological agents underlying these diseases [28,41,42]. Early predictors used architectures different from ML, but the difficulty in identifying the contribution weight of each feature to the final decision limited their performance. Thanks to recent advances in ML and bioinformatics, an increasing number of tools based on ML have been released, as it is becoming the standardized method to generate predictive algorithms.

The success of ML methods in predicting peptide properties has led to the development of a vast amount of tools that intend to be useful resources for the scientific community. In this sense, researchers generally prefer the development of user-friendly web servers that are easy to use rather than complex code repositories that only people with some bioinformatics and programming background can run. Based on our analysis, it seems that providing open-source code does not render any benefit citation-wise, and as a consequence, researchers may opt to keep their original code for future research without losing significant impact. We believe this is the origin of the reproducibility crisis we have described in this minireview since only 27 % of all tools (38 out of 140) present the minimum information required to reproduce the ML algorithm from scratch.

A similar tendency applies to DL algorithms, which have become the standard "advanced" models for peptide property prediction for some researchers. The ability of these algorithms to learn relevant features directly from the data led many experts in data science and related disciplines to translate this knowledge into the peptide prediction field. Although DL was successfully applied to identify AMPs in metagenomics data [43], it is not clear whether the success DL has seen in other fields of science has yet

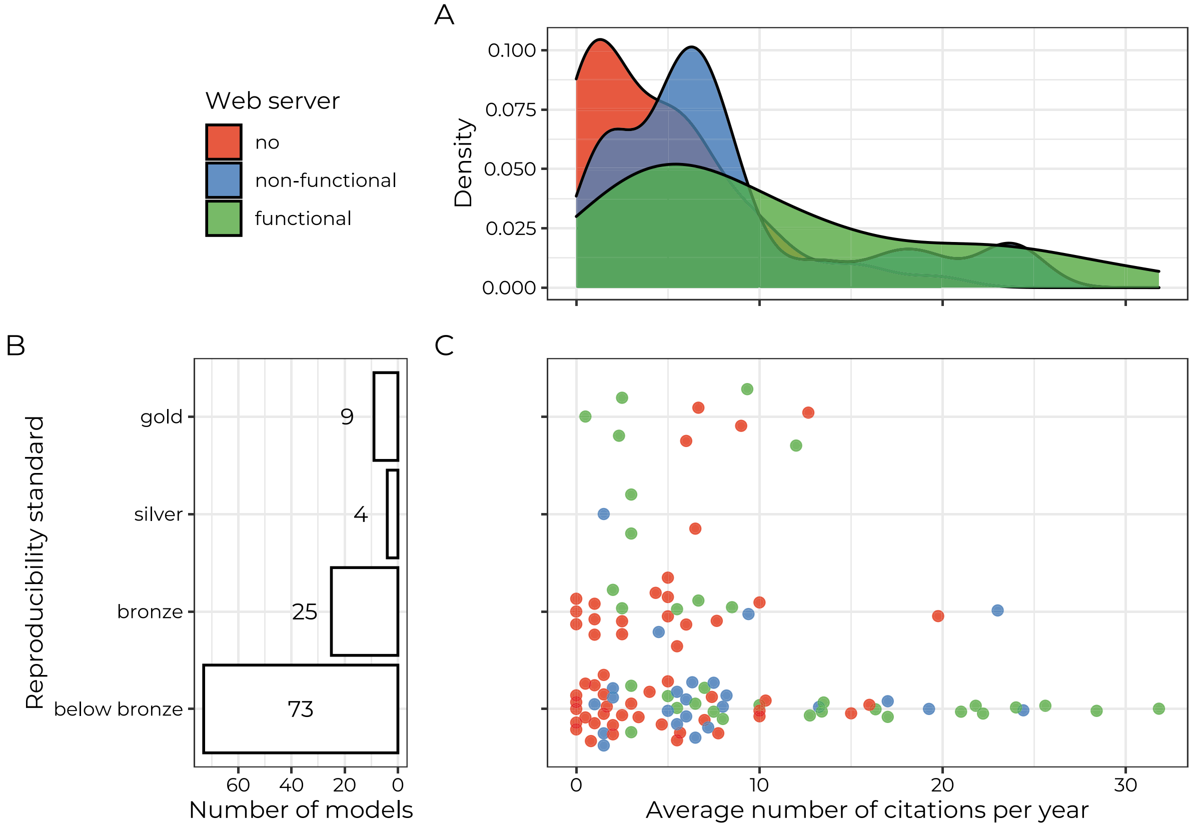

Fig. 3. Comparison of tools published from 2018 onwards according to the average number of citations per year, web server availability and their reproducibility standard. (A) The density chart of models according to the number of citations. Density colors represent the availability of web servers. (B) Bar chart representing the number of models fulfilling the given reproducibility standard. (C) The beeswarm chart of models. The y-axis represents the reproducibility standard and x-axis represents the average number of citations per year. The color of the point represents the availability of the web server.

Table 1
Tools granted with the gold or silver standard of reproducibility classification as adapted from Heil et al. [35]. Most of them are designed to predict anticancer or antimicrobial activities, which account for the majority of peptide tools as observed in Fig. 1.

Tool name	Activity prediction	Reference
ACP-MHCNN	Anticancer	[20]
CancerGram	Anticancer	[18]
DeepACP	Anticancer	[19]
iACP-FSCM	Anticancer	[21]
amPEPpy 1.0	Antimicrobial	[16]
AmpGram	Antimicrobial	[14]
AMPlify	Antimicrobial	[17]
PredAPP	Antiparasitic	[22]
PPTPP	Antiangiogenic; Antibacterial; Anticancer; Antiinflamatory; Antiviral; Cell-penetrating; Quorum sensing; Surface-binding	[24]
Ampir	Antimicrobial	[15]
Macrel	Antimicrobial	[13]
MLBP	Anticancer; Antihypertensive; Antiinflamatory; Antimicrobial	[25]
PreAntiCoV	Antiviral	[23]

been translated to peptide property prediction. Indeed, some studies suggest that DL models so far have failed to improve performance over shallow models for AMP identification [38].

An influx of new tools for developing deep and shallow models significantly reduced the entry barrier for peptide activity prediction. Still, these tools can not replace researchers in assessing the correctness of their models. Here, biological domain knowledge helps in the design of robust benchmarks for a fair evaluation of the performance of proposed models. For example, it was reported that some AMP predictors yield higher probabilities of antimicrobial activity for long peptides which makes little sense from the biological point of view [10]. As the majority of the effort in the field is directed toward proposing new tools and not evaluating existing ones, we are still unaware of the similar shortcomings of available solutions.

Altogether, we argue that there is a need for reorganization and clarification in the peptide property prediction field. The redundancy of specific tools and the difficulty of finding optimal bioinformatics resources keep adding complexity to the standard use of these methods. To at least partially alleviate this issue, we enhance our minireview with an online supplementary web resource that collects all available peptidic tools to scrutinize accessible models and help users find the best for their needs. We aim to raise awareness among researchers and developers about the reproducibility crisis observed in peptide property prediction and encourage them to work together to promote open science and cooperation between researchers.

Funding

C.P.-G was supported by the Secretariat of Universities and Research of the Catalan Government and the European Social Fund (2021 FI_B 00087). F.T. was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). K.S. was partially supported by the National Science Centre grant (2018/31/N/

NZ2/01338). S.V. was funded by the Spanish Ministry of Science and Innovation (PID2019-105017RB-I00) and by ICREA, ICREA-Academia 2020. M. B. was supported by the Maria Zambrano grant funded by the European Union-NextGenerationEU.

CRediT authorship contribution statement

Oriol Bárcenas: Data curation, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Carlos Pintado-Grima:** Data curation, Investigation, Validation, Writing – original draft, Writing – review & editing. **Katarzyna Sidorczuk:** Data curation, Investigation, Methodology, Software, Writing – review & editing. **Felix Teufel:** Writing – review & editing. **Henrik Nielsen:** Writing – review & editing. **Salvador Ventura:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Michał Burdukiewicz:** Conceptualization, Data curation, Methodology, Software, Visualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors want to thank Dominik Rafacz (Warsaw University of Technology) for his help in the online supplementary website development and the anonymous reviewers for their valuable comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.043.

References

[1] Ghosh C, Sarkar P, Issa R, Haldar J. Alternatives to conventional antibiotics in the era of antimicrobial resistance. Trends Microbiol 2019;27:323–38. https://doi.org/10.1016/j.tim.2018.12.010.

[2] Zhou X, Smith QR, Liu X. Brain penetrating peptides and peptide–drug conjugates to overcome the blood–brain barrier and target CNS diseases. WIREs Nanomed Nanobiotechnol 2021;13:e1695. https://doi.org/10.1002/wnan.1695.

[3] Habault J, Poyet J-L. Recent advances in cell penetrating peptide-based anticancer therapies. Molecules 2019;24:927. https://doi.org/10.3390/molecules24050927.

[4] Vandergriff A, Huang K, Shen D, Hu S, Hensley MT, Caranasos TG, et al. Targeting regenerative exosomes to myocardial infarction using cardiac homing peptide. Theranostics 2018;8:1869–78. https://doi.org/10.7150/thno.20524.

[5] Saghatelian A, Couso JP. Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol 2015;11:909–16. https://doi.org/10.1038/nchembio.1964.

[6] 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 2020;40:1276–314. https://doi.org/10.1002/med.21658.

[7] de Groot NS, Parella T, Aviles FX, Vendrell J, Ventura S. Ile-Phe dipeptide self-assembly: clues to amyloid formation. Biophys J 2007;92:1732–41. https://doi.org/10.1529/biophysj.106.096677.

[8] Gnanasekaran K, Korpanty J, Berger O, Hampu N, Halperin-Sternfeld M, Cohen-Gerassi D, et al. Dipeptide nanostructure assembly and dynamics via in situ liquid-phase electron microscopy. ACS Nano 2021;15:16542–51. https://doi.org/10.1021/acsnano.1c06130.

[9] Clark S, Jowitt TA, Harris LK, Knight CG, Dobson CB. The lexicon of antimicrobial peptides: a complete set of arginine and tryptophan sequences. Commun Biol 2021;4:1–14. https://doi.org/10.1038/s42003-021-02137-7.

[10] Gabere MN, Noble WS. Empirical comparison of web-based antimicrobial peptide prediction tools. Bioinformatics 2017;33:1921–9. https://doi.org/10.1093/bioinformatics/btx081.

[11] Shi G, Kang X, Dong F, Liu Y, Zhu N, Hu Y, et al. DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res 2022;50:D488–96. https://doi.org/10.1093/nar/gkab651.

[12] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018;34:2740–7. https://doi.org/10.1093/bioinformatics/bty179.

[13] Santos-Júnior CD, Pan S, Zhao X-M, Coelho LP. Macrel: antimicrobial peptide screening in genomes and metagenomes. PeerJ 2020;8:e10555. https://doi.org/10.7717/peerj.10555.

[14] Burdukiewicz M, Sidorczuk K, Rafacz D, Pietluch F, Chilimoniuk J, Rödiger S, et al. Proteomic screening for prediction and design of antimicrobial peptides with AmpGram. Int J Mol Sci 2020;21:4310. https://doi.org/10.3390/ijms21124310.

[15] Fingerhut LCHW, Miller DJ, Strugnell JM, Daly NL, Cooke IR. ampir: an R package for fast genome-wide prediction of antimicrobial peptides. Bioinformatics 2020;36:5262–3. https://doi.org/10.1093/bioinformatics/btaa653.

[16] Lawrence TJ, Carper DL, Spangler MK, Carrell AA, Rush TA, Minter SJ, et al. amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btaa917.

[17] Li C, Sutherland D, Hammond SA, Yang C, Taho F, Bergman L, et al. AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. BioRxiv 2020:155705. https://doi.org/10.1101/2020.06.16.155705.

[18] Burdukiewicz M, Sidorczuk K, Rafacz D, Pietluch F, Bąkała M, Słowik J, et al. CancerGram: an effective classifier for differentiating anticancer from antimicrobial peptides. Pharmaceutics 2020;12:1045. https://doi.org/10.3390/pharmaceutics12111045.

[19] Yu L, Jing R, Liu F, Luo J, Li Y. DeepACP: A novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. Mol Ther - Nucleic Acids 2020;22:862–70. https://doi.org/10.1016/j.omtn.2020.10.005.

[20] Ahmed S, Muhammod R, Khan ZH, Adilina S, Sharma A, Shatabda S, et al. ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. Sci Rep 2021;11:23676. https://doi.org/10.1038/s41598-021-02703-3.

[21] Charoenkwan P, Chiangjong W, Lee VS, Nantasenamat C, Hasan MM, Shoombuatong W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci Rep 2021;11:3017. https://doi.org/10.1038/s41598-021-82513-9.

[22] Zhang W, Xia E, Dai R, Tang W, Bin Y, Xia J. PredAPP: predicting anti-parasitic peptides with undersampling and ensemble approaches. Interdiscip Sci Comput Life Sci 2022;14:258–68. https://doi.org/10.1007/s12539-021-00484-x.

[23] Pang Y, Wang Z, Jhong J-H, Lee T-Y. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. Brief Bioinform 2021;22:1085–95. https://doi.org/10.1093/bib/bbaa423.

[24] Zhang YP, Zou Q. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics 2020;36:3982–7. https://doi.org/10.1093/bioinformatics/btaa275.

[25] Tang W, Dai R, Yan W, Zhang W, Bin Y, Xia E, et al. Identifying multi-functional bioactive peptide functions using multi-label deep learning. Brief Bioinform 2022;23:bbab414. https://doi.org/10.1093/bib/bbab414.

[26] Lin W, Xu D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. Bioinformatics 2016;32:3745–52. https://doi.org/10.1093/bioinformatics/btw560.

[27] Santos J, Gracia P, Navarro S, Peña-Díaz S, Pujols J, Cremades N, et al. α-Helical peptidic scaffolds to target α-synuclein toxic species with nanomolar affinity. Nat Commun 2021;12:3752. https://doi.org/10.1038/s41467-021-24039-2.

[28] Santos J, Pallarès I, Ventura S. Is a cure for Parkinson's disease hiding inside us? Trends Biochem Sci 2022;47:641–4. https://doi.org/10.1016/j.tibs.2022.02.001.

[29] Santos J, Ventura S, Pallarès I. LL-37 and CsgC exemplify the crosstalk between anti-amyloid, antimicrobial, and anti-biofilm protein activities. Neural Regen Res 2023;18:1027–8. https://doi.org/10.4103/1673-5374.355757.

[30] Sidorczuk K, Gagat P, Pietluch F, Kała J, Rafacz D, Bąkała L, et al. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. Brief Bioinform 2022;23:bbac343. https://doi.org/10.1093/bib/bbac343.

[31] Jurtz VI, Johansen AR, Nielsen M, Armenteros A, Juan J, Nielsen H, et al. An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics 2017;33:3685–90. https://doi.org/10.1093/bioinformatics/btx531.

[32] Kern F, Fehlmann T, Keller A. On the lifetime of bioinformatics web services. Nucleic Acids Res 2020;48:12523–33. https://doi.org/10.1093/nar/gkaa1125.

[33] Papin JA, Gabhann FM, Sauro HM, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. PLOS Comput Biol 2020;16:e1007881. https://doi.org/10.1371/journal.pcbi.1007881.

[34] Wang G, Vaisman II, van Hoek ML. Machine learning prediction of antimicrobial peptides. Methods Mol Biol Clifton NJ 2022;2405:1–37. https://doi.org/10.1007/978-1-0716-1855-4_1.

[35] Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods 2021;18:1132–5. https://doi.org/10.1038/s41592-021-01256-7.

[36] Gaspar D, Veiga AS, Castanho MARB. From antimicrobial to anticancer peptides. A review. Front Microbiol 2013:4. https://doi.org/10.3389/fmicb.2013.00294.

[37] Deslouches B, Di YP. Antimicrobial peptides with selective antitumor mechanisms: prospect for anticancer applications. Oncotarget 2017;8:46635–51. https://doi.org/10.18632/oncotarget.16743.

[38] García-Jacas CR, Pinacho-Castellanos SA, García-González LA, Brizuela CA. Do deep learning models make a difference in the identification of antimicrobial peptides? Brief Bioinform 2022;23:bbac094. https://doi.org/10.1093/bib/bbac094.

[39] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23:40–55. https://doi.org/10.1038/s41580-021-00407-0.

[40] Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar SR. AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. Brief Bioinform 2021;22:bbab242. https://doi.org/10.1093/bib/bbab242.

[41] Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. Chem Biol Drug Des 2013;81:136–47. https://doi.org/10.1111/cbdd.12055.

[42] Xiao Y-F, Jie M-M, Li B-S, Hu C-J, Xie R, Tang B, et al. Peptide-based treatment: A promising cancer therapy. J Immunol Res 2015;2015:e761820. https://doi.org/10.1155/2015/761820.

[43] Ma Y, Guo Z, Xia B, Zhang Y, Liu X, Yu Y, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat Biotechnol 2022;40:921–31. https://doi.org/10.1038/s41587-022-01226-0.