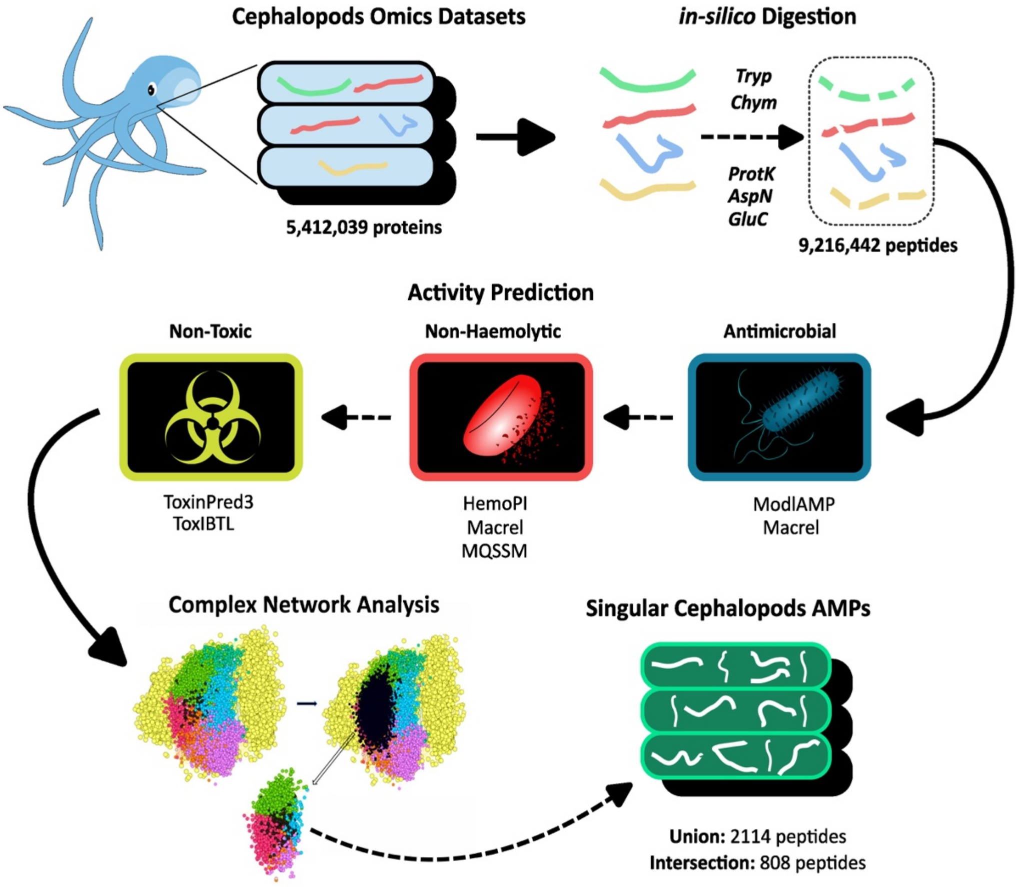
Unveiling Encrypted Antimicrobial Peptides from Cephalopods' Salivary Glands: A Proteolysis-Driven Virtual Approach

Guillermin Agüero-Chapin[1,2†*], Dany Domínguez-Pérez[3,1†], Yovani Marrero-Ponce[4,5*] Kevin Castillo-Mendieta[6], and Agostinho Antunes[1,2]

[1] CIIMAR – Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208, Porto, Portugal.
[2] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007, Porto, Portugal.
[3] PagBiOmicS - Personalised Academic Guidance and Biodiscovery-integrated OMICs Solutions. 4200-603, Porto, Portugal.
[4] Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas; and Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles y vía Interoceánica, Quito, 170157, Pichincha, Ecuador
[5] Departamento de Ciencias de la Computación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, Baja California, México.
[6] School of Biological Sciences and Engineering, Yachay Tech University, Hda. San José s/n y Proyecto Yachay, Urcuquí 100119, Ecuador

* Correspondence to: G. Agüero-Chapin (gchapin@ciimar.up.pt) & Y. Marrero-Ponce (ymarrero@usfq.edu.ec)
† These authors contributed equally to this work



Abstract: Antimicrobial peptides (AMPs), with their versatile actions, offer promise against antimicrobial resistance and as templates for novel therapeutic agents. While existing AMP databases primarily feature AMPs from terrestrial eukaryotes, marine sources are gaining attention, with cephalopods emerging as a promising but still underexplored source. This study unveils the potential reservoir of AMPs encrypted within the proteome of cephalopods' salivary glands using *in silico* proteolysis. A composite protein database comprising canonical and non-canonical proteins from cephalopods' salivary apparatus was used as the substrate for five proteases involved in three digestion protocols. The resulting millions of peptides were screened using machine learning, deep learning, multi-query similarity-based models, and complex networks. The screening prioritizes antimicrobial activity, the absence of haemolytic and toxic attributes, and structural distinctiveness compared to characterized AMPs. Diverse publicly accessible AMP datasets are produced, catering to various research needs, ranging from those focused solely on antimicrobial activity to refined datasets of non-haemolytic and non-toxic AMPs. Comparative analyses and network science principles were applied to identify singular and representative subsets from non-haemolytic and non-toxic AMPs. All these sets of AMPs and the proposed mining tools serve as valuable assets for peptide drug developers.

Keywords: Cephalopods, salivary glands, omics data, virtual screening, complex networks, AMPs datasets.

1. Introduction

Antimicrobial resistance (AMR) poses a significant global public health threat, prompting the urgent need for novel antimicrobial agents. The diminishing effectiveness of conventional antibiotics against a wide range of resistant pathogens has driven the search for alternative solutions.[1] Antimicrobial peptides (AMPs) have emerged as promising candidates to address this crisis, offering versatile antimicrobial activities and diverse modes of action. Their therapeutic potential extends beyond the development of new antibiotics to combat multidrug-resistant bacteria.[2, 3] AMPs also hold promise in the creation of agents with antitumoral, antiviral, antifungal, and other therapeutic properties.[4]

To fully harness the potential of AMPs, extensive efforts have been made to compile and organize AMP-related information into specialized databases. Notable among these are databases like APD (Antimicrobial Peptide Database)[5], CAMP (Collection of Antimicrobial Peptides)[6], and DBAASP (Database of Antimicrobial Activity and Structure of Peptides)[7], which have been continuously updated. In addition to these, the StarPep database (StarPepDB) stands out as one of the most comprehensive curated repositories of AMPs, integrating unique entries from 42 AMP databases with their metadata.[8] These databases facilitate the study of AMP sequences, structures, activities, and other relevant information, significantly enhancing their potential translation into therapeutic interventions.

The origins distribution of AMPs has also been facilitated by databases. Most AMPs reported to date stem from eukaryotic origins, notably plants, animals, and fungi.[9] Antimicrobial properties have been attributed to various bodily fluids since 1885, including blood, sweat, saliva, plasma, white blood cell secretions, and granule extracts.[10] Historically, terrestrial eukaryotes have been a primary source of AMPs. However, more recently, marine organisms, particularly invertebrates, have gained prominence due to their robust and effective innate immune systems, enabling their survival for over 450 million years in diverse ecological niches.[11, 12] The immense ecological diversity of marine environments provides a promising landscape for the discovery of AMPs with unique structures and potent antimicrobial activities. Notably, AMPs from marine invertebrates constitute a significant proportion, approximately 67% of all marine AMPs (statistics as of December 2022).[12]

Marine invertebrates, including shrimp, oysters, and horseshoe crabs, are known to consistently express AMPs.[13, 14] For instance, horseshoe crabs produce highly effective AMPs like tachyplesin and polyphemusin, exhibiting antibacterial and antifungal properties at low micromolar levels.[11] Notably, polyphemusin, similar to several other AMPs, demonstrates antiviral activity against human immunodeficiency virus (HIV).[15] More recently, the exploration of marine invertebrates has expanded through omics techniques, offering greater sensitivity in detecting the presence of AMPs.[16] In this context, our research group identified AMPs within the ascidian's tunic and the salivary glands of *Octopus vulgaris* through shotgun proteomics analyses[17, 18], and more recently others found AMPs common within octopus skin mucus proteome.[19] The comprehensive discovery of AMPs in *O. vulgaris* became feasible through the application of an optimized methodological workflow and the utilization of a composite protein database constructed from proteomic and transcriptomic data of the cephalopods' salivary apparatus.[18, 20, 21] Cephalopods, known for their efficient predatory tactics involving a diverse array of substances, predominantly cephalotoxins and neurotoxins to immobilize prey[22], possess omics data characterizing their salivary apparatus that holds the potential for venom-related proteins, toxins, and AMPs, as substantiated in previous research.[18, 21] However, AMPs with encrypted sequences within longer transcripts or proteins, exemplified by cases such as histones[23], those not constitutively expressed, or potentially disregarded by the computational omics workflow (e.g., small-size transcripts or protein fragments less than 100 amino acids)[16] can be unveiled by a comprehensive examination of a composite protein database sourced from the cephalopods' salivary apparatus.[20] This composite protein database was purposefully built for a proteome-wide AMPs discovery by including "non-canonical" proteins, exploring all the ORFs from cephalopods' salivary glands transcriptomes, and proteins shorter than the TransDecoder default minimum protein length threshold of 100 amino acids.[20]

In this context, this study focused on a privileged marine source represented by the cephalopods' salivary glands, where a potentially abundant reservoir of hidden AMPs is believed to exist. Our approach to unveil these cryptic AMPs involves the *in-silico* proteolysis of the composite protein database originating from cephalopods' salivary apparatus. This enzymatic digestion is performed using proteases commonly employed in proteomics. Subsequently, the resulting peptide libraries, comprising millions of peptides, were subjected to *in-silico* screening. During this screening, we consider essential AMP characteristics relevant for drug development and pay particular attention to their structural distinctiveness within the chemical space.

The resulting mining workflow yields various AMP datasets, catering to a spectrum of research needs. These datasets range from those solely focusing on antimicrobial activity to a refined, distinct dataset consisting of non-haemolytic AMPs devoid of toxic attributes. These datasets are publicly accessible and offer valuable resources for peptide drug developers, adaptable to their specific requirements.

2. Datasets and Methods

2.1. Omics data as a substrate for in silico proteolysis

A version of the composite protein database, comprising various omics datasets sourced from the cephalopods' salivary apparatus, as reported in Ref[20], served as the substrate for the in silico proteolysis. This composite database includes five distinct datasets originally labelled and referenced as follows:

- *Database A* — 19,087 proteins derived from proteogenomic analyses of the *O. vulgaris* salivary apparatus, as reported by Fingerhut *et al.* (2018).[21]
- *Database C* — 2,427 proteins corresponding to the post-salivary glands (PSGs) of three *O. vulgaris* specimens, as detailed by Almeida *et al.* (2020).[18]
- *Database D* — 84,778 proteins identified through 16 publicly-available transcriptomes from cephalopods' PSGs by TransDecoder.[18, 20]
- *Database E* — 5,106,635 six-frame translated proteins shorter than the TransDecoder default minimum protein length threshold of 100 amino acids, which were not included in Database D.[18, 20]
- *Database F* — 720,910 six-frame translated proteins extracted from the open reading frames (ORFs) from *O. vulgaris* PSGs transcriptomes that were not part of Database A.[20, 21]

Database B was not considered for proteolysis because it contained characterized AMPs from the StarPepDB.[8] Putative duplicates in each database were removed, and then the databases were fused into a composite protein database, followed by a redundancy removal process with the cd-hit tool at 0.98 sequence identity (https://github.com/weizhongli/cdhit).[24] The seqkit tool (https://bioinf.shenwei.me/seqkit/) was used to assist in both duplicates removal and finding common sequences between two databases[25], which allowed for an *all-vs-all* comparison among databases. The Jaccard index was used as a pairwise similarity metric.[26]

2.2. In silico proteolysis and peptidomes characterization

Five main proteases commonly used in proteomics: trypsin, chymotrypsin, proteinase K, AspN, and GluC were applied.[27] Peptidomes were generated using 13 distinct proteolysis protocols involving the action of one enzyme (OE) or two enzymes (TE), which could be applied in a sequential (S) or concurrent (C) mode. We performed the in silico enzymatic digestion using the Rapid Peptides Generator (RPG) tool (https://rapid-peptide-generator.readthedocs.io/en/latest/index.html).[28] The previously-mentioned proteases were involved in the three digestion protocols:

One Enzyme	Two Enzyme Sequential Mode	Two Enzyme Concurrent Mode
Tryp	Tryp-Chym	Tryp-Chym
Chym	Tryp-Proteinase-K	Tryp-Proteinase-K
Proteinase-K	Tryp-GluC	Tryp-GluC
GluC	Tryp-AspN	Tryp-AspN
AspN		

The peptide libraries or peptidomes resulting from each proteolysis protocol were filtered following these steps: (i) retaining only peptides that were 6-40 AAs in length, (ii) removing duplicates, (iii) removing peptides sharing above 0.98 of sequence identity, (iv) leaving out peptides with non-standard amino acids. The seqkit and cd-hit tools were used to perform this pre-screening. Then, each peptide library was characterized based on its global peptide features, such as sequence length, AA frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment. The PDAUG package (https://github.com/jaidevjoshi83/pdaug) was used to calculate these features.[29]

2.3. Antimicrobial and toxicity screening

Each resulting peptide library is subsequently screened for promising AMPs, which had been revealed by the proteolysis step. To ensure accurate detections, we determined the final prediction output by consensus agreement of three prediction models/tools. The screening of the 13 peptidomes started by the prediction of the antimicrobial activity using one model implemented in Macrel: (Meta)genomic AMP Classification and Retrieval[30] and other two from modlAMP[31]. The subcommand "macrel peptides" were used to run marcel on peptide libraries (https://github.com/BigDataBiology/macrel) while modlAMP used the data "AMPvsUniProt" for training its two implemented machine learning (ML)-based classifiers: modlAMP_Random Forest (RF) and modlAMP_Support Vector Machine (SVM) (https://modlamp.org).

Subsequently, the toxicity which is the most undesired property of AMPs for drug development, was assessed by the prediction of their haemolytic potential and their content of toxic signatures. The haemolysis prediction was also performed by macrel[30], HemoPI[32], and by a multi-query similarity searching model (MQSSM) developed in Ref.[33] Since macrel output also provides haemolytic predictions for detected AMPs, "macrel peptides" were run as before (https://github.com/BigDataBiology/macrel). The standalone version of HemoPI was used (https://webs.iiitd.edu.in/raghava/hemopi/standalone.php), particularly its virtual screening option where the hybrid model is selected. The hybrid model considers the integration of motif- and SVM-based predictions. The MQSSM **I1**, the best model reported in Ref.[33], was constructed using the half-space proximal network (HSPN) projecting the chemical space of 2,004 haemolytic peptides from StarPepDB. The HSPN was constructed without similarity cutoff, and the angular separation was used as a pairwise similarity metric. Subsequently, a representative haemolytic subset was extracted from the HSPN using the following parameters: hub-bridge centrality, global alignment and a similarity cutoff of 0.8. This representative subset was further improved as described in Ref.[33], and it was finally used to build a MQSSM model using global alignment and a similarity cutoff of 0.40.

The detection of toxic signatures was performed by two models from ToxinPred3[34] and by ToxIBTL.[35] ToxinPred3 has implemented two prediction model types, a ML-based classifier trained with compositional features of peptides and the other a hybrid model combining two or more models including motif- and ML-based predictions (https://github.com/raghavagps/toxinpred3). ToxIBTL is a deep learning approach based on the integration of evolutionary information and physicochemical properties of peptides into the information bottleneck principle, and transfer learning to predict the toxicity of peptides (https://server.wei-group.net/ToxIBTL/Server.html). Venn diagrams were used for identifying consensus predictions among the outputs of the three prediction models.

In summary, the 13 proteolysis protocols rendered the following datasets: (i) peptide libraries (peptidomes), (ii) AMP consensus, (iii) non-haemolytic AMPs, (iv) non-haemolytic/non-toxic AMPs. Peptide subsets corresponding to the 13 digestion protocols within each of the four datasets were concatenated and sequence redundancy was removed at 0.98 of identity with cd-hit.

2.4. Selection of cephalopods singular AMPs

A comparison of the non-redundant non-haemolytic and non-toxic AMPs to StarPepDB[8], one of the most comprehensively reported peptide databases, was performed using the cd-hit-2d tool at 0.40, 0.50, 0.60, 0.70, 0.80, and 0.90 identity cutoffs. This was done to identify new peptide representations encoded in the cephalopods' proteome that differ from the previously reported peptides. Cephalopod singular peptides (CSPs) are considered those sharing sequence identities below the 0.40 threshold with StarPepDB members, while peptides with an equal or higher threshold were considered similar. Prior to the comparison, the StarPepDB's original space of 45,120 peptides was reduced to 32,863 by applying the cd-hit tool at 0.98 identity and retaining only peptides that ranged from 5 to 100 AAs in length and contained standard AAs.

Validating the singularity of Cephalopods' AMPs using complex networks: To validate the no relatedness of CSPs with respect to the known chemical space from StarPepDB[8], both chemical spaces were represented as HSPNs.[36] The non-redundant non-haemolytic and non-toxic AMPs from cephalopods were divided into a set of CSPs (identity < 0.40 with StarPepDB space) and a more-closely related set to StarPepDB (identity > 0.40). These sets were then used together with the 32,863 peptides from StarPepDB to build HSPNs using the StarPep Toolbox.[37] Each peptide/node was represented by an optimized set of molecular descriptors, and the Euclidean distance metric with min-max normalization were applied to determine the pairwise similarity relationships among them. AMPs within the HSPNs were clustered using the modularity optimization algorithm based on the Louvain method.[38] Peptides sharing similar features are grouped together, thus occupying the same chemical space in the network.

2.5. Representativeness from Cephalopods singular AMPs by complex networks

To further reduce the CSPs at selecting the most representative ones, centrality analyses were performed. A HSPN was constructed using only the CSPs (identity < 0.40 with StarPepDB space). The HSPN construction followed the same procedure described above, but a similarity cutoff of 0.75 was applied to improve network topology for mining information. Clusters, also known as communities, were identified using the Louvain method[38], and then two centrality measures were calculated: hub-bridge centrality (HB)[39] and harmonic centrality (HC).[40] Centrality values measure the importance of a node in a network. Additionally, pairwise similarity comparisons were performed using the Smith-Waterman method.[41] Using the peptide's centrality and a sequence similarity cutoff of 0.30, the least redundant yet most important peptides in the network were identified. This process is described in Ref.[36] Afterwards, two datasets were recovered: the union and the intersection of the sets recovered using both HB and HC centralities.

2.6. Computer Resources

The *in silico* proteolysis of 5,412,039 proteins and the subsequent screening of resulting peptidomes were managed using a high-performance desktop computer with the following specifications: CPU: Dual 20-core Intel Xeon Gold 6148 processors with (min/max) speed 1010/1000/3700 MHz, RAM: 256 GB, SSD: NVMe KINGSTON SNV2S/2000G (2 TB - M.2 - 3500 MB/s), Operating System: Linux kernel 5.15.0-72-generic for x86_64 architecture, Processors: 880.

2.7. Workflow focusing cephalopods' omics data to AMP datasets

The diagram representing how cephalopods' omics data have been focused to different AMP datasets to cater to a spectrum of research needs is displayed in Figure 1**.**

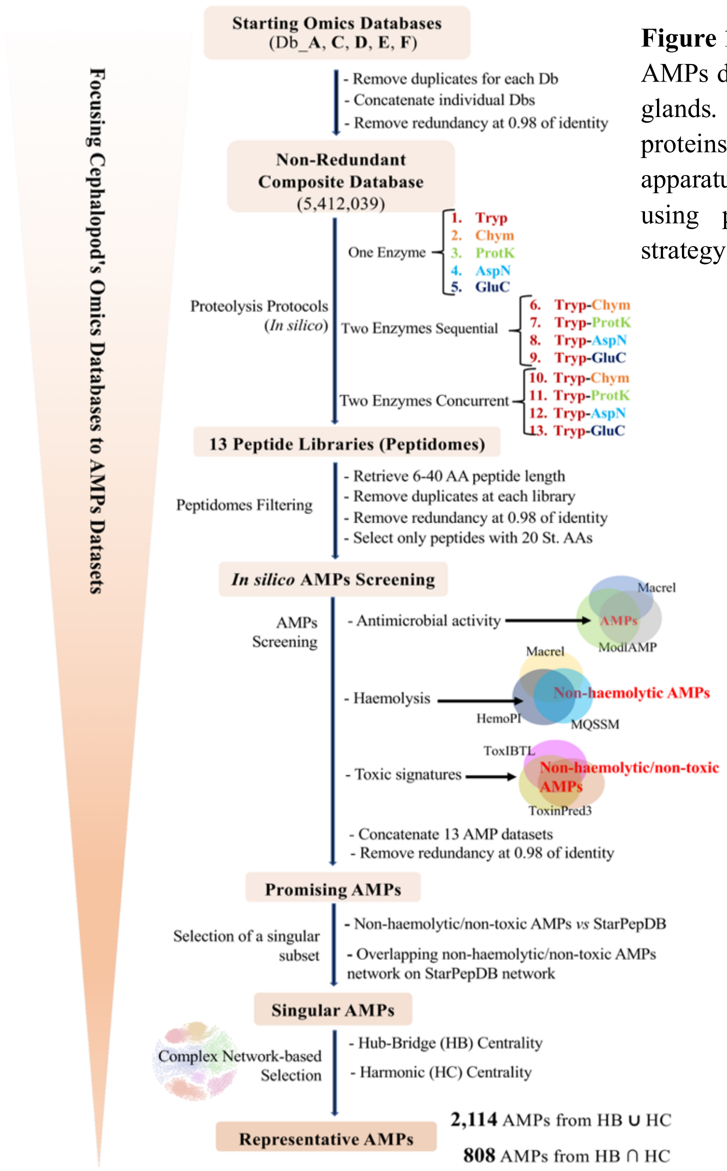

Focusing Cephalopod's Omics Databases to AMPs Datasets
Starting Omics Databases (Db_A, C, D, E, F)
- Remove duplicates for each Db
- Concatenate individual Dbs
- Remove redundancy at 0.98 of identity
Non-Redundant Composite Database (5,412,039)
Proteolysis Protocols (In silico)
One Enzyme
1. Tryp
2. Chym
3. ProtK
4. AspN
5. GluC
Two Enzymes Sequential
6. Tryp-Chym
7. Tryp-ProtK
8. Tryp-AspN
9. Tryp-GluC
Two Enzymes Concurrent
10. Tryp-Chym
11. Tryp-ProtK
12. Tryp-AspN
13. Tryp-GluC
13 Peptide Libraries (Peptidomes)
Peptidomes Filtering
- Retrieve 6-40 AA peptide length
- Remove duplicates at each library
- Remove redundancy at 0.98 of identity
- Select only peptides with 20 St. AAs
In silico AMPs Screening
AMPs Screening
- Antimicrobial activity
Macrel
AMPs
ModlAMP
- Haemolysis
Macrel
Non-haemolytic AMPs
HemoPI
MQSSM
- Toxic signatures
ToxIBTL
Non-haemolytic/non-toxic AMPs
ToxinPred3
- Concatenate 13 AMP datasets
- Remove redundancy at 0.98 of identity
Promising AMPs
Selection of a singular subset
- Non-haemolytic/non-toxic AMPs vs StarPepDB
- Overlapping non-haemolytic/non-toxic AMPs network on StarPepDB network
Singular AMPs
Complex Network-based Selection
- Hub-Bridge (HB) Centrality
- Harmonic (HC) Centrality
Representative AMPs
2,114 AMPs from HB ∪ HC
808 AMPs from HB ∩ HC

Figure 1. Workflow proposed to uncover several AMPs datasets encrypted in cephalopod salivary glands. This scheme shows how millions of proteins that characterize the cephalopod salivary apparatus are focused into several AMP databases using proteolysis and a rational screening strategy.

3. Results

3.1. Construction of the starting composite database from cephalopods salivary glands.

The composite protein database integrating transcriptomic and proteomic data used for the wide-proteome discovery of AMPs in *O. vulgaris*[18, 20], is re-utilized here, for uncovering AMPs encrypted within the salivary apparatus of cephalopods. The scheme for building such composite database is depicted in Figure 2, where it is evident that characterized AMPs originally integrated as database B are leaving out. The original smaller databases (A, C, D, E and F) that integrated the composite were analysed by considering pairwise similarities based on common sequences, which are encoded by the Jaccard index (Figure 2). Overall, the individual databases were rather unique compared to each other, except for database C and D, which shared a Jaccard index of 0.44, and database E and F, which shared a Jaccard index of 0.48. Databases C and D were somewhat related because the latter was used within the reference to detect the 2,427 proteins registered in database C. Similarly, databases E and F were sourced from non-standard ORFs from cephalopod PSGs transcriptomes (Figure 2). Redundancy was also explored within each individual database, and duplicates were found in 4 out of 5 databases. The resulting individual databases after removing duplicates can be found at doi:10.17632/hgwkkmms3h.1. Such individual databases were concatenated, and a more stringent redundancy reduction was carried out at 0.98 sequence identity on the resulting composite database. Thus, a non-redundant composite database made up of 5,412,039 proteins was created (doi:10.17632/gxmkytwdhx.1), to be used as substrate at the *in silico* proteolysis.

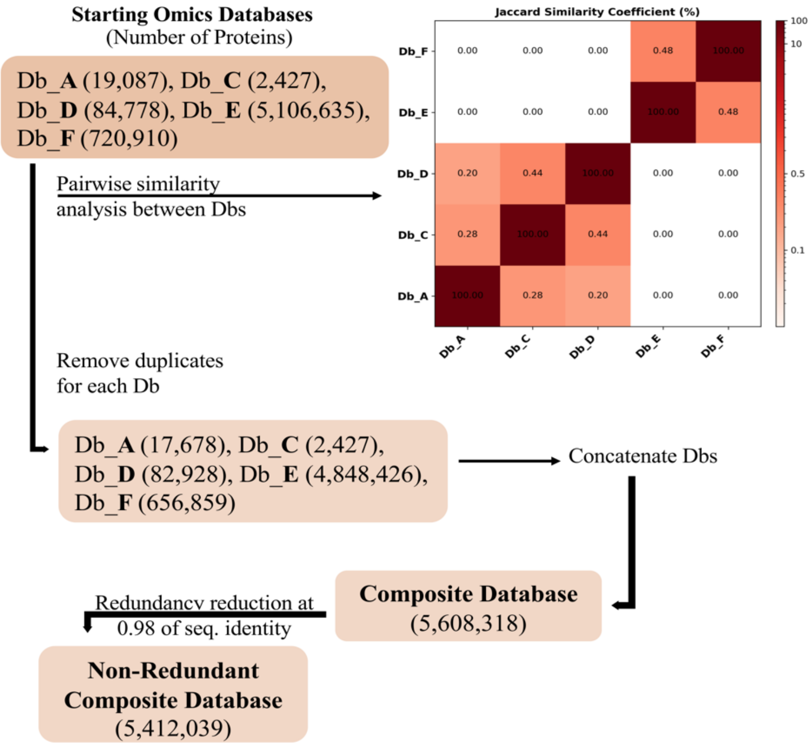
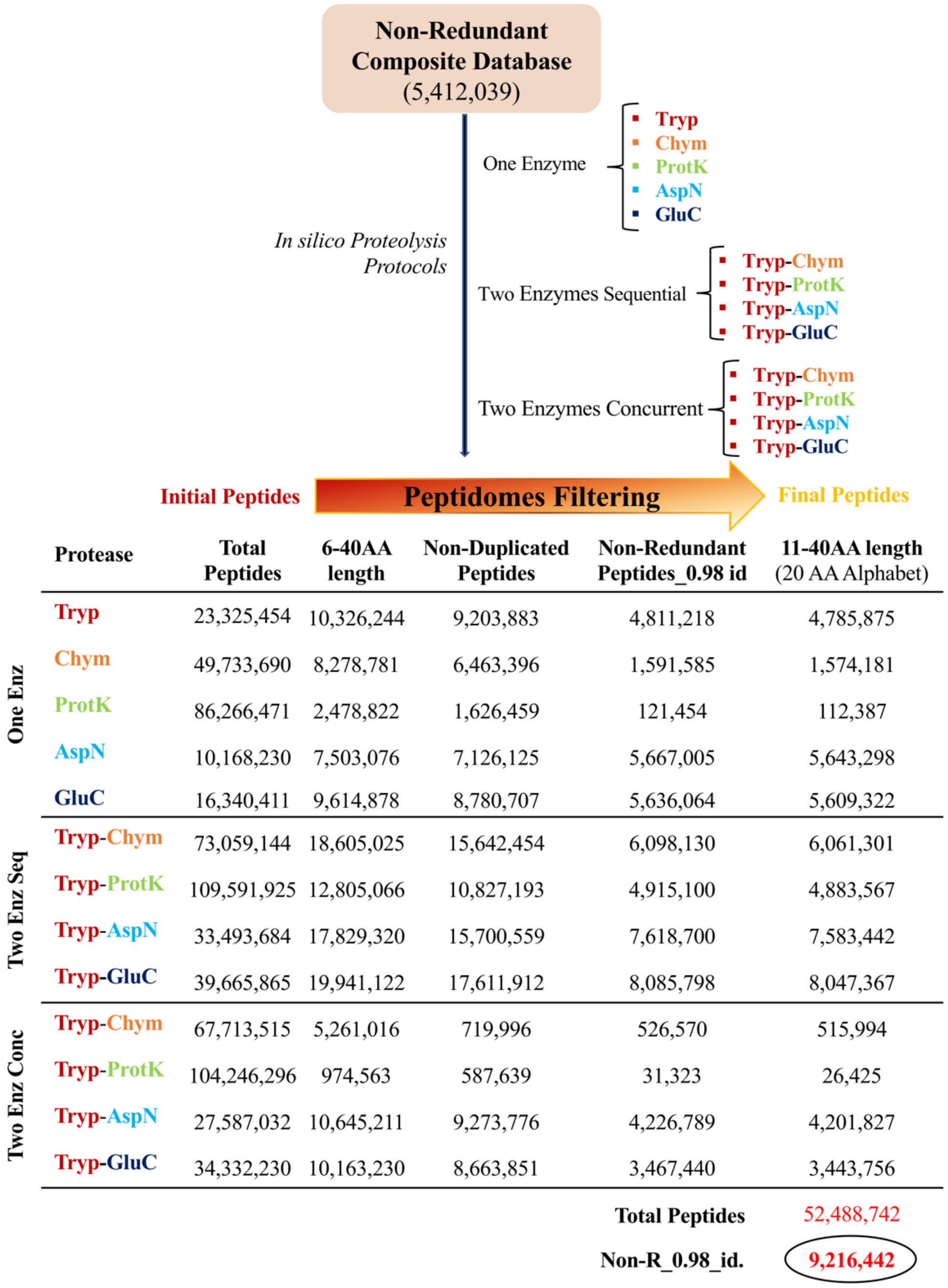

Starting Omics Databases
(Number of Proteins)
Db_A (19,087), Db_C (2,427), Db_D (84,778), Db_E (5,106,635), Db_F (720,910)
Pairwise similarity analysis between Dbs
Jaccard Similarity Coefficient (%)
Db_F 0.00 0.00 0.00 0.48 100.00
Db_E 0.00 0.00 0.00 100.00 0.48
Db_D 0.20 0.44 100.00 0.00 0.00
Db_C 0.28 100.00 0.44 0.00 0.00
Db_A 100.00 0.28 0.20 0.00 0.00
Db_A Db_C Db_D Db_E Db_F
100
10
0.5
0.1
Remove duplicates for each Db
Db_A (17,678), Db_C (2,427), Db_D (82,928), Db_E (4,848,426), Db_F (656,859)
Concatenate Dbs
Composite Database (5,608,318)
Redundancy reduction at 0.98 of seq. identity
Non-Redundant Composite Database (5,412,039)

Figure 2. Building a non-redundant composite database with 5,412,039 proteins for the *in silico* proteolysis**.** This figure illustrates the scheme followed for concatenating and analysing the starting omics database from cephalopods' salivary apparatus to generate the final composite protein database.

3.2. In silico proteolysis of cephalopods omics data and filtering of virtual peptidomes

This composite protein database was used as the substrate for the intended *in silico* digestion. The digestion used the five main proteases used in proteomics: trypsin, chymotrypsin, proteinase K, AspN, and GluC. Since trypsin is the most commonly used protease in proteomics, it was combined with the remaining four proteases in a sequential and concurrent mode. This combination was aimed at complementing the trypsin action with other cutting sites in order to obtain a higher diversity within the virtual peptidomes.

As previously mentioned, three main digestion protocols were applied involving five enzymes, so a total of 13 distinct enzymatic digestions were performed on the non-redundant composite database (Figure 3). Consequently, 13 virtual peptide libraries were generated, offering a wide peptide diversity from cephalopods to explore in the field of peptide science. Such peptidomes are publicly available at doi:10.17632/c3zhzgwsnw.1.



	Protease	Total Peptides	6-40AA length	Non-Duplicated Peptides	Non-Redundant Peptides_0.98 id	11-40AA length (20 AA Alphabet)
One Enz	Tryp	23,325,454	10,326,244	9,203,883	4,811,218	4,785,875
	Chym	49,733,690	8,278,781	6,463,396	1,591,585	1,574,181
	ProtK	86,266,471	2,478,822	1,626,459	121,454	112,387
	AspN	10,168,230	7,503,076	7,126,125	5,667,005	5,643,298
	GluC	16,340,411	9,614,878	8,780,707	5,636,064	5,609,322
Two Enz Seq	Tryp-Chym	73,059,144	18,605,025	15,642,454	6,098,130	6,061,301
	Tryp-ProtK	109,591,925	12,805,066	10,827,193	4,915,100	4,883,567
	Tryp-AspN	33,493,684	17,829,320	15,700,559	7,618,700	7,583,442
	Tryp-GluC	39,665,865	19,941,122	17,611,912	8,085,798	8,047,367
Two Enz Conc	Tryp-Chym	67,713,515	5,261,016	719,996	526,570	515,994
	Tryp-ProtK	104,246,296	974,563	587,639	31,323	26,425
	Tryp-AspN	27,587,032	10,645,211	9,273,776	4,226,789	4,201,827
	Tryp-GluC	34,332,230	10,163,230	8,663,851	3,467,440	3,443,756
					Total Peptides	52,488,742
					Non-R_0.98_id.	9,216,442

Figure 3. Tracking the filtering of the peptidomes resulting from each proteolysis protocol. The scheme illustrates how the number of peptides decreased as the screening steps increased. Initial peptides were produced by directly applying proteases and were filtered to satisfy mainly sequence length (6-40 amino acids) and redundancy (no duplicates and representative peptides at 0.98 sequence identity) criteria

Each resulting peptidomes was filtered to approach them to AMPs features. In this sense, only peptides ranging from 6-40 amino acids (AAs) in length were initially selected, followed by the removal of duplicates and a more stringent redundancy reduction at 0.98 sequence identity using cd-hit. At this stage, the length range for the peptides varied from 11 to 40 AA and non-standard AAs were also removed to facilitate further screenings.

Figure 3 shows how much each peptide library varied at each filtering step, arriving at the final libraries containing peptides with standard AAs ranging 11-40 AAs in length (doi:

10.17632/6fjsdnvygb.1). These 13 final peptide libraries were concatenated to give a total of 52,488,742 peptides, subsequently reduced to 9,216,442 peptides when applying redundancy removal at 0.98 sequence identity (doi:10.17632/v67g7r8nf2.1). This extensive but non-redundant peptidome sourced from cephalopods' salivary glands will be of great utility for those researchers who want discover new bioactive peptides by computational and vitro screenings.

3.3. Focusing cephalopods peptidomes to several AMP datasets.

The final peptidomes corresponding to each proteolysis protocol (last column of table shown in Figure 3) were screened individually against antimicrobial activity. To determine whether a query peptide is an AMP, consensus prediction agreement among three models was considered. The Figure 1A-SM contains 13 Venn diagrams corresponding to the screened peptidomes, showing the AMPs detected solely by macrel, modelAMP_RF, and modlAMP_SVM, respectively, as well as the agreement/intersection among the three prediction tools. The FASTA files containing AMPs libraries, identified by consensus across three prediction models for each proteolysis protocol, can be accessed freely at doi:10.17632/wwk7zzcfhv.1 Additionally, the results of predictions on the 13 individual peptidomes by the three models are available in file 1SM.

The consensus AMP libraries were then filtered by considering their toxicity potential, expressed by their haemolytic activity and the presence of toxic signatures. The haemolytic activity was first evaluated by three prediction tools. The Venn diagram representing non-haemolytic predictions from Macrel, HemoPI, and MQSSM is illustrated in Figure 1B-SM. The definitive predictions for non-haemolytic AMPs are found at the intersections. The corresponding FASTA files for the 13 non-haemolytic AMP consensus libraries can be accessed at doi:10.17632/pvptjh7kmv.1. Additionally, the raw predictions made by each individual model are available in file 2SM. Subsequently, these consensus non-haemolytic AMPs were screened against toxic signatures using ML-based and hybrid models implemented in ToxinPred3 and the deep learning tool ToxIBTL. Similarly, Venn diagrams were employed to establish consensus predictions for non-haemolytic/non-toxic AMPs, as depicted in Figure 1C-SM. The libraries containing non-haemolytic/non-toxic AMPs, identified through the agreement of the three models, can be accessed publicly at doi:10.17632/ccp94tgcp2.1. Furthermore, the raw predictions from each model are available for consultation in file 3SM. The tracking of this screening process from the peptidomes generated by the 13 proteolysis protocols to the generation of the datasets corresponding to non-haemolytic/non-toxic AMPs is summarized in Table 1.

Table 1. Focusing peptidomes resulting from each proteolysis protocol to AMP datasets. The table illustrates how the peptides libraries are rationally reduced by the robust detection of AMPs, non-haemolytic AMPs and non-haemolytic/non-toxic AMPs by three prediction tools at each screening step.

Proteolysis protocol	Peptidomes (No. peptides)	AMPs_Consensus	Non-Hem. AMPs_Cons	Non-Hem/Non-Tox. AMPs_Cons
Tryp	4,785,875	46,615	9,897	7,478
Chym	1,574,181	21,801	3,970	2,604
ProtK	112,387	775	310	157
AspN	5,643,298	294,959	33,108	22,978
GluC	5,609,322	404,990	43,955	31,756
Tryp-Chym_S	6,061,301	67,811	13,558	9,875
Tryp-ProtK_S	4,883,567	47,316	10,142	7,599

Tryp-AspN_S	7,583,442	307,767	36,455	25,454
Tryp-GluC_S	8,047,367	413,108	70,043	42,514
Tryp-Chym_C	515,994	1,179	600	**430**
Tryp-ProtK_C	26,425	168	148	**52**
Tryp-AspN_C	4,201,827	46,713	9,670	**7,270**
Tryp-GluC_C	3,443,756	57,691	10,335	7,696
Total	52,488,742	1,710,893	242,191	165,863
NonR_0.98_SeqId	**9,216,442**	**542,485**	**104,242**	**68,694**

Table 1 also displays in bold the total number of non-redundant AMPs after all libraries within each column were concatenated and sequence redundancy was removed with cd-hit at 0.98 of sequence identity. The resulting 542,485 AMP sequences from cephalopods, which represent a potential reservoir of novel AMPs, are promising for additional screenings to uncover peptide candidates for drug development (doi:10.17632/tr7xbp2pyt.1). This subset was further filtered by extracting 104,242 non-haemolytic AMPs, which may have an increased relevance for drug development (doi:10.17632/6gsdfj9876.1). However, the most promising dataset, made up of privileged AMPs, was obtained after toxic signatures were removed from non-haemolytic AMPs, rendering 68,694 non-haemolytic/non-toxic AMPs (doi:10.17632/8mttp4pvmc.1).

The evolution of the 13 virtual peptidomes at the key AMPs mining points, which are shaded in Table 1, is monitored by changes in the distribution of six global peptide features, such as length, amino acid (AA) frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment, within each peptide library class (Figure 2A-, 2B-, 2C-SM). Changes in the distribution of the global peptide feature values can be observed from the peptidomes (Figure 2A-SM) to the non-haemolytic/non-toxic AMPs (Figure 2C-SM). While median peptide length at the peptidomes are generally below 15 AAs, there is a shift to higher than 15 AAs with a top around 28 AAs in mostly of the non-haemolytic/non-toxic AMPs libraries. A similar shift to increased values is shown for the distribution of the pI and global charge. The median pI values distribution at peptidomes changed to be roughly around 8 to be consistently distributed around 10 at intermediate AMPs (Figure 2B-SM) and final AMPs datasets (Figure 2C-SM). Similarly, the global charge is completely shifted to the right at the AMPs and non-haemolytic/non-toxic AMPs libraries, where most of the AMPs take charges above 0. On the other hand, the hydrophobicity holds its values in a range from -1 to 1 for the peptidomes, intermediate, and final AMPs libraries, while the hydrophobic moment values slightly moved from a median of 0.35 at the peptidomes to higher values than 0.4 in the non-haemolytic/non-toxic AMPs libraries. The AA frequency did not change significantly from the peptidomes to AMP libraries, even focusing attention on the positively charged AAs, which are key for the antimicrobial activity.

The singularity of the peptide libraries generated at the AMPs mining points highlighted in Table 1, is also inspected by all-*vs*-all comparison using the Jaccard index. The Jaccard index quantifies how many peptides are shared by two libraries, namely the intersection of two sets. Thus, it is used as a pairwise similarity metric to evaluate the diversity among peptide libraries from each proteolysis protocol at three key AMPs mining steps. The Jaccard index heatmaps corresponding to each proteolysis protocol for the generated peptidomes, predicted AMPs consensus, and predicted non-haemolytic/non-toxic AMPs are shown in Figure 4.

Generally, the heatmaps show a striking singularity among the digestion protocols at each of the evaluated mining steps (Figure 4). The Jaccard index only reached values above 60% among the peptidomes when trypsin was compared to a combination of trypsin-chymotrypsin and trypsin-proteinase K in a sequential mode, or when these last proteolysis protocols were compared to each other. A significant library redundancy is also observed when comparing the proteolysis with AspN to its sequential action after trypsin (Figure 4A).

Similarly, redundancy among AMP libraries is mostly observed between trypsin and its sequential counterparts, trypsin-chymotrypsin and trypsin-proteinase K. However, additional pairs from concurrent mode, such as trypsin-AspN and trypsin-GluC, also show significant similarities with trypsin proteolysis. GluC proteolysis shows a high AMP redundancy with its trypsin-GluC sequential counterpart. The same sequential pairs that shared redundancy with trypsin, trypsin-chymotrypsin and trypsin-proteinase K, also share redundancy with the concurrent action of trypsin-AspN and trypsin-GluC (Figure 4B).

Finally, a similar redundancy pattern is displayed for the non-haemolytic/non-toxic AMPs (Figure 4C), including the high peptide redundancy derived from the action of AspN and the sequential proteolysis of trypsin-AspN.

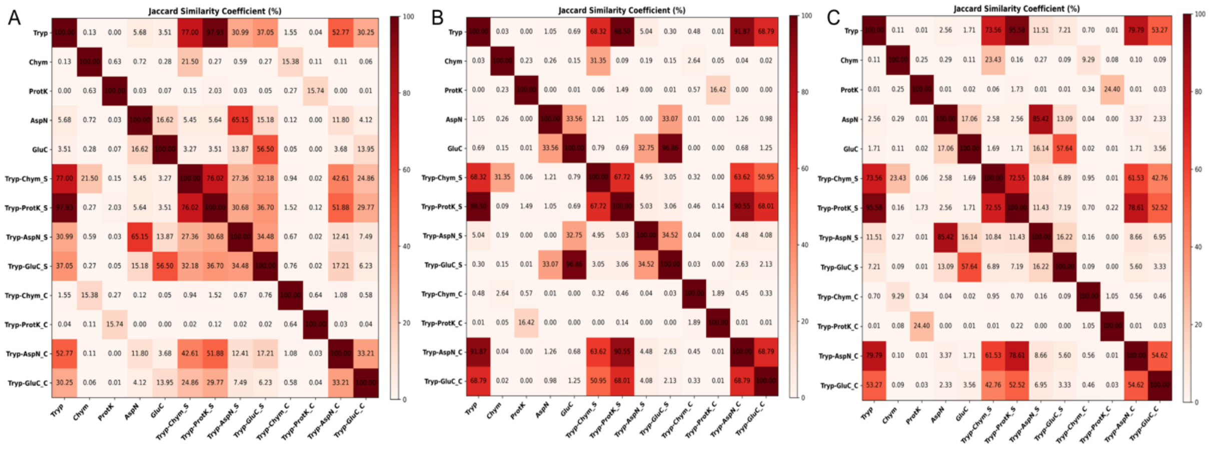


Figure 4. Peptide diversity among 13 proteolysis protocols at three steps of AMP mining on cephalopod salivary glands. **A**. Virtual peptidomes generated by 13 proteolysis protocols. **B.** AMPs detected by the consensus of three prediction models from peptidomes shown in A. **C**. Non-haemolytic/non-toxic AMPs detected by the consensus of three prediction models from AMPs libraries shown in B. Jaccard index is used as a pairwise similarity metric.

While the heatmaps allowed for comparative analyses even between proteolysis protocol pairs not originally intended in the primary design, this analysis suggests that sequential application of chymotrypsin and proteinase K after trypsin leads to high peptide redundancy at all mining stages. Similarly, but in a less consistent manner, this is observed for the concurrent action of trypsin with AspN and GluC, respectively.

Therefore, for future proteolysis-driven virtual mining efforts aimed at AMP discovery using the proposed enzymatic digestion protocols, it is not recommended to employ the sequential application of chymotrypsin and proteinase K following trypsin. Similarly, though with less emphasis, the concurrent action of trypsin with AspN and GluC should be avoided. These two recommendations are further supported by the observed recurrent similarity in the distribution pattern of global peptide features between trypsin and its sequential action with chymotrypsin and proteinase K, as well as between trypsin and its concurrent action with AspN and GluC at the same mining AMP stages (Figure 2A-, 2B-, and 2C-SM).

The singularity of the sequence space represented by the 68,694 non-haemolytic/non-toxic AMPs from cephalopods' salivary glands was evaluated against the 32,863 characterized AMPs registered in StarPepDB. To achieve this, both databases were compared using cd-hit-2d to identify how many and which AMPs from cephalopods were clustered to StarPepDB's members above identity cutoffs of 0.40, 0.50, 0.60, 0.70 and 0.80. The similarity clusters resulting from the comparison were parsed to extract the cephalopods AMP sequences satisfying the previously mentioned identity cutoffs. Both the similarity clusters and the FASTA files corresponding to the extracted subsets are shown in file 4SM. Out of 68,694 cephalopod AMPs, 63,228 were clustered to StarPepDB members above the threshold of 0.40 sequence identity, suggesting that these AMPs are more closely related to the known chemical space of characterized AMPs.

The remaining 5,466 non-haemolytic/non-toxic AMPs are denoted by the acronym CSPs (Cephalopods Singular Peptides), as explained in the Materials and Methods section. Both sets of AMPs are accessible at doi:10.17632/8mttp4pvmc.1, along with additional datasets that categorize the similarity with StarPepDB based on identity percentages within the following ranges: 40-50, 50-60, 60-70, 70-80, and greater than 80. These datasets consist of 26,744, 30,217, 5,716, 453, and 98 AMPs, respectively.

Given that the 5,466 CSPs share less than 40% of sequence identity with the characterized chemical space of AMPs, their internal diversity was also explored by all-*vs*-all global alignments (Figure 5). Figure 5A illustrates the heatmap of the pairwise sequence identities from all-*vs*-all global alignments above, while figure 5B shows the distribution/frequency of the peptide pairs satisfying sequence identities at ranges increasing in 0.10 units. This analysis was also applied to characterize the 63,228 cephalopods AMPs displaying similarities above 40% of identity with StarPepDB, using the aforementioned datasets discretized by identity ranges (Figure 3SM).

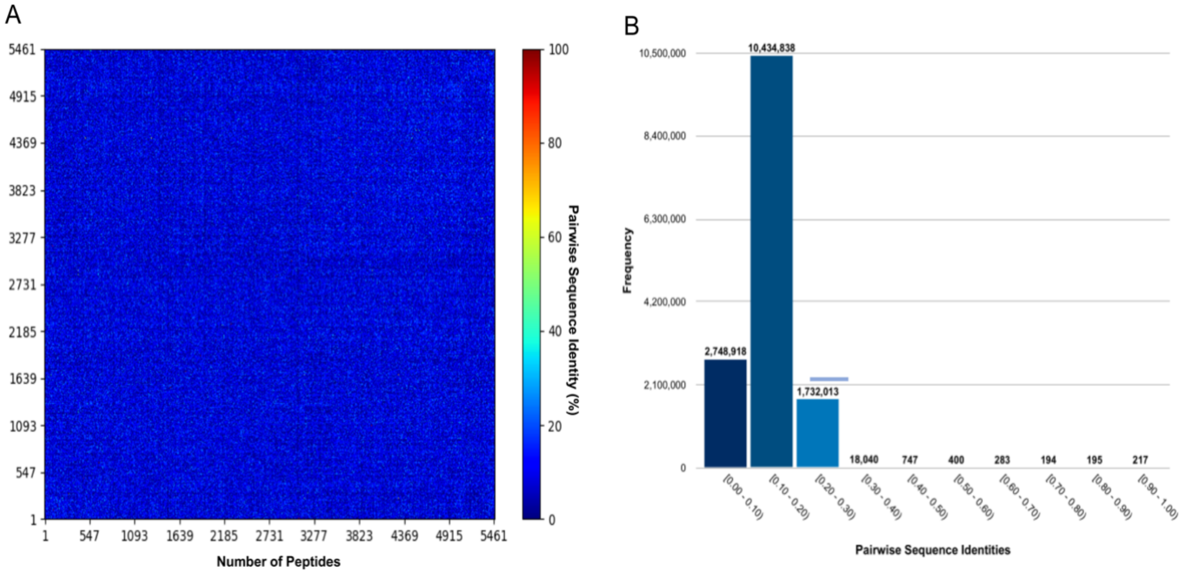


Figure 5. A. Heat map and B. histogram of pairwise sequence identity of the 5,466 CSPs. The heat map and histogram were built with in-house tools SeqDivA (https://github.com/eancedeg/SeqDivA)[42] and Dover Analyzer (http://mobiosd-hub.com/doveranalyzer/).[43]

As shown in Figure 5, the internal sequence diversity among the 5,466 CSPs is high, also indicating the structural singularity among its members. This singularity among these virtual scaffolds bearing privileged antimicrobial potentials is a strong point for peptide drug development.

3.4. The singularity of Cephalopods' AMPs from the outlook of complex networks

Based on the previous comparison, where 63,228 out of 68,694 promising cephalopod AMPs were identified as more closely related to characterized StarPepDB members, while the remaining 5,466 appear to be unique with respect to the known chemical space, the relatedness of the cephalopod AMPs with the characterized chemical space of AMPs can also be demonstrated using HSPNs, which are less computationally demanding at considering all peptides but not all pairwise similarity relationships.[36, 44] A HSPN was constructed from the 32,863 StarPepDB peptides and the 68,694 cephalopod AMPs, including the two subsets with different degrees of relatedness to the StarPepDB chemical space. A clustering algorithm was then performed over the network topology to delineate network communities that should group peptides with similar features. Figure 6 illustrates how the cephalopods chemical space represented by 68,694 promising AMPs are overlapped on the know sequence space represented by the 32,863 peptides from StarPepDB. HSPNs were used to project such chemical/sequence spaces.

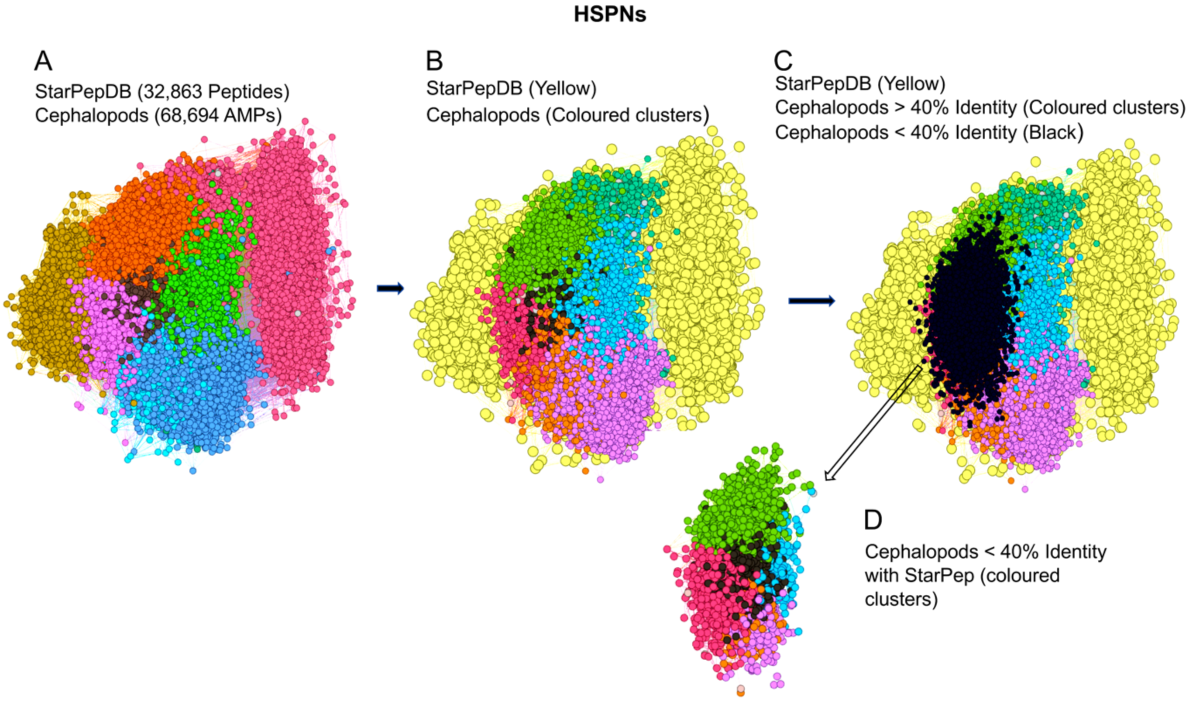

HSPNs
A
StarPepDB (32,863 Peptides)
Cephalopods (68,694 AMPs)
B
StarPepDB (Yellow)
Cephalopods (Coloured clusters)
C
StarPepDB (Yellow)
Cephalopods > 40% Identity (Coloured clusters)
Cephalopods < 40% Identity (Black)
D
Cephalopods < 40% Identity
with StarPep (coloured
clusters)

Figure 6. Superposition of the 68,694 non-haemolytic/non-toxic AMPs from cephalopods on the known sequence space represented by 32,863 peptides from StarPepDB, projected through Half-Proximal Similarity Networks (HSPNs). **A**. HSPN constructed with cephalopod and StarPep datasets. Clusters are delineated using different colours. **B**. HSPN projecting the superposition of cephalopods AMPs on StarPepDB members in yellow. **C**. HSPN projecting the overlapping of three AMP datasets: (i) StarPepDB in yellow, (ii) cephalopod AMPs sharing higher than 40% of sequence identity with StarPepDB, coloured by clusters, and (iii) cephalopod AMPs sharing less than 40% of sequence identity (black) with StarPepDB. **D**. HSPN projecting cephalopod AMPs sharing less than 40% of sequence identity with StarPepDB, highlighting the network clusters or communities

Figure 6C supports the findings of the comparison of the 68,694 non-haemolytic/non-toxic AMPs from cephalopods to StarPepDB using the cd-hit-2d tool. The chemical space corresponding to the cephalopod AMPs sharing higher than 40% of sequence identity with

StarPepDB is closer to the known sequence space of StarPepDB (coloured in yellow), while the sequence space occupied by CSPs sharing less than 40% of sequence identity (black) with StarPepDB is somewhat spatially disconnected from the yellow zone.

3.5. The singularity of Cephalopods AMPs as Seen from Physicochemical Characterization of Network Clusters

The 5,466 CSPs were studied in the context of the 32,863 peptides from StarPepDB. The HSPN consisting of both peptide datasets revealed nine clusters (Figure 7).

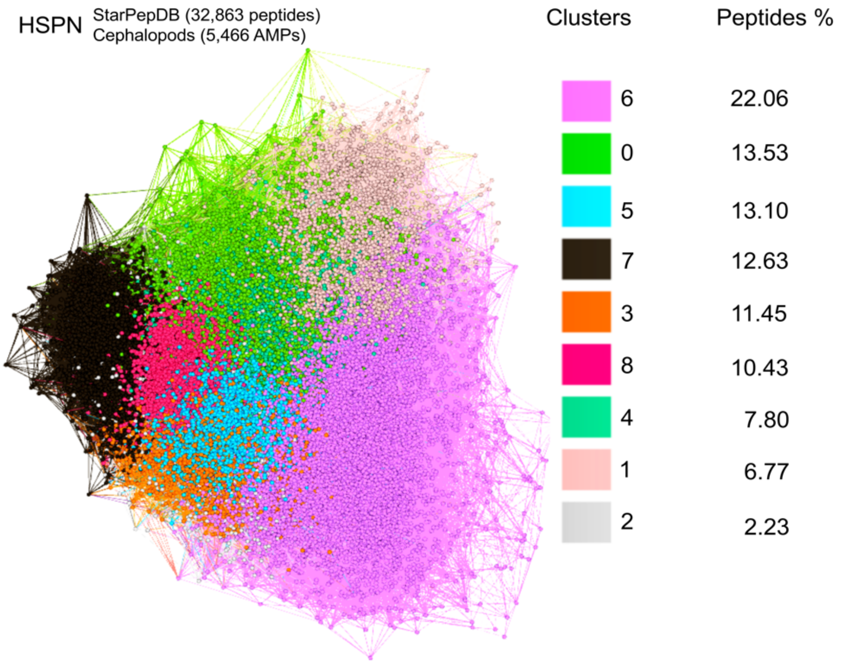


Figure 7. HSPN corresponding to the clustering of 32,863 peptides from StarPepDB and the 5,466 non-haemolytic/non-toxic AMPs from cephalopods. Nine clusters (0-8) were identified, and their peptide content is displayed as a percentage.

The peptides from each cluster were identified and physicochemically characterized. The detailed composition of peptide clusters and their physicochemical characterization can be found in file 5SM. Of the nine clusters, two were highly represented by CSPs: cluster 8 (50.7%) and cluster 5 (44.0%). The remaining clusters were only represented by 0.04%–18.40% CSPs (Figure 8).

Cluster 8 is characterized by having an intermediate peptide length (~36 AAs), low hydrophobicity (-0.21), high net charge (4.03), intermediate amphiphilicity (1.02), high isoelectric point (9.65), and a high Boman index (1.99). On the other hand, peptides from cluster 5 are shorter (~28 AAs) and more hydrophobic (-0.07), but they are also less charged (1.54), with a lower amphiphilicity (0.82), isoelectric point (8.54), and Boman index (0.95).

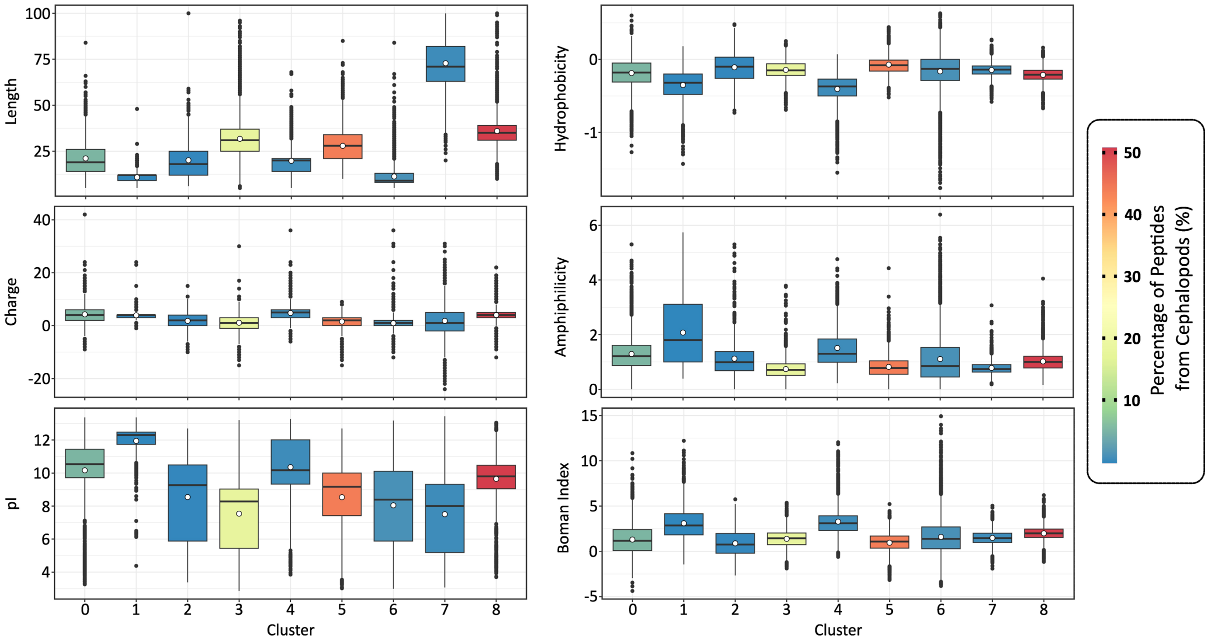


Figure 8. Physicochemical characterization of peptide clusters. This figure shows the distribution of the physicochemical properties of the peptides belonging to different network clusters or communities. The colour of each cluster represents the percentage of CSPs that it contains. Only clusters 8 and 5 are mostly represented by CSPs. The clusters were obtained after building a HSPN with the StarPepDB peptides and the CSPs.

Overall, peptides from clusters 8 and 5 differ from other peptide clusters in their sequence length, as they are neither as long as in cluster 7 (~73 AAs) nor as short as in cluster 1 (~11 AAs). Additionally, they tend to have higher net charge, hydrophobicity, and isoelectric point values. These findings provide more evidences that the CSPs are novel peptide representations.

3.6. Complex networks for extracting representativeness from the CSPs

The 5,466 CSPs were further reduced by extracting the most representative ones using network science. First, an HSPN projecting the chemical space of the CSPs was constructed. However, to achieve effective extraction of representative CSPs, the HSPN projecting the most informative topology should be used. This HSPN was found by applying an optimal similarity cutoff of 0.75 to produce a reasonable trade-off between the number of communities and singletons, considering the diversity of the CSPs set. A community or cluster within the network is considered when at least two nodes/peptides are connected, and the singletons are those that are not connected with any other in the network. Singletons are atypical peptides with singular structures that may represent privileged scaffolds for designing peptide drugs.

The optimal cutoff of 0.75 was determined by exploring network density at different similarity cutoffs. From 0.70 to 0.80, a significant change in network density is observed, reaching the desired value of 0.001 for HSPNs at a similarity cutoff of 0.75 (Figure 9A). At this similarity cutoff of 0.75, the number of communities/clusters increased to 60, while the network density decreased to 0.001, as mentioned before. Additionally, the number of disconnected peptides increased to 763, the so-called singletons, with a degree of 0 (File 6SM). The HSPN representing this topology is visualized in Figure 9B, after applying the Fruchterman-Reingold layout.

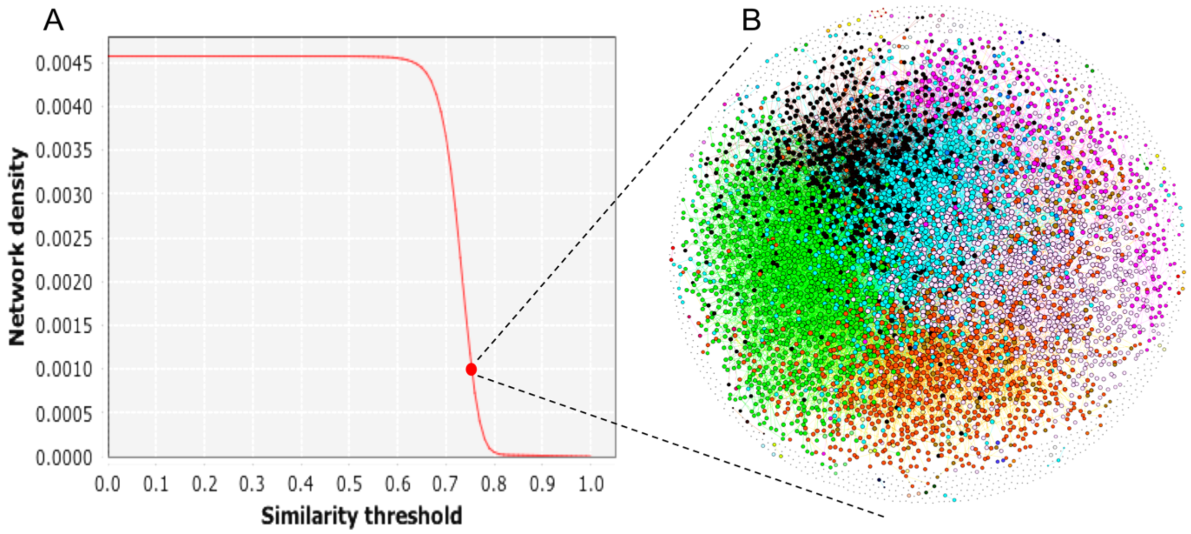


Figure 9. Selection of the most informative HSPN projecting the chemical space of the 5,466 CSPs, by applying an optimal similarity cutoff. **A**. Network density plot at different similarity thresholds. The similarity cutoff of 0.75, indicated in the plot, was selected as optimal. **B**. HSPN topology resulting from applying the optimal similarity threshold. The HSPN topology is formatted according to the Fruchterman-Reingold layout.

From this optimal HSPN topology, the most representative peptides were extracted using the procedure described in Ref.[36] Two subsets of non-redundant and representative peptides were extracted based on their harmonic (HC) and hub-bridge (HB) centrality measures. HC centrality weights the relevance or popularity of each peptide in the entire network, while HB centrality measures the relevance at the community level. Thus, two subsets of 1,469 and 1,453 CSPs were extracted using HC and HB centralities, respectively. File 6SM contains the sequences corresponding to these two subsets, the HSPN characterization at a 0.75 similarity cutoff, and the properties of its 5,466 nodes (CSPs), including the HC and HB values.

Finally, the union and intersection of these two subsets resulted in 2,114 and 808 non-haemolytic/non-toxic AMPs, respectively (Figure 10). These two final datasets are freely available at doi:10.17632/vv5fcxk5rn.2. The larger final dataset is a non-redundant but comprehensive representative subset of CSPs, while the smaller one is composed of the representative CSPs commonly identified by each centrality metric.

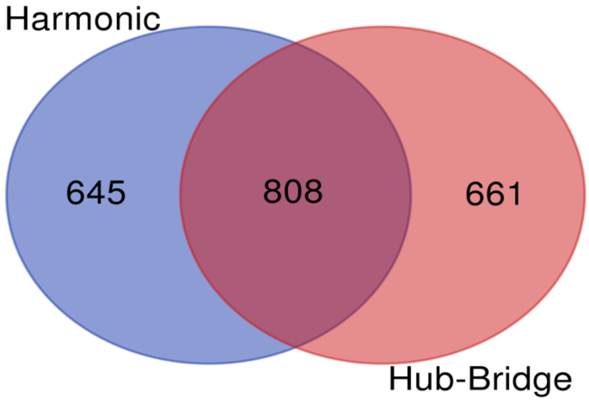


Figure 10. Venn diagram illustrating the union and intersection of the 1,469 and 1,453 non-haemolytic/non-toxic AMPs that were extracted using harmonic and hub-bridge centralities, respectively. From the union and intersection of these two subsets, the final AMP datasets from this study were obtained: 2,114 and 808 non-haemolytic/non-toxic AMPs from cephalopods.

4. Discussion

In silico proteolysis has been mostly applied to protein families from plants to identify promising bioactive with clinical potential.[45-49] However, this approach has not been extended to omics data for the same purpose. This proteolysis-based exploration has been limited to small protein datasets, likely due to the high dimensionality and diversity of peptides resulting from protease application, despite trypsin being the most commonly used protease and targeting 10.7% of the AAs.[27] Trypsin is the preferred protease for (MS)-based proteomics. It cleaves carboxy-terminal to arginine and lysine residues, resulting in a positive charge at the peptide C-terminus, which is beneficial for MS analysis. Nonetheless, other proteases are frequently used to gather supplementary data, such as AspN and GluC, which target acidic AAs, and chymotrypsin, which primarily targets aromatic AAs.[27]

The sequential use of these proteases following trypsin has recently been shown to enhance the identification of proteins and peptides by MS, even encompassing less commonly used proteases in proteomics like proteinase K due to its broad specificity, targeting 53.3% of AAs [27]. Inspired by these findings and the growing need to utilize omics data to identify new AMPs, we evaluated trypsin, chymotrypsin, AspN, GluC, and proteinase K *in silico*, as well as the activity of these last four proteases following trypsin in a sequential and concurrent manner, using a composite protein database that incorporates all proteomic and transcriptomic data from cephalopods salivary glands.[20]

One of the primary challenges of this work was addressing the "curse of dimensionality," which is exacerbated when generating peptidomes through *in silico* proteolysis of 5,412,039 proteins representing a comprehensive proteome characterizing the cephalopods' salivary apparatus. The total number of non-redundant peptides (9,216,442) from the 13 proteolysis protocols significantly exceeded the initial number of proteins (5,412,039). The selection of appropriate proteolysis and AMPs mining tools, capable of exploiting high-performance computing resources and integrated into a rational screening strategy combining machine learning, deep learning, multi-query similarity searches, and complex networks for AMP discovery, enabled the processing of millions of proteins/peptides until manageable AMP datasets were obtained. The RPG tool played a pivotal role in the AMPs mining process by facilitating the execution of the intended proteolysis protocols involving five proteases and, crucially, enabling the processing of millions of protein sequences from the composite database.[28]

The use of an encompassing omics database characterizing the cephalopods' salivary apparatus for the proteolysis-based AMPs exploration is a strong point of the study. This comprehensive database integrates 16 translated transcriptomes from cephalopods' PSGs using six ORF translations, considering non-canonical transcripts. Additionally, proteins shorter than 100 amino acids, often disregarded by the TransDecoder coding-region identifier tool, are included. Therefore, the *in silico* proteolysis not only revealed encrypted AMPs from existing proteins but also brought to light potential AMPs hidden in non-canonical proteins or in those typically methodologically discarded.

The rational *in silico* reduction from 9,216,442 unannotated peptides to various AMP datasets/libraries with varying relevance for further screenings aimed at discovering/developing new peptide drugs is depicted in Figure 11. This rational mining strategy yielded AMPs datasets that were subsequently narrowed down to a privileged subset of 5,466 CSPs, which could be represented by either 2,114 or 808 non-haemolytic/non-toxic AMPs according to network centralities. These AMPs datasets are publicly accessible and can be utilized by drug developers according to their specific requirements.

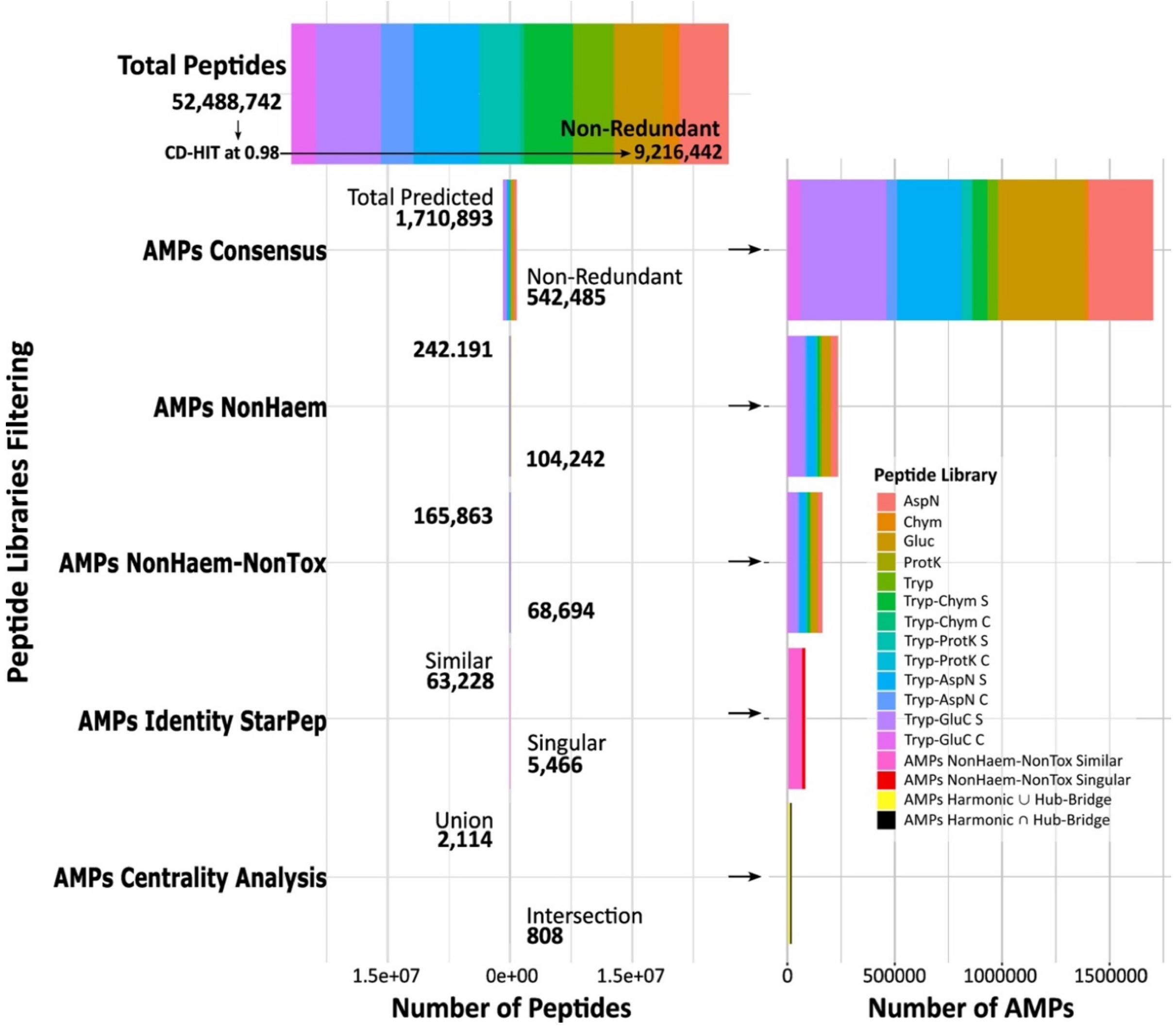


Figure 11. Tracking the screening of non-redundant cephalopods peptides (9,216,442) derived from the application of 13 proteolysis protocols. Different AMPs libraries were generated, considering the (i) antimicrobial activity (AMPs consensus), (ii) non-haemolytic potential (AMPs NoHaem), (iii) no presence of toxic signatures (AMPs NoHaem-NonTox), (iv) AMPs singularity regarding the known sequence space of StarPep (similar and singular NoHaem-NonTox AMPs), (v) Representative subsets according to network centrality analyses (union and intersection of the subsets extracted with HC and HB centralities).

The sequential application of chymotrypsin, AspN, GluC, and proteinase K after trypsin has been demonstrated to enhance peptide detection by MS.[27] However, our study shows that the sequential use of chymotrypsin and proteinase K following trypsin does not significantly increase peptide diversity compared to trypsin alone, despite both enzymes having more cleavage sites than trypsin. Furthermore, the concurrent action of AspN and GluC with trypsin does not significantly contribute to the diversity of the resulting libraries compared to trypsin alone. This is evident as AspN and GluC proteases are highly specific, targeting only aspartic (D) and glutamic (E) acids, which represent only 5.4% and 6.8% of AAs in proteins.[27]

5. Conclusions

Cephalopod salivary glands harbour a remarkable reservoir of AMPs, including non-haemolytic and non-toxic AMPs, underscoring the remarkable biological diversity of these marine invertebrates and their potential as antimicrobial agents. A significant portion of these AMPs exhibits unique sequences that expand the chemical space for exploration beyond existing databases.

Omics data integration and advanced *in silico* analyses provide a powerful strategy for AMP identification. This multifaceted strategy has the potential to uncover a vast array of AMPs, including those encrypted within existing and non-canonical proteins, as well as those present in smaller proteins often overlooked by standard translation tools. The proteolysis-driven mining strategy, coupled with rigorous virtual screening steps aimed to effectively identify promising AMPs based on their characteristic signatures, non-toxic nature, and sequence singularity expands the potential for AMP discovery in proteogenomic data.

Thus, the peptide datasets provided lay the foundation for further exploration of cephalopod salivary glands as a rich source of novel AMPs with therapeutic potential. These findings contribute significantly to the field of AMP research, being our approach extensive to other organisms, which hold promise for combating antimicrobial resistance and promoting peptide-based drug development.

Supplementary Materials: The following supporting information is available free of charge at:

- Figure 1SM – Venn diagrams representing the prediction results from the three evaluated models. The consensus prediction for 1A- AMPs detection, 2B- Non-haemolytic AMPs and 3C- Non-haemolytic/non-toxic AMPs.
- Figure 2SM – Distribution of global peptide features (length, amino acid (AA) frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment) within each peptide library class. 2A- AMPs consensus, 2B- Non-haemolytic AMPs, 2C- Non-haemolytic/non-toxic AMPs.
- Figure 3SM – Histograms of pairwise sequence identity for datasets sharing similarity with StarPepDB at following identity percentage ranges: 40-50, 50-60, 60-70, 70-80, and greater than 80.
- File 1SM – Raw prediction results for AMPs detection on the 13 individual peptidomes by each of the three models.
- File 2SM – Raw prediction results for non-haemolytic AMPs detection on the 13 individual peptidomes by each of the three models.
- File 3SM – Raw prediction results for non-haemolytic AMPs deprived of toxic signatures detection (non-haemolytic/non-toxic AMPs) on the 13 individual peptidomes by each of the three models.
- File 4SM – Similarity clusters resulting from the comparison between 68,694 non-haemolytic/non-toxic AMPs from cephalopods versus StarPepDB members at different identity cutoffs. It also contains FASTA sequences from cephalopods extracted from similarity clusters at different identity cutoffs.

- File 5SM – HSPN projecting the clustering of CSPs with StarPepDB. Clusters composition and their characterization through peptide length, charge, pI, hydrophobicity, amphiphilicity, Boman Index.
- File 6SM – HSPN that projects the chemical/sequence space of the 5,466 CSPs at 0.75 of similarity cutoff. HSPN properties and CSPs' representative subsets extracted with network centralities.

Author Contributions: G.A.-C and D.D.-P were involved in the design and conduction of all peptide mining analyses. K.C.-M worked mainly on the MQSS for haemolysis prediction and on HSPN clusters characterization. G.A.-C and Y.M.-P worked on the conceptualization, supervision, writing and reviewing of the manuscript. A.A participated in funding acquisition, writing and reviewing the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: G.A.-C. and A.A. acknowledge the support of the FCT - Foundation for Science and Technology under UIDB/04423/2020 and UIDP/04423/2020. Y.M.-P was supported by the USFQ Collaboration Grant (Project ID16911).

References

(1) Antimicrobial Resistance, C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **2022**, *399* (10325), 629-655. DOI: 10.1016/S0140-6736(21)02724-0 From NLM Medline.

(2) Miethke, M.; Pieroni, M.; Weber, T.; Bronstrup, M.; Hammann, P.; Halby, L.; Arimondo, P. B.; Glaser, P.; Aigle, B.; Bode, H. B.; et al. Towards the sustainable discovery and development of new antibiotics. *Nat Rev Chem* **2021**, *5* (10), 726-749. DOI: 10.1038/s41570-021-00313-1 From NLM PubMed-not-MEDLINE.

(3) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; et al. The value of antimicrobial peptides in the age of resistance. *Lancet Infect Dis* **2020**, *20* (9), e216-e230. DOI: 10.1016/S1473-3099(20)30327-3 From NLM Medline.

(4) Lei, J.; Sun, L.; Huang, S.; Zhu, C.; Li, P.; He, J.; Mackey, V.; Coy, D. H.; He, Q. The antimicrobial peptides and their potential clinical applications. *Am J Transl Res* **2019**, *11* (7), 3919-3931. From NLM PubMed-not-MEDLINE.

(5) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* **2016**, *44* (D1), D1087-1093. DOI: 10.1093/nar/gkv1278 From NLM Medline.

(6) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res* **2016**, *44* (D1), D1094-1097. DOI: 10.1093/nar/gkv1051 From NLM Medline.

(7) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for

development of new therapeutics. *Nucleic Acids Res* **2021**, *49* (D1), D288-D297. DOI: 10.1093/nar/gkaa991 From NLM Medline.

(8) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. *Bioinformatics* **2019**, *35* (22), 4739-4747. DOI: 10.1093/bioinformatics/btz260 From NLM Medline.

(9) Kumar, P.; Kizhakkedathu, J. N.; Straus, S. K. Antimicrobial Peptides: Diversity, Mechanism of Action and Strategies to Improve the Activity and Biocompatibility In Vivo. *Biomolecules* **2018**, *8* (1). DOI: 10.3390/biom8010004 From NLM Medline.

(10) Skarnes, R. C.; Watson, D. W. Antimicrobial factors of normal tissues and fluids. *Bacteriol Rev* **1957**, *21* (4), 273-294. DOI: 10.1128/br.21.4.273-294.1957 From NLM Medline.

(11) Tincu, J. A.; Taylor, S. W. Antimicrobial peptides from marine invertebrates. *Antimicrob Agents Chemother* **2004**, *48* (10), 3645-3654. DOI: 10.1128/AAC.48.10.3645-3654.2004 From NLM Medline.

(12) Wang, S.; Fan, L.; Pan, H.; Li, Y.; Qiu, Y.; Lu, Y. Antimicrobial peptides from marine animals: Sources, structures, mechanisms and the potential for drug development. *Frontiers in Marine Science* **2023**, *9*, Review. DOI: 10.3389/fmars.2022.1112595.

(13) Bachere, E.; Gueguen, Y.; Gonzalez, M.; de Lorgeril, J.; Garnier, J.; Romestand, B. Insights into the anti-microbial defense of marine invertebrates: the penaeid shrimps and the oyster Crassostrea gigas. *Immunol Rev* **2004**, *198*, 149-168. DOI: 10.1111/j.0105-2896.2004.00115.x From NLM Medline.

(14) Iwanaga, S.; Kawabata, S. Evolution and phylogeny of defense molecules associated with innate immunity in horseshoe crab. *Front Biosci* **1998**, *3*, D973-984. DOI: 10.2741/a337 From NLM Medline.

(15) Masuda, M.; Nakashima, H.; Ueda, T.; Naba, H.; Ikoma, R.; Otaka, A.; Terakawa, Y.; Tamamura, H.; Ibuka, T.; Murakami, T.; et al. A novel anti-HIV synthetic peptide, T-22 ([Tyr5,12,Lys7]-polyphemusin II). *Biochem Biophys Res Commun* **1992**, *189* (2), 845-850. DOI: 10.1016/0006-291x(92)92280-b From NLM Medline.

(16) Aguero-Chapin, G.; Galpert-Canizares, D.; Dominguez-Perez, D.; Marrero-Ponce, Y.; Perez-Machado, G.; Teijeira, M.; Antunes, A. Emerging Computational Approaches for Antimicrobial Peptide Discovery. *Antibiotics (Basel)* **2022**, *11* (7). DOI: 10.3390/antibiotics11070936 From NLM PubMed-not-MEDLINE.

(17) Matos, A.; Dominguez-Perez, D.; Almeida, D.; Aguero-Chapin, G.; Campos, A.; Osorio, H.; Vasconcelos, V.; Antunes, A. Shotgun Proteomics of Ascidians Tunic Gives New Insights on Host-Microbe Interactions by Revealing Diverse Antimicrobial Peptides. *Mar Drugs* **2020**, *18* (7). DOI: 10.3390/md18070362 From NLM Medline.

(18) Almeida, D.; Dominguez-Perez, D.; Matos, A.; Aguero-Chapin, G.; Osorio, H.; Vasconcelos, V.; Campos, A.; Antunes, A. Putative Antimicrobial Peptides of the Posterior Salivary Glands from the Cephalopod Octopus vulgaris Revealed by Exploring a Composite

Protein Database. *Antibiotics (Basel)* **2020**, *9* (11). DOI: 10.3390/antibiotics9110757 From NLM PubMed-not-MEDLINE.

(19) Perez-Polo, S.; Imran, M. A. S.; Dios, S.; Perez, J.; Barros, L.; Carrera, M.; Gestal, C. Identifying Natural Bioactive Peptides from the Common Octopus (Octopus vulgaris Cuvier, 1797) Skin Mucus By-Products Using Proteogenomic Analysis. *Int J Mol Sci* **2023**, *24* (8). DOI: 10.3390/ijms24087145 From NLM Medline.

(20) Almeida, D.; Domínguez-Pérez, D.; Matos, A.; Agüero-Chapin, G.; Castaño, Y.; Vasconcelos, V.; Campos, A.; Antunes, A. Data Employed in the Construction of a Composite Protein Database for Proteogenomic Analyses of Cephalopods Salivary Apparatus. *Data* **2020**, *5* (4), 110.

(21) Fingerhut, L.; Strugnell, J. M.; Faou, P.; Labiaga, A. R.; Zhang, J.; Cooke, I. R. Shotgun Proteomics Analysis of Saliva and Salivary Gland Tissue from the Common Octopus Octopus vulgaris. *J Proteome Res* **2018**, *17* (11), 3866-3876. DOI: 10.1021/acs.jproteome.8b00525 From NLM Medline.

(22) Gonçalves, C.; Costa, P. M. Cephalotoxins: A Hotspot for Marine Bioprospecting? *Frontiers in Marine Science* **2021**, *8*, Mini Review. DOI: 10.3389/fmars.2021.647344.

(23) Cho, J. H.; Sung, B. H.; Kim, S. C. Buforins: histone H2A-derived antimicrobial peptides from toad stomach. *Biochim Biophys Acta* **2009**, *1788* (8), 1564-1569. DOI: 10.1016/j.bbamem.2008.10.025 From NLM Medline.

(24) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658-1659. DOI: 10.1093/bioinformatics/btl158 From NLM Medline.

(25) Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **2016**, *11* (10), e0163962. DOI: 10.1371/journal.pone.0163962 From NLM Medline.

(26) Reina, D.; Toral, S.; Johnson, P.; Barrero, F. Improving discovery phase of reactive ad hoc routing protocols using Jaccard distance. *The Journal of Supercomputing* **2014**, *67*, 131-152.

(27) Dau, T.; Bartolomucci, G.; Rappsilber, J. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal Chem* **2020**, *92* (14), 9523-9527. DOI: 10.1021/acs.analchem.0c00478 From NLM Medline.

(28) Maillet, N. Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR Genom Bioinform* **2020**, *2* (1), lqz004. DOI: 10.1093/nargab/lqz004 From NLM PubMed-not-MEDLINE.

(29) Joshi, J.; Blankenberg, D. PDAUG: a Galaxy based toolset for peptide library analysis, visualization, and machine learning modeling. *BMC Bioinformatics* **2022**, *23* (1), 197. DOI: 10.1186/s12859-022-04727-6 From NLM Medline.

(30) Santos-Junior, C. D.; Pan, S.; Zhao, X. M.; Coelho, L. P. Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ* **2020**, *8*, e10555. DOI: 10.7717/peerj.10555 From NLM PubMed-not-MEDLINE.

(31) Muller, A. T.; Gabernet, G.; Hiss, J. A.; Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **2017**, *33* (17), 2753-2755. DOI: 10.1093/bioinformatics/btx285 From NLM Medline.

(32) Chaudhary, K.; Kumar, R.; Singh, S.; Tuknait, A.; Gautam, A.; Mathur, D.; Anand, P.; Varshney, G. C.; Raghava, G. P. A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci Rep* **2016**, *6*, 22843. DOI: 10.1038/srep22843 From NLM Medline.

(33) Castillo-Mendieta, K.; Agüero-Chapin, G.; Marquez, E. A.; Perez-Castillo, Y.; Barigye, S. J.; Pérez-Cárdenas, M.; Peréz-Giménez, F.; Marrero-Ponce, Y. A New Robust Method for Predicting Hemolytic Toxicity from Peptide Se-quence. **2023**.

(34) Rathore, A. S.; Arora, A.; Choudhury, S.; Tijare, P.; Raghava, G. P. S. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *bioRxiv* **2023**, 2023.2008.2011.552911. DOI: 10.1101/2023.08.11.552911.

(35) Wei, L.; Ye, X.; Sakurai, T.; Mu, Z.; Wei, L. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* **2022**, *38* (6), 1514-1524. DOI: 10.1093/bioinformatics/btac006 From NLM Medline.

(36) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Garcia-Jacas, C. R.; Chavez, E.; Beltran, J. A.; Guillen-Ramirez, H. A.; Brizuela, C. A. Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Sci Rep* **2020**, *10* (1), 18074. DOI: 10.1038/s41598-020-75029-1 From NLM Medline.

(37) Aguilera-Mendoza, L.; Ayala-Ruano, S.; Martinez-Rios, F.; Chavez, E.; Garcia-Jacas, C. R.; Brizuela, C. A.; Marrero-Ponce, Y. StarPep Toolbox: an open-source software to assist chemical space analysis of bioactive peptides and their functions using complex networks. *Bioinformatics* **2023**, *39* (8). DOI: 10.1093/bioinformatics/btad506 From NLM Medline.

(38) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, *2008* (10), P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

(39) Ghalmane, Z.; Hassouni, M. E.; Cherifi, H. Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining* **2019**, *9*, 1-22.

(40) Boldi, P.; Vigna, S. Axioms for centrality. *Internet Mathematics* **2014**, *10* (3-4), 222-262.

(41) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **1981**, *147* (1), 195-197. DOI: 10.1016/0022-2836(81)90087-5 From NLM Medline.

(42) Aguero-Chapin, G.; Galpert, D.; Molina-Ruiz, R.; Ancede-Gallardo, E.; Perez-Machado, G.; de la Riva, G. A.; Antunes, A. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* **2019**, *10* (1). DOI: 10.3390/biom10010026 From NLM Medline.

(43) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Tellez-Ibarra, R.; Llorente-Quesada, M. T.; Salgado, J.; Barigye, S. J.; Liu, J. Overlap and diversity in antimicrobial peptide databases: compiling a non-redundant set of sequences. *Bioinformatics* **2015**, *31* (15), 2553-2559. DOI: 10.1093/bioinformatics/btv180 From NLM Medline.

(44) Aguero-Chapin, G.; Antunes, A.; Mora, J. R.; Perez, N.; Contreras-Torres, E.; Valdes-Martini, J. R.; Martinez-Rios, F.; Zambrano, C. H.; Marrero-Ponce, Y. Complex Networks Analyses of Antibiofilm Peptides: An Emerging Tool for Next-Generation Antimicrobials' Discovery. *Antibiotics (Basel)* **2023**, *12* (4). DOI: 10.3390/antibiotics12040747 From NLM PubMed-not-MEDLINE.

(45) Prasertsuk, K.; Prongfa, K.; Suttiwanich, P.; Harnkit, N.; Sangkhawasi, M.; Promta, P.; Chumnanpuen, P. Computer-Aided Screening for Potential Coronavirus 3-Chymotrypsin-like Protease (3CLpro) Inhibitory Peptides from Putative Hemp Seed Trypsinized Peptidome. *Molecules* **2022**, *28* (1). DOI: 10.3390/molecules28010050 From NLM Medline.

(46) Qiao, L.; Li, B.; Chen, Y.; Li, L.; Chen, X.; Wang, L.; Lu, F.; Luo, G.; Li, G.; Zhang, Y. Discovery of Anti-Hypertensive Oligopeptides from Adlay Based on In Silico Proteolysis and Virtual Screening. *Int J Mol Sci* **2016**, *17* (12). DOI: 10.3390/ijms17122099 From NLM Medline.

(47) Guo, H.; Richel, A.; Hao, Y.; Fan, X.; Everaert, N.; Yang, X.; Ren, G. Novel dipeptidyl peptidase-IV and angiotensin-I-converting enzyme inhibitory peptides released from quinoa protein by in silico proteolysis. *Food Sci Nutr* **2020**, *8* (3), 1415-1422. DOI: 10.1002/fsn3.1423 From NLM PubMed-not-MEDLINE.

(48) Udenigwe, C. C. Towards rice bran protein utilization: In silico insight on the role of oryzacystatins in biologically-active peptide production. *Food Chem* **2016**, *191*, 135-138. DOI: 10.1016/j.foodchem.2015.01.043 From NLM Medline.

(49) Langyan, S.; Khan, F. N.; Yadava, P.; Alhazmi, A.; Mahmoud, S. F.; Saleh, D. I.; Zuan, A. T. K.; Kumar, A. In silico proteolysis and analysis of bioactive peptides from sequences of fatty acid desaturase 3 (FAD3) of flaxseed protein. *Saudi J Biol Sci* **2021**, *28* (10), 5480-5489. DOI: 10.1016/j.sjbs.2021.08.027 From NLM PubMed-not-MEDLINE.