antibiotics

Review

Emerging Computational Approaches for Antimicrobial Peptide Discovery

Guillermin Agüero-Chapin [1,2,*], Deborah Galpert-Cañizares [3], Dany Domínguez-Pérez [1,4], Yovani Marrero-Ponce [5], Gisselle Pérez-Machado [6], Marta Teijeira [7,8] 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; dany.perez@ciimar.up.pt

[2] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

[3] Departamento de Ciencia de la Computación, Universidad Central Marta Abreu de Las Villas (UCLV), Santa Clara 54830, Cuba; deborah@uclv.edu.cu

[4] Proquinorte, Unipessoal, Lda, Avenida 5 de Outubro, 124, 7º Piso, Avenidas Novas, 1050-061 Lisboa, Portugal

[5] Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Translacional (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, Ecuador; ymarrero@usfq.edu.ec

[6] EpiDisease S.L—Spin-Off of Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), 46980 Valencia, Spain; giselle.perez@epidisease.com

[7] Departamento de Química Orgánica, Facultade de Química, Universidade de Vigo, 36310 Vigo, Spain; qomaca@uvigo.es

[8] Instituto de Investigación Sanitaria Galicia Sur, Hospital Álvaro Cunqueiro, 36213 Vigo, Spain

[*] Correspondence: gchapin@ciimar.up.pt (G.A.-C.); aantunes@ciimar.up.pt (A.A.); Tel.: +351-22-340-1813 (G.A.-C. & A.A.)

check for updates

Citation: Agüero-Chapin, G.; Galpert-Cañizares, D.; Domínguez-Pérez, D.; Marrero-Ponce, Y.; Pérez-Machado, G.; Teijeira, M.; Antunes, A. Emerging Computational Approaches for Antimicrobial Peptide Discovery. *Antibiotics* **2022**, *11*, 936. https://doi.org/10.3390/antibiotics11070936

Academic Editor: Mire Zloh

Received: 14 June 2022
Accepted: 8 July 2022
Published: 13 July 2022

Abstract: In the last two decades many reports have addressed the application of artificial intelligence (AI) in the search and design of antimicrobial peptides (AMPs). AI has been represented by machine learning (ML) algorithms that use sequence-based features for the discovery of new peptidic scaffolds with promising biological activity. From AI perspective, evolutionary algorithms have been also applied to the rational generation of peptide libraries aimed at the optimization/design of AMPs. However, the literature has scarcely dedicated to other emerging non-conventional in silico approaches for the search/design of such bioactive peptides. Thus, the first motivation here is to bring up some non-standard peptide features that have been used to build classical ML predictive models. Secondly, it is valuable to highlight emerging ML algorithms and alternative computational tools to predict/design AMPs as well as to explore their chemical space. Another point worthy of mention is the recent application of evolutionary algorithms that actually simulate sequence evolution to both the generation of diversity-oriented peptide libraries and the optimization of hit peptides. Last but not least, included here some new considerations in proteogenomic analyses currently incorporated into the computational workflow for unravelling AMPs in natural sources.

Keywords: artificial intelligence; machine learning; AMPs; evolutionary algorithms; molecular descriptors; complex networks; proteogenomics

1. Introduction

The rise of resistance to antimicrobial agents evidenced in the last decades have caused excess healthcare costs worldwide [1]. The microbial natural resistance process, moved by evolutionary events, has been accelerated by the over-prescription and misuse of antibiotics [2]. This worrying situation has encouraged the search of new antibiotics from antimicrobial peptides (AMPs) with the ability to overcome resistance, mainly given by their versatile mode of action [3].Indeed, AMPs are not only considered for the development of antibiotics to treat multi-resistant bacterial strains [4,5], but also they are promising for the developing of antitumoral [6], antiviral [7], antifungal agents [8] and so on.

The discovery of peptides with relevant biological activities is a real challenge considering the great diversity of AMPs in terms of origin, structure, mode of action, activity, and, on the other hand by considering the overabundance of natural-occurring non-bioactive peptides [8]. Thus, several AMP databases with associated machine learning (ML)-based classifiers have been developed for over one decade, in order to assist wet-lab researchers in the long development process of peptide-based drugs [9]. AMP databases such as DAMPD [10], CAMPR3 [11], LAMP [12], DRAMP [13], ADAM [14], DBAASP [15] have incorporated ML predictors trained with alignment-free (AF) protein features such as amino acid (aa) and pseudo-aa composition, structural features, word frequency-based features, physicochemical aa properties with influence on the AMP activity, and some others [16,17] (Table 1). Figure 1 illustrates how databases and ML algorithms have been integrated to assist the discovery/design of AMPs for the developing of peptide drugs.

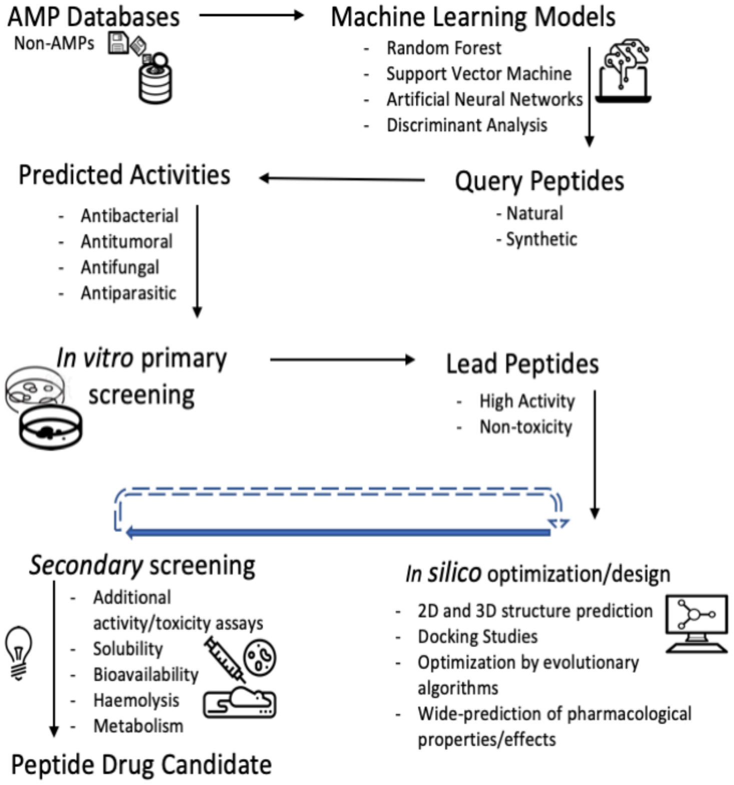

Figure 1. Workflow illustrating peptide drug discovery. The strategy involves the screening of query peptides from either natural or synthetic sources by applying ML models trained with the information stored in AMP databases. ML algorithms also assist the optimization/design step of lead peptides by means of a fitness/selection criterion [18,19].

The prediction tools built up with Support Vector Machine (SVM) and Random Forest (RF) based classifiers have been widely applied, but hardly considered the natural imbalance between the AMPs and non-AMPs [18]. On the other hand, emerging ML techniques such as Deep-Learning Neural Networks [18–21] and those based on the Rough Set Theory [22,23] have been applied to improve certain classification pitfalls like the quality in the learning phase and the classification boundaries between AMPs and non-AMPs, respectively. Although most of the classical predictive tools have focused on if a query peptide is an AMP or not, without targeting a specific biological activity among the reported for the AMPs [24], the current tendency is to address a hierarchical multi-level classification by downstream considering the specific biological activities of the AMPs as labels e.g., the antibacterial, antifungal, antiviral and antitumoral among others.

The most popular hierarchical multi-label classifiers, also listed in Table 1, are the following: (i) the iAMP-2L, a two-level classifier trained with Chou's pseudo amino acid composition (PseACC) [25], aimed at identifying AMPs and their five functional types [26], (ii) the iAMPpred predictor that combines compositional, physicochemical, and structural features into Chou's general PseACC for training a SVM multi-classifier [16], (iii) the MLAMP, a RF-based classifier built up with a non-classical PseACC sequence formulation incorporating a Grey Model that firstly discriminates AMP from non-AMPs, and then subclassify their biological activities into antibacterial, anti-cancer, antifungal, antiviral, and anti-HIV [27], (iv) the Antimicrobial Activity Predictor (AMAP) [28], a hierarchical multi-label classifier targeting 14 biological activities that is built up with SVM and XGboost tree [29] algorithms trained with amino acid composition (ACC) features, (v) the AMPfun webserver containing RF-based models that firstly classify AMPs and non-AMPs and afterwards address the prediction of AMPs functional activities including their possible target types [30], and more recently, the (vi) AMPDiscover [31] and the (vii) ABPFinder webservers (https://protdcal.zmb.uni-due.de/ABP-Finder/index.php; accessed on 7 March 2022) containing hierarchical RF-based classifiers built up with protein descriptors from the ProtDCal software [32] to firstly detect AMPs and antibacterial peptides (ABPs), respectively. While the AMPDiscover uses several downstream RF models to predict AMPs specific functions (antibacterial, antifungal, antiparasitic and antiviral), the ABP-Finder sub-classifies ABPs according to the Gram staining type of the potential targets (Gram-positive, Gram-negative bacteria, or broad-spectrum peptides with expected activity against both types of bacteria) by using a multi-classifier. The high success classification rates of both tools stems from considering the StarPep database [33] which is probably the most comprehensive curated repository of AMPs so far, and from performing an applicability domain (AD) analysis for the proposed ML models [31]. Both, AMPDiscover and ABP-Finder defined ADs for their corresponding RF-based models, however, AMPDiscover perform a rigours AD analysis at applying a consensus-based decision from five different approaches [31].

Despite the great number of reported ML-based tools for AMPs prediction, only few ones have considered the lack of balance among either the specific activities of AMPs or among their putative targets, as well as the AD of their corresponding models. The imbalance among AMPs and non-AMPs as well as the existing one among AMP activities was addressed by applying the synthetic minority over-sampling technique (SMOTE) during the IAMPE and MLAMP building [27,34] while the ABP-Finder addressed the imbalance among the bacterial target types of the ABPs (Gram+, Gram- and Gram+/- bacteria) by training a RF multi-classifier with a cost matrix weighting the different types of misclassified cases according to the imbalance ratio between the two classes (https://protdcal.zmb.uni-due.de/ABP-Finder/index.php; accessed on 7 March 2022).

On the other side, artificial intelligence (AI)-derived approaches like evolutionary algorithms have been applied to optimize lead candidates retrieved from the high-throughput screening in drug discovery. Evolutionary algorithms are inspired on several evolutionary events occurring in nature; they generally start with a small population of peptides identified as putative leads due to its relevant biological activities. The optimization is carried out by the generation of offspring peptides from these initial peptides by applying several operators simulating natural evolutionary process like cross-over and mutation operators, a parent and survival selection algorithms [40,41]. A parent selection algorithm is firstly applied on the initial peptide population to select the best parent peptides for the offspring generation. The survival aims at selecting a subset of good individuals (new population) from the generated offspring peptides. Then, the new peptide population will be iteratively subjected to the parent selection algorithm, evolutionary operators and the survival selection until finding an offspring peptide meeting a termination condition (selection criteria in Figure 2). The selection criteria can be represented by a fitness function which can be a ML model scoring peptide bioactivity. This selection process may be accompanied to experimental evaluations against the desired biological activities [42] (Figure 2).

Table 1. Summary of the most relevant ML approaches, from the classical to the emerging ones, for assisting the discovery of bioactive peptides from AMPs.

Classical AMP Prediction Tools				
Integrated to Database	**ML Algorithm**	**Peptide features**	**Implementation**	**Ref.**
CAMP_{R3}	RF, SVM, ANN, DA	AAC, net charge, hydrophobicity	http://www.camp3.bicnirrh.res.in/prediction.php	[11]
DRAMP 3.0	ANN, SVM, RF	Secondary structure features	http://shicrazy.pythonanywhere.com/	[13]
ADAM	SVM	AAC	http://bioinformatics.cs.ntou.edu.tw/adam/tool.html	[14]
DBAASPv3.0	Threshold value-based discrimination	Physicochemical properties accounting for the interaction with membrane	https://dbaasp.org/tools?page=general-prediction	[15]
Independent Tools				
ClasssAMP *	RF, SVM	Sequence-based features	http://www.bicnirrh.res.in/classamp/predict.php	[35]
iAMPpred *	SVM	compositional, physicochemical, and structural features	http://cabgrid.res.in:8080/amppred/server.php	[16]
iAMP-2L **	k-NN	PseAAC	http://www.jci-bioinfo.cn/iAMP-2L	[25]
AmPEP	RF	Sequence-based features	https://cbbio.online/software/AmPEP/	[36]
amPEPpy	RF	Global protein sequence descriptors	https://github.com/tlawrence3/amPEPpy	[37]
AMPScannerv1 **	RF	Physicochemical features	https://www.dveltri.com/ascan/v1/index.html	[38]
AMPfun **	RF	AAC-based features, physicochemical features and word frequency-based features	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	[30]
AMAP **	SVM and XGboost tree	AAC-based features	http://amap.pythonanywhere.com/	[28]
Emerging AMP prediction tools				
MLAMP **	RF	Non-classical PSeAAC	http://www.jci-bioinfo.cn/MLAMP	[27]
IAMPE	RF, k-NN, SVM, XGboost	NMR-based features	http://cbb1.ut.ac.ir/AMPClassifier/Index	[34]
AMPDiscover **	RF/DNN	Non-classical protein features (ProtDCal)	https://biocom-ampdiscover.cicese.mx/	[31,39]
ABP-Finder **	RF	Non-classical protein features (ProtDCal)	https://protdcal.zmb.uni-due.de/ABP-Finder/index.php	[Unpub]
AMPScannerv2	DNN	AA alphabet	https://www.dveltri.com/ascan/v2/ascan.html	[19]
ACP-DL	DNN	Binary profile feature and K-mer sparce matrix	https://github.com/haichengyi/ACP-DL (Standalone)	[20]
xDeep-AcPEP *	DNN	Physicochemical, biochemical, evolutionary and positional	https://app.cbbio.online/acpep/home	[21]

Methods listed in Table 1 are currently active (Accessed on 7 March 2022) * Multi-label classifiers allowing the prediction of specific biological activities (antibacterial, antifungal, antiviral, antitumoral and others) from AMPs ** Hierarchical multi-label classifiers addressing firstly AMPs detection and in the second level their specific biological activities. ACC: amino acid composition, ANN: artificial neural networks, DA: discriminant analysis, DNN: deep neural networks, k-NN: k- nearest neighbours, NMR: nuclear magnetic resonance, PseAAC: pseudo amino acid composition, RF: random forest, SVM: support vector machine.

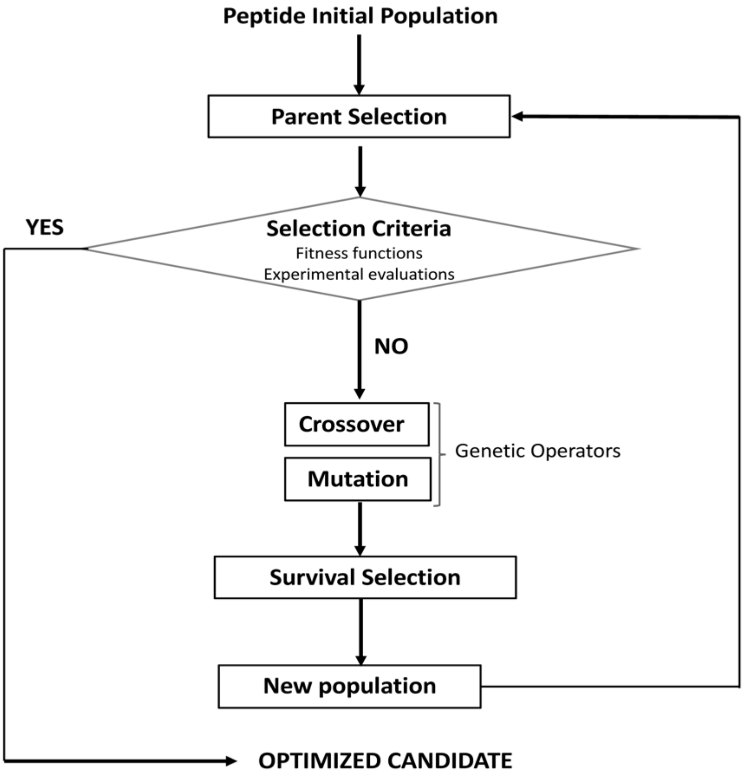

Figure 2. Workflow illustrating the main steps of evolutionary and genetic algorithms. Both approaches are very similar, in fact the use of evolutionary and genetic terms have been interchangeable. Genetic algorithms particularly use a fixed-length binary array to represent peptides as genes into a chromosome-like structure.

The genetic algorithm is the most popular technique among the evolutionary approaches where the peptides with promising biological properties (initial solution) are encoded as binary strings into chromosome-like structures, called genotypes. The optimization process is performed by evolving each chromosome toward optimized solutions by iteratively applying genetic recombination (crossover) operators and survival fitness functions that is somehow similar to the parent selection mechanism [40,42–44]. Optimized solutions in the case of peptides consist in generating structural entities with optimized biological properties e.g., peptides showing a trade-off among their pharmaceutical potency, solubility, haemolytic and toxicity properties [42] (Figure 2).

Despite AI-derived approaches have been largely applied to the rational search and design of bioactive peptides; most of them are represented by classical ML and evolutionary algorithms that frequently also use canonical sequence-based features as peptide descriptors and therefore have been documented in literature [18,45,46]. However, there is a growing number of emerging computational approaches effectively applied to the search/design of bioactive peptides that are comprehensively revisited here (Table 1).

Most of the non-standard approaches are represented by classical ML algorithms which are either trained with non-conventional peptide features [31] or combined with sequence alignment methods [47]. In addition to the singularity of these predictors; preprocessing steps managing the natural imbalance between bioactive and inactive peptides have been hardly applied to the AMPs predictions [27,34] as well as no big data solutions have been implemented yet to address scalability problems. As mentioned before, other less-known ML algorithms in the field of protein/peptide science like those based on the Rough Sets Theory (RST) are being currently intended for peptide classification/design [22,48]. Moreover, a non-conventional methodology that analyses the known chemical space of bioactive peptides by similarity networks was developed to identify the most relevant ones for each specific biological activity [33]. Such representative peptides were recently used in multi-query similarity searches against the StarPep database to repurpose AMPs for specific activities such as antiparasitic and tumour homing [49,50].

By other side, evolutionary algorithms that simulate sequence evolution have been recently applied to design/optimize peptides having a pharmaceutical activity [51,52]. Last but not least, computational tools used in proteogenomic analyses are being modified for uncovering cryptic peptides with biological activities in natural sources [53,54]. From now on we go deeper into these emerging approaches in peptide search and design

2. Non-Classical Peptide Features for Bioactivity Prediction

2.1. Peptide Features Inspired in Molecular Descriptors Used in Cheminformatics

There is a set of chemoinformatics-derived peptide features considered as “non-conventional” because of its in-house development; however, have been successfully applied in the recognition of bioactive peptides by ML-based classifiers [31,55–58]. The definition of these peptide/protein features is generally inspired on the mathematical formalisms applied to the calculation of molecular descriptors for small organic molecules [59,60], which have been traditionally used to Quantitative-Structure-Activity Relationship (QSAR) studies for drug design/search. Most of them are classified as topological descriptors since they consider the connectivity either between adjacent amino acids (aas) or between aa groups by using both algebraic and statistic invariants [32,61,62].

Those based on algebraic forms express protein/peptide structural topology through the definition of connectivity or adjacency matrices. The elements of these matrices (n_{ij} or e_{ij}) reflect topological relationships between the aas or aa groups, they are equal to 1 if i and j are adjacent otherwise take the value of 0. Topological indices (TIs) are estimated by applying several algorithms on the connectivity/adjacency matrix. The most common algorithms for the TIs calculation involve the powers of the topological matrix, the multiplication of a property vector by the topological matrix and the multiplication of vector-matrix-vector (Figure 3). Many of the most popular TIs within the cheminformatic have been defined by these algebraic formalisms, such as the Winner index (W) [63], the Randić invariant (χ) [64], Broto–Moreau autocorrelation (ATSd) [65], the Balaban index (J) [66], and the spectral moments introduced by Estrada [59]. Thus, many of them were reformulated to describe the spatial topology of aa sequences at different structural levels, e.g., linear sequences (1D), pseudo-secondary structure (2D) and the 3D-dimensional space [61,62] (Figure 3).

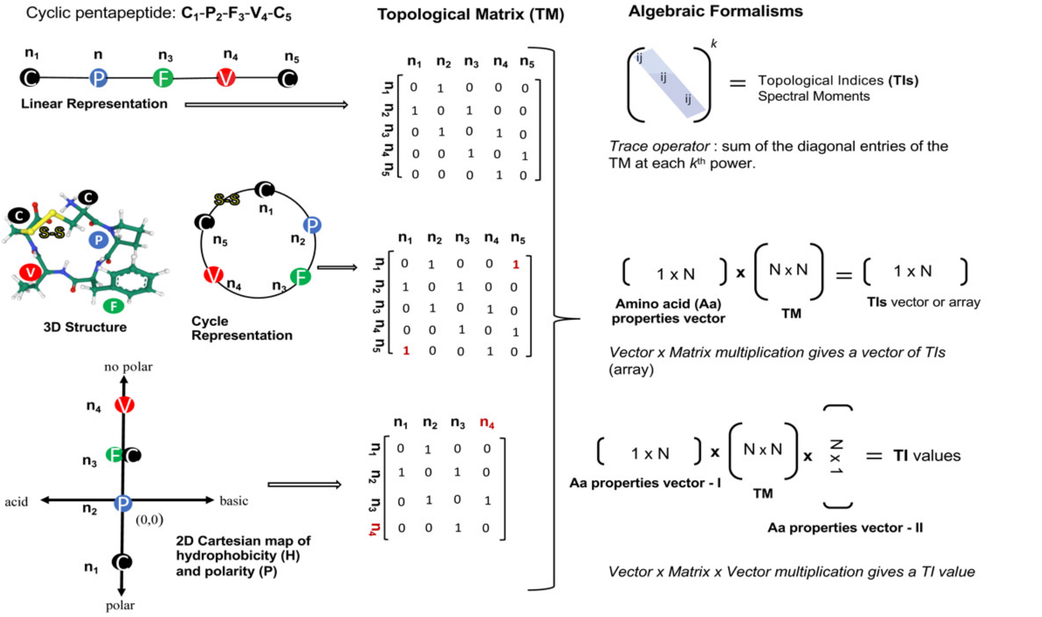
Cyclic pentapeptide: C_1-P_2-F_3-V_4-C_5
Topological Matrix (TM)
Algebraic Formalisms
n1 n n3 n4 n5
C P F V C
Linear Representation
3D Structure
S-S
Cycle Representation
no polar
polar
acid
basic
(0,0)
2D Cartesian map of hydrophobicity (H) and polarity (P)
n1 n2 n3 n4 n5
ij
k
= Topological Indices (TIs) Spectral Moments
Trace operator : sum of the diagonal entries of the TM at each k^{th} power.
(1 x N) x (N x N) = (1 x N)
Amino acid (Aa) properties vector
TM
TIs vector or array
Vector x Matrix multiplication gives a vector of TIs (array)
(1 x N) x (N x N) x N x 1 = TI values
Aa properties vector - I
Aa properties vector - II
Vector x Matrix x Vector multiplication gives a TI value

Figure 3. Workflow for the calculation of topological indices from several representation types of the cyclopentapeptide [CPFVC] with promising antiviral activity against the hantavirus cardiopulmonary

syndrome [67]. Each peptide representation defines a singular topological matrix (TM) encoding structural features at different degrees. In addition to the several ways to represent the topology of a peptide (linear, circular, 2D-Cartesian), several algebraic formalisms/operators can be applied on the TM to calculate different topological indices (TIs) types. n represents the nodes in the peptidic representations (linear, circular, and Cartesian) as well as in their corresponding TMs, which may contain some elements in red font (e.g., n_4 and 1) to highlight differences in structural encoding from the cyclopentapeptide. N indicates the number of rows and columns of matrices involved in TI calculation.

On the other hand, there is another set of topological descriptors that also comes from the chemoinformatic field that have been applied to the identification and design of AMPs [31,52,56,58]. They are not formulated by using algebraic forms but rather they rely on descriptive statistics as invariant operators on the aa properties either along the sequence or the 3D protein structure. In this case, the 1D or 3D topology is encoded by the application of classic cheminformatics algorithms that consider the neighborhood such as autocorrelation [65], Kier-Hall's electro-topological state [68], Ivanshiuc-Balaban [69], and Gravitational-like operators [70].

2.1.1. Topological Indices from Algebraic Forms

Among the TIs defined for small molecules, the spectral moments formalism probably is one of the most extended to characterize proteins and peptides structures [61,62,71,72]. The spectral moments may encode peptide structures through the definition of their corresponding topological matrixes and the application of the trace operator on the k-th power of such matrixes (Figure 3).

A sort of stochastic spectral moments applied to the electronic or charge delocalization of the aas within the peptide backbone and the entropy involved on such delocalization, were applied to model the bitter tasting threshold of dipeptides by linear discriminant and regression analyses [57]. These non-standard peptide features provided accuracies higher than 83% in the detection of bitter taste, and the regression models could explain the experimental variance of the bitter tasting threshold in more than 80%. It was shown the non-standard peptide descriptors correlate with the bitter taste as good as or even better than other well-known peptide features like the z-scale [73].

The spectral moments have been also applied to characterize bacteriocins. Bacteriocins are peptidic toxins produced and exported by bacteria as a defense mechanism to kill or inhibit the grow of other strains but the producer. The bacteriocins are very attractive for the development of new antibiotics and anticancer agents, however their high structural diversity represents a challenge for alignment-based predictive tools. Since the hydrophobicity and basicity of bacteriocins are relevant for their antibacterial activity, Agüero-Chapin et al. introduced the 2D-Hydrophobicity and Polarity (2D-HP) maps to pseudo-fold bacteriocin protein sequences in order to derive a set of spectral moments encoding information beyond the linear sequence [74] (Figure 4). These TIs are implemented in the Topological Indices to Biopolymers (TI2BioP) software [75] and were useful to build an AF model based on Linear Discriminant Analysis with a higher sensitivity (66.7%) than the attained by InterProScan (60.2%). In addition, they could detect cryptic bacteriocins, ignored by alignment methods [74].

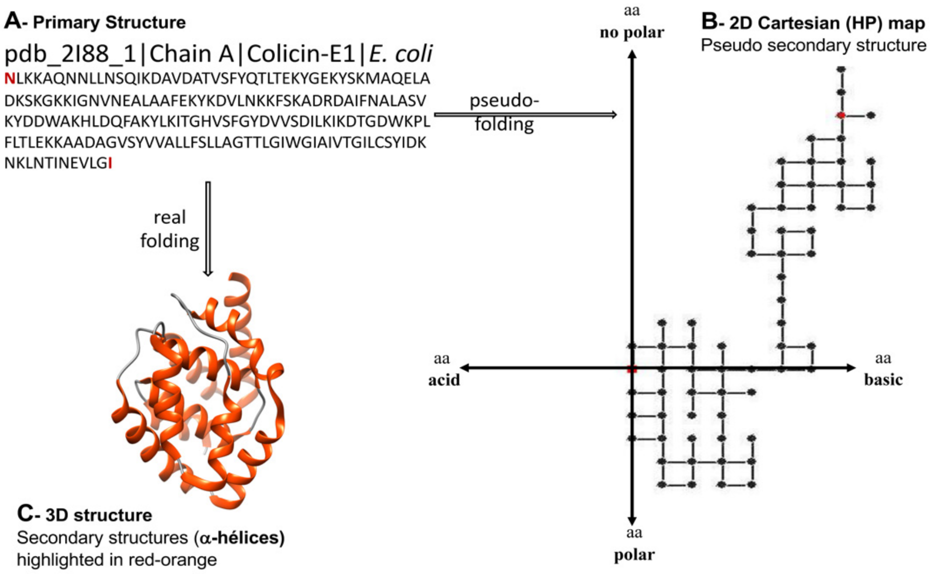
A- Primary Structure
pdb_2I88_1|Chain A|Colicin-E1|*E. coli*
NLKKAQNNLLNSQIKDAVDATVSFYQTLTEKYGEKYSKMAQELA
DKSKGKKIGNVNEALAAFEKYKDVLNKKFSKADRDAIFNALASV
KYDDWAKHLDQFAKYLKITGHVSFGYDVVSDILKIKDTGDWKPL
FLTLEKKAADAGVSYVVALLFSLLAGTTLGIWGIAIVTGILCSYIDK
NKLNTINEVLGI
real folding
pseudo-folding
aa no polar
B- 2D Cartesian (HP) map
Pseudo secondary structure
aa acid
aa basic
aa polar
C- 3D structure
Secondary structures (α-hélices) highlighted in red-orange

Figure 4. Different structural representations for the channel-forming domain of Colicin E1 (pdb 2I88). **A**—Primary structure, **B**—Pseudo secondary Cartesian map of hydrophobicity (H) and polarity (P) (2D Cartesian (HP) map), **C**—Three-dimensional structure. The 2D Cartesian protein map is an arbitrary bidimensional arrangement (pseudo-folding) of the protein/peptide sequences bearing higher-order useful patterns than contained in linear sequences.

2.1.2. Topological Indices from Descriptive Statistics

The cheminformatic-derived protein descriptors that have been widely applied to the prediction and design of bioactive peptides were developed and implemented by Ruiz-Blanco et al. in the ProtDCal software [32]. ProtDCal provides a great diversity of protein/peptide descriptors thanks to its divide-and-conquer methodology that considers both the aa properties and those estimated for groups, which can be modified by the neighbourhood through the application of classic previously-mentioned chemoinformatics algorithms. The modified properties of the aas or their resulting groups are later aggregated using statistical operators to estimate local or global descriptors either at sequence or 3D structural level. Although a more detailed description of ProtDCal's protein descriptors can be found in [32], the Figure 5 shows an schematic representation of the protein descriptor generation process of ProtDCal. The diversity of ProtDCal's protein descriptors represented by different families stems from combinatorically applying different aa properties, the ways to consider the vicinity to the target aa by several operators, the criteria used to group the aas as well as the invariant operator used for aggregating aa properties within the same array (Figure 5).

ProtDCal's descriptors have been involved in the discovery of antibacterial peptides by developing a non-conventional multi-target QSAR models [56]. Despite the AMPs selected for training were evaluated against multiple targets (Gram-positive bacterial strains), they could be integrated in the same model by modifying their ProtDCal's descriptors through the Box-Jenkins moving average operator. This operator allows modifying the sequence-based descriptors by subtracting the corresponding mean of the descriptors of all AMPs assayed against the same Gram-positive bacterial strain. This is a way to particularize a sequence-based descriptor by incorporating information about the experimental conditions or biological assays. With this kind of descriptors, the multi-target cheminformatic model displayed percentages of correct classification higher than 90.0% in both training and prediction (test) sets [56].

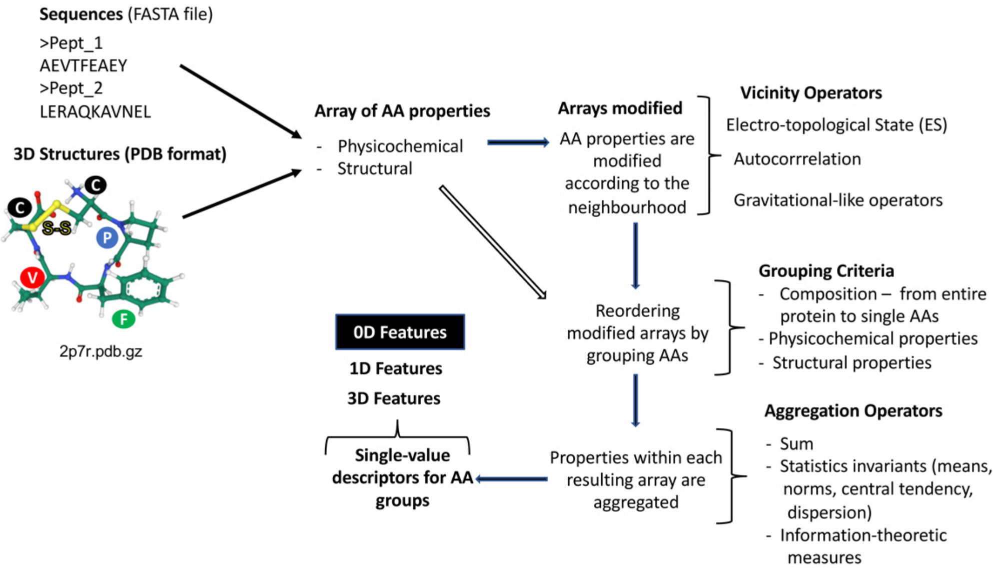
Sequences (FASTA file)
>Pept_1
AEVTFEAEY
>Pept_2
LERAQKAVNEL
3D Structures (PDB format)
C
C
S-S
P
V
F
2p7r.pdb.gz
Array of AA properties
- Physicochemical
- Structural
Arrays modified
AA properties are modified according to the neighbourhood
Vicinity Operators
Electro-topological State (ES)
Autocorrrelation
Gravitational-like operators
Reordering modified arrays by grouping AAs
Grouping Criteria
- Composition – from entire protein to single AAs
- Physicochemical properties
- Structural properties
0D Features
1D Features
3D Features
Single-value descriptors for AA groups
Properties within each resulting array are aggregated
Aggregation Operators
- Sum
- Statistics invariants (means, norms, central tendency, dispersion)
- Information-theoretic measures

Figure 5. Schematic representation of ProtDCal's descriptors calculation. 1D and 3D protein features implies the application of vicinity operators to modify amino acid (aa) properties while 0D features estimation go straightforward to group the original aa properties according to several grouping criteria.

Similarly, the same authors also applied the Box-Jenkins moving average operator to develop non-conventional multi-task QSAR models able to predict simultaneously antibacterial activity and toxicity [58]. This time, the continuous response variables measured on AMPs such as minimum inhibitory concentration (MIC), cytotoxic concentration at 50% (CC50), and haemolytic concentration at 50% (HC50) were transformed in a binary variable labelled as (1) referred to high antibacterial activity/low cytotoxicity, and (−1) assigned to low antibacterial activity/high cytotoxicity. The ProtDCal's descriptors that usually encodes only peptide features were modified by the Box-Jenkins moving average operator in order to consider the variability implying the evaluation of the antimicrobial activity and toxicity on different biological systems. Thus, a multi-task QSAR model displayed an accuracy higher than 96% for classifying/predicting peptides was built by using LDA discriminant [58].

ProtDCal's descriptors have also been involved in the design of new peptides that inhibit the *E. coli* ATP synthase, as putative antibiotics [52,76]. ProtDCal's descriptors, implemented in PPI-Detect [77], were applied to predict interactions between peptides and the main subunits of *E. coli*'s (Ec) and human's (Hs) F1Fo-ATP synthase. Those peptide with a maximum and a minimum interaction likelihood with EcF1Fo and HsF1Fo were selected for in vitro assays. An overall of three peptides resulted attractive for further optimization steps in the design of new antibiotics [52,76].

More recently, ProtDCal's protein descriptors were successfully applied to improve the prediction performance of the existing alignment-free models by using the largest experimentally validated non-redundant peptide dataset reported to date, the StarPepDB [78], together with Random Forest (RF) classifiers [31]. Pinacho-Castellanos at al. not only built RF-based models for identifying AMPs, but also addressed the main biological activities reported for them (antibacterial, antifungal, antiparasitic, and antiviral) as endpoints. The specific functions of AMPs were either directly predicted or by a hierarchical classification that first consider the antimicrobial activity. RF-based models, developed with ProtDCal's descriptors aimed to predict specific activities of AMPs, showed a higher effectivity and reliability than 13 freely available prediction tools. The best reported models were implemented in the AMPDiscover tool [31], publicly available at https://biocom-ampdiscover.cicese.mx/ (accessed on 7 March 2022). Ruiz-Blanco et al. also applied successfully ProtDCal's descriptors to predict antibacterial peptides by using

RF-based models trained with StarPepDB instances and, in a second step they are predicted on what bacterial targets according to their Gram-staining classification could be active by using a multi-classifier. These two RF-based models were implemented in the web server ABP-Finder: https://protdcal.zmb.uni-due.de/ABP-Finder/ (accessed on 7 March 2022) which is freely available but unpublished yet.

2.2. Integration of Peptide Features from Heterogenous Sources

Considering previous experiences in protein functional classification where protein features from heterogeneous sources have been integrated to improve classification rates; we wonder if this strategy has been applied to peptide classification? In this sense, the integration/combination of alignment-based (AB) and alignment-free (AF) protein features in machine learning models have been evaluated for such purpose. For example, Galpert et al. improved orthologs classification at the twilight zone (<30% of identity) by combining AB and AF protein similarity measures in supervised big data classifiers [79]. It has also been shown that the integration of AB and AF methods gives the best exploration of highly diverse protein classes, such as the nonribosomal peptide synthases (NRPS) represented by their A-domains [80]. Other examples of feature integration methods for remote homology detection can be found in [81], and the one of Borozan's et al. [82], based on weighted aggregation which is a very inclusive approach avoiding the loss of information.

Regarding AMPs classification improvements by integrating AB and AF peptide features, an algorithm applying AB measures and the SVM algorithm trained with AF pairwise measures was published for increasing AMPs prediction sensitivity [47]. The algorithm consists in two stages. Firstly, AMPs are identified by Basic Local Alignment Search Tool (BLAST) scores, and those peptides that cannot be unequivocally identified by pairwise alignments were inputted in an SVM-based classifier built with AF pairwise similarity scores. The AF similarity scores were estimated with the Lempel–Ziv's complexity algorithm [83]. The integrative algorithm achieved higher sensitivity performance for AMPs prediction than the prediction tools implemented within the first version of CAMPR3 database [11] and the integrated method proposed by Wang et al. [84]. Wang and colleagues had previously proposed a similar algorithmic workflow where BLAST is used to firstly classify a query peptide against a training set made up by 870 AMPs and 8661 non-AMPs. Classification label is transferred to the query peptide from the matching with highest similarity score. Query peptides that did not match with any within the training set were encoded by protein features like ACC and PseACC and the aas by five of their physicochemical and biochemical properties. As the number of generated features were relatively high, a rigorous feature selection step was performed by applying both the Maximum Relevance, Minimum Redundancy (mRMR) method [85] and the Incremental Feature Selection method [86] before building a Nearest Neighbour (NN)-based predictor. The NN algorithm assign the label AMP or non-AMP to a query peptide according to the class of the nearest neighbour.

Despite the efforts for integrating AB and AF features in a classification peptide system; they have actually been combined through their corresponding algorithms and have not been included in the same model or function. In this sense, AB and AF similarity scores could be combined to build an unique classifier for AMP prediction, as Galpert et al. did it for ortholog detection [79].

2.3. NMR-Based Features for Peptides

In 2020, the IAMPE webserver (http://cbb1.ut.ac.ir/; accesed on 17 March 2022) was released for an accurate prediction of AMPs by using classical ML-based classifiers trained with both conventional and ^{13}CNMR-based features. The non-conventional ^{13}CNMR-based features for peptides were defined from the quantitative NMR spectra for ^{13}C isotope of the naturally-occurring aas. Firstly, ^{13}CNMR-based features for each aa were calculated using ^{13}CNMR spectra signals. Secondly the aas were grouped according to their ^{13}CNMR-based features by applying Fuzzy c-means clustering algorithm. The resulting aa clusters were used to extract feature vectors along the peptide sequences according to classical "composition", "transition" and "distribution" patterns. Despite the new information provided by such non-conventional peptide descriptors, authors suggested their combination with physicochemical features to yield higher accuracy for the prediction of active AMP sequences [34].

3. Breakthroughs of ML Algorithms in the AMP Prediction

3.1. Data Imbalance and Multi-Label Classification in the Prediction of AMPs—New Algorithm Approaches

As mentioned in the Introduction, data imbalance is an issue to tackle in the classification of potential peptide sequences. Here, we collected some other reported solutions combining two-level classifiers with imbalance management in both, the first level binary AMP/non-AMP problem, and the second level multi-label functional type problem. For example, the authors of MAMP-Pred [87] proposed two alternative imbalance management methods: (i) under-sampling of the non-AMP class, and (ii) weighting sequences according to the imbalance ratio; the second one being eligible after the experiment process. Then, they used pruned sets and label combinations, considering label correlations, to transform the RF binary classification. For the classification assessment, the Matthew's correlation coefficient was selected for the first level, and the multi-label metrics: Exact-Match Ratio (EMR), Hamming-Loss (H-Loss), Accuracy (Acc), Precision (Precision, Recall), Ranking-Loss (RL), Log-Loss, One-error (OE), F1-Measure (F1-Mic, F1-Mac), for the second level. As they assessed, MAMP-Pred outperformed iAMP-2L (proposed in 2013 as a two-level multi-label classifier) because of the feature extraction process involved ACC and its eight physicochemical selected properties, besides the classification process.

Another example of imbalance management can be found in [88] where the authors tried to identify peptides with dedicated anti-CoV antimicrobial function on an imbalanced dataset with relatively insufficient positive data. They used NearMiss under-sampling and balanced RF to build the classification model, and the sensitivity, specificity and geometric mean for the unbiased evaluation.

Ensemble learning has also been used to cope with class imbalance in the binary AMP/non-AMP prediction tool Ensemble-AMPPred [89]. The prediction model based on ensemble methods (RF, max probability voting, majority voting, adaptive boosting, or extreme gradient boosting) was combined with feature extraction (vectors of 517 numerical descriptors representing peptide sequences), feature engineering (hybrid feature generation by the fusion of various selected features using a logistic regression model) and feature selection to improve classification accuracy after the application of a balancing clustering-based proportionate stratified random sampling that selected peptide sequences representing the positive and negative data. Thus, representative sequences selected from each cluster were used as training data, while the other remaining sequences, as testing data.

A recent report in [90] presents a multi-label framework HMD-AMP to hierarchically annotate peptide sequences into AMP/non-AMP, and then, into eleven functional classes that can be small and extremely imbalanced classes. The classification framework includes an embedding layer of protein sequences, a protein language encoder, a feature transformer and a hierarchical deep forest model. An ablation study and a reduced feature test demonstrate the effectiveness of the framework based on the detailed structural information of AMPs to improve the accuracy of the prediction model and to manage data imbalance problem. At each function prediction level, the model demonstrates a cascade forest structure where each cascade level is an ensemble of decision tree forests, and different types of forests are included to make the model diverse. It's worth noting that deep forest does not rely on backpropagation, so it is suitable for training data with either imbalance labels or small sample sizes, hence preventing the model from overfitting.

3.2. Deep-Learning in the Recognition of AMPs

The lack of samples in the positive class, as well as, the ambiguity in the negative class are key issues concerning deep learning models in AMP prediction as stated in review [91]. The starting point for knowledge discovery in this rough scenario is the correct representation of raw data. Precisely, deep learning provides a solution to the human expert dependence problem of featurization, which is known as representation learning; but also allows the application of some widely-used features in peptide machine learning by means of unsupervised embeddings (pretrained representations that can be fine-tuned with specific downstream supervised tasks), learned embeddings (usually one-hot or one-letter encoding on the amino-acid level, producing a dimension-reduced dense vector for subsequent layers), or engineered features (physicochemical or evolution-based properties).

In generative approaches for AMP discovery, recently reviewed in [92], the reliance on expertise-engineered features may limit the generation of candidates qualitatively distinct from known AMPs, or the limited number of known structures of the annotated peptides may reduce the effectiveness of structured-based models [93]. On the contrary, those attribute-controlled models based on recurrent neural networks, variational autoencoders, adversarial autoencoders, generative adversarial networks may encourage novelty of designed sequences. That is the case of the specific bidirectional conditional generative adversarial network developed in AMPGAN v2 [94] that learns data driven priors through generator-discriminator dynamics and controls generation using conditioning variables. Thus, a learned encoder mapping data samples into the latent space of the generator implements the bidirectional component that aids iterative manipulation of novel, diverse, and application-tailored candidate peptides.

The diversity target in generative models has been also tackled with a semi-supervised learning approach combined with a variational autoencoder (VAE) that can simultaneously learn from the large unlabelled peptide sequence databases and a limited number of labelled sequences as in PepCVAE [95]. In this case, a controlled generative model is learned from large unlabelled peptide database for the encoder and decoder losses, together with a much smaller labelled dataset (peptides with reported antimicrobial annotation) for the classifier loss, that is, using a large unlabelled corpus to capture the distribution with VAE, and a small labelled corpus to learn a certain controlling attribute code.

Also with VAE generation, the report in [96] used the Giant Repository of AMP Activity (GRAMPA) [97] to apply an improved automated semi-supervised approach based on stochastic long short-term memory (LSTM) encoder-decoder networks for generating promising new sequences and an experimental investigation, resulting in low minimal inhibitory concentration (MIC) AMPs against *Escherichia coli*, *Staphylococcus aureus*, and *Pseudomonas aeruginosa*. In this approach, the decoding from the same point in the latent space may result in a different peptide being generated and is dependent on the random seed set prior to running. Thus, the VAE is trained on a curated AMP dataset followed by the development of a regression model for activity prediction and the subsequent development of the latent space. Then, new AMP sequences are identified from the latent space (by sampling) and, subsequently, the AMPs are produced and characterized with their corresponding MIC values. This method produces peptides with similar MICs as the input reference peptides, but with novel sequences not found in the training set; at the same time, without imposing thresholds on peptide characteristics or otherwise biasing output post-sequence generation. As a result, a list of newly generated active peptides includes non-canonical AMPs of low helicity and low net charge.

An alternative data augmentation method is presented in [98] to improve the recognition of neurotoxic peptides via a convolutional neural network model. Novel potential neurotoxic peptides were discovered from the best performed model in a simulation dataset among the transcriptome of an endemic spider of South Korea, *Callobius koreanus* (*C. koreanus*). The BLAST-based augmentation method was intended to improve the generalization property of the model.

Specifically, for candidate short peptide generation, the authors in [99] combined LSTM generation and bidirectional LSTM classification to design short novel AMP sequences with potential antibacterial activity against *E. coli*. The models were trained using sequences with proven low MICs and tuned with Bayesian hyperparameter optimization.

Some other deep learning methods are reviewed in [100] as a promising approach to meet short-length peptides requirements [101] where they combine deep convolutional neural network with reduced aa composition comprising clustered aas on the basis of evolutionary information, substitution score, hydrophobicity, and contact potential energy. As a result, a short peptide of 20 aa was selected by Deep-AmPEP30 from sequences extracted from the gut commensal fungus *C. glabrata* genome and experimentally validated to have antibacterial activities similar to ampicillin.

In a recent review [39], the authors presented some reasons to select ML approaches over deep learning ones in AMP prediction and design, when a fair balance is required among high accuracy and generalization capability, interpretability and low computational cost. However, some improvements like parameter tuning or model hybridization may lead to more robust deep learning classifiers in this field.

3.3. Rough Sets Theory in the Classification of AMPs

As an example of model hybridization, the authors in [48] presented a codon-based genetic algorithm combined with rough set theory methods to find a peptide active against *S. epidermidis*. Their rough set theory method provided explicit boundaries between physicochemical properties that active sequences possess and inactive sequences do not possess. Since this method produced explicit decision components, they could test sequences containing multiple components. They were inspired in their previous publication [22] where they tried to reduce false discovery rate with a rough set-based classification method generating similarity rule set boundaries between active and non-active peptides based on their physicochemical properties.

Another example of the rough set theory application can be found in [102] where they implemented a rough set classification framework together with a Rough Set Quick Reduct and Rough Set Relative Reduct based on an improved Harmony Search algorithm to classifyAnti-HIV-1 peptides. Specifically, they hybridized a rough set-based feature selection technique, with population-based meta-heuristic algorithms (Particle Swarm Optimization), to classify the peptide sequences and solve dimensionality problems. Besides, a fuzzy set classification framework [23] was also intended to cope with limited and severely skewed high-dimensional space for short (<30 aa) AMP activity prediction.

4. Other Methodologies Than Classical ML for Identifying and Modelling AMPs

4.1. Homology-Based Prediction and Modelling of AMPs

The most popular approaches in addition to classical machine learning algorithms for the identification of AMPs in databases are local alignments which are represented by BLAST and FASTA tools [103,104]. Although local alignments have been successfully applied by using iterative rounds and filters such as the presence of signal peptides, aa patterns and gene vicinity during AMP searches [105–107], they can fail in identifying some AMP sequences [55], if compared to pattern-matching searches [107,108]. There are two main ways for searching for sequences by patterns: hidden markov models (profile-HMM) [109] or regular expressions (REGEX) [110]. Both the REGEX and profile-HMM methodologies work similarly for the identification of AMPs. Firstly, a set of homologous sequences are aligned and the multiple sequence alignment (MSA) is inputted to a specific program such as Pratt [111] or HMMER [112] for the identification of REGEX patterns or profile-HMM, respectively. Currently, instead of building REGEX patterns and profile-HMMs, they are available for many protein families at the Prosite [113] and Pfam [114] databases where a query sequence/peptide can be identified. The pattern/profile-based searches for AMPs can be complemented with the identification of signal peptides and other structural filters. In fact, improved versions of databases have incorporated MSA, profiles-HMM and molecular modelling for AMPs detection [11,115,116]. Even so, when a query peptide could be high-scored against profile-HMMs from different peptide families, it is advisable to use a prediction tool combining different protein signature recognition methods such as InterProScan [117].

As we previously mentioned, the molecular modelling complements AMPs pattern-based searches by confirming expected three-dimensional (3D) structural features characterizing them. The 3D structure can be also integrated into homology-based searches to identify homologous sequences sharing low identity but retaining a great structural conservation. Such structural similarities have enabled the detection of AMPs in databases with higher accuracy [9]. When the structures of peptides are not experimentally elucidated, two modelling techniques are suggested: homology-based and *ab initio* modelling. The homology-based modelling uses the structure experimentally-determined from available homologous as template to infer the 3D structure of novel peptides, but rather using structural than sequence similarities, especially if the query and template are remote homologous [118]. By contrast, the *ab initio* method is used to predict the structures of peptides with yet unknown homologs. The prediction of the 3D protein structure starts from scratch requiring an energy model describing the main factors that contribute to the stability of the folding process and an efficient method for the conformational space exploration of the peptide chain [119]. However, homology-based approaches are more suitable for peptides when homologs are identified. In fact, the second release of BACTIBASE [115] incorporated the MODELLER program [120], as a tool for the 3D structure prediction of query peptides by homology to known bacteriocins [115]. Besides, the incorporation of 3D structure prediction tools to AMP databases provide another filter for an accurate identification of query AMPs, the 3D structure can be used for scoring peptide-cellular target interactions which is a crucial step for the *in-silico* design of novel AMPs [121].

Especially, since classical ML algorithms were recently reviewed in [18], we have addressed here, traditional homology-based approaches applied to the search and the modelling of AMPs, and will describe next, the most singular algorithms.

4.2. Emerging ML-Independent Methodologies for AMP Prediction/Design

In this section, we will address other emerging methodologies regardless of ML approaches and classical homology-based approaches for AMP discovery. Firstly, we want to highlight the AMPA webserver (http://tcoffee.crg.cat/apps/ampa, accessed on 7 March 2022), developed to detect antimicrobial stretches within the protein sequences. The antimicrobial regions detected in proteins can serve as new templates for AMP design, especially those uncovered within proteins no related with the defense function. AMPA algorithm does not depend on homology-based searches since it estimates an antimicrobial index (AI) to each aa, derived from half-maximal inhibitory concentration (IC_{50}) values in high-throughput screening experiments, encoding the propensity of each aa to be present in an AMP sequence. As low IC_{50} values correspond to high activity, aas with low AIs are more likely to be part of an AMP. By applying a sliding-windows analysis along the protein sequence, AMPA generates an antimicrobial profile based on the AIs. Those regions scored below certain threshold are considered putative antimicrobial domains [122]. The singularity of this approach is that it doesn't either rely on building machine learning models or similarity searches against AMP databases. However, potentially conserved antimicrobial regions can be checked in conjunction with the T-coffee alignment tool [123].

On the other hand, complex networks have been applied to explore the chemical space of AMPs aimed to discover structural entities with promising biological activities that also could serve as template for peptide drugs design/optimization. In this sense, Marrero-Ponce et al. were the pioneers on this topic by publishing a seminal of related works [33,78,124]. Firstly, Marrero-Ponce et al. analyzed both the diversity among 25 AMP databases and the showed within each one. The study revealed some AMP databases contained common sequences showing certain overlapping degree. After removing duplicates among AMP databases, a representative set of 16 990 non-redundant

AMPs was collected, which probably was the most comprehensive and exhaustively curated AMP dataset at that moment [124]. This relevant dataset was further enriched and structured in a graph database called StarPepDB (http://mobiosd-hub.com/starpep/; accessed on 17 March 2022) integrating 45 120 unique peptide sequences from 42 AMPs databases (Figure 6), with their metadata (origin organisms, function, biological target, source database, chemical modifications, cross-referenced entries to UniProt, PDB and PubMed) [78].

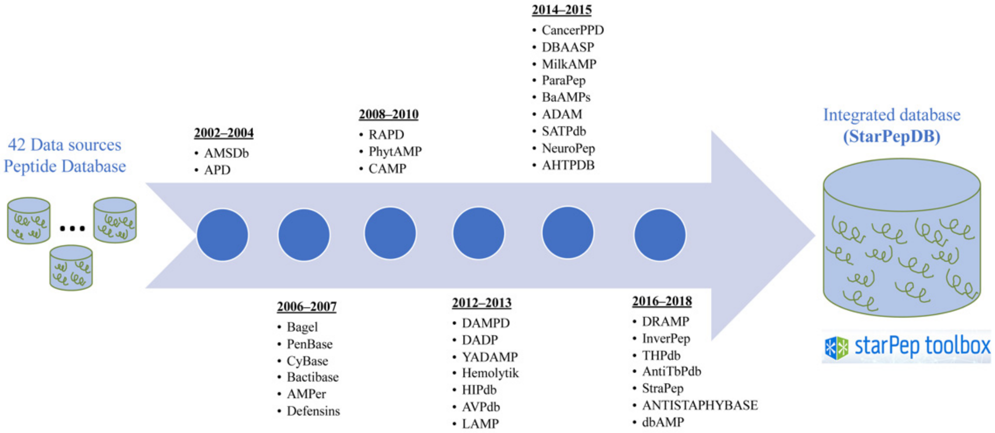

Figure 6. Chronological listing of AMP databases used in StarPep Database (StarPepDB) compilation. After collecting web pages from a large variety of bioactive peptide databases (see Table 1 in Ref. [78]), their contents were integrated into a graph database that holds total of 71.310 nodes and 348.505 relationships. In this graph structure, there are 45.120 nodes representing peptides (unique sequences) and the rest of the nodes are connected to peptides for the describing metadata.

StarPepDB has a star-like network architecture where a central node represents the peptide sequence and is connected to neighbour nodes labelled with the metadata. The edges depict a relational and unidirectional connection of the central node by a using a set of selection criteria "produced by", "assessed against", "*related to*", "*compiled in*" with its corresponding metadata nodes such as the origin, target, function and database, respectively. Peptide nodes besides the sequence also contain peptide's ID and length, while the metadata nodes have the 'name' property and relationships have the 'db-ref' property (referred as source database) [78]. Finally, different network topologies can be visualized by applying filtering criteria on StarPepDB. For example, it is possible to display a network of those peptides (central nodes) "*related to*" (edges) function "antibacterial" (metadata node) and "*compiled in"* (edges) the ADP database (metadata node).

Thus, the StarPepDB structure together with the StarPep toolbox allows building customized networks and their visualization. The visual and analytics exploration of the network by extracting some centralities measures (e.g., weighted degree or harmonic centralities) allows identifying the most relevant bioactive peptides in the network (Figure 7). Furthermore, peptide subsets can be either retrieved from the graph database by sequence identity searches or by applying filtering criteria such as peptide length, sequence motifs/patterns, physicochemical properties, and other metadata.

More recently, the same research group encoded each peptide sequence with a set of molecular descriptors bearing non-redundant structural information to set alignment-free (AF) pairwise similarity/distance relationships among the peptide nodes of the network by using a general pipeline as show in Figure 7. The resulting chemical space represented by these AF similarity networks are explored by visual inspection in combination with clustering and network science techniques [49,50].

Here, we show the chemical space network (CSN) of 174 non-redundant Anti-Biofilm Peptides (ABPs) (Figure 8) by applying the StarPep Toolbox flowchart represented above. Networks become more interpretable through visual inspection if having a community structure. Note that communities of ABPs may represent some biologically relevant regions from the chemical space where bioactive compounds reside. Hence, we have explored the CSNs by varying the similarity threshold until a well-defined community structure emerged. In this way, a final CSN has been analyzed by adjusting the similarity threshold to 0.65, at network density of 0.0068, achieving 20 ABP outliers (singletons) with atypical or unique sequences (Figure 8). Also, for each peptide discovered to be a relevant node, additional information (metadata) is available in Supplementary Materials (File S1, SI1-A and B).

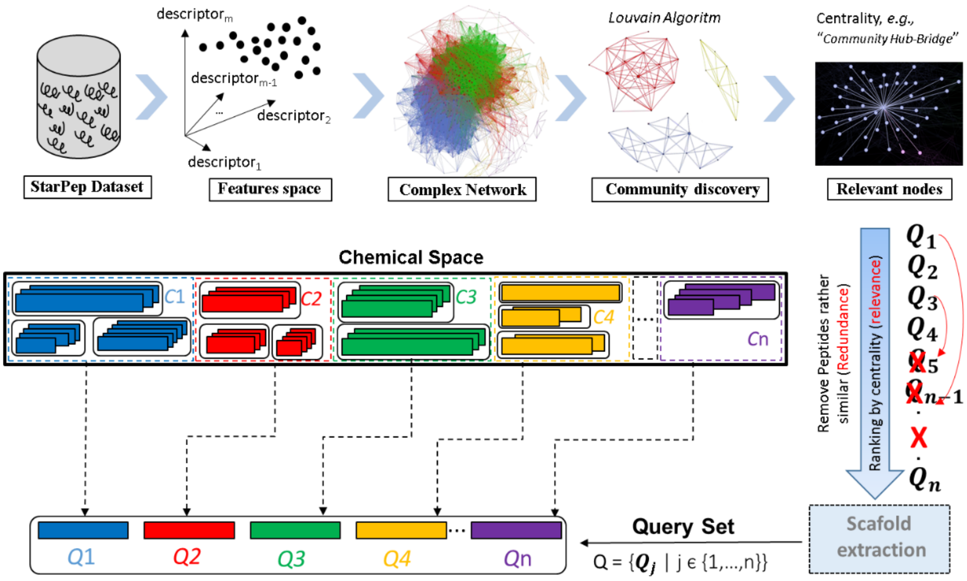

Figure 7. StarPep Toolbox flowchart. A flow diagram guiding the automatic construction and visual graph mining of similarity networks (see Figure 1 in Ref. [33]). Networks can be clustered, and communities are optimized using the Louvain method [125]. Moreover, the centrality of each node can be particularly measured by harmonic, community hub-bridge, betweenness, and weighted degree. Centrality is crucial to perform scaffold extractions because peptides are ranked according to their centrality score, and then redundant sequences are removed, prioritizing the most central. Thus, scaffold extractions depend on the type of centrality applied.

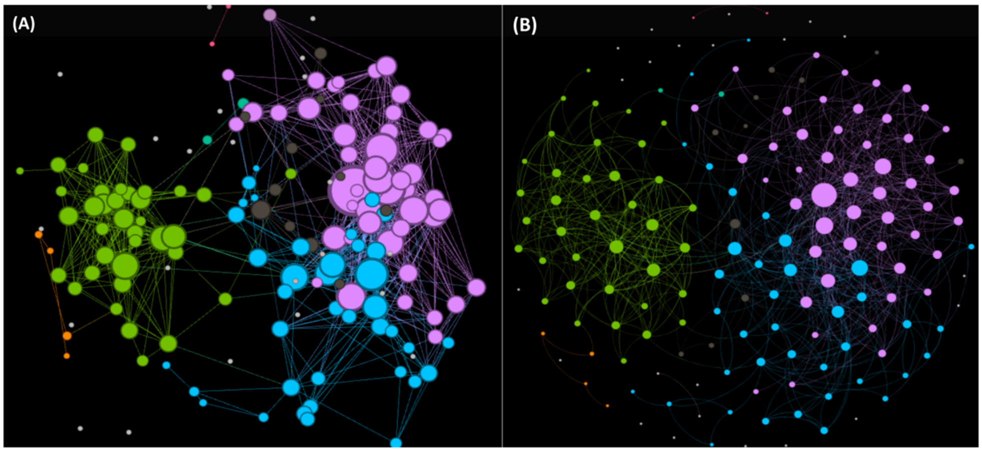

Figure 8. Visualizing the similarity network (Chemical Space Network, CSN) of a set of 174 non-redundant Anti-Biofilm Peptides (ABP_98% identity) at threshold t = 0.65 and density = 0.068, using the (**A**) three main PCAs as coordinated of each ABPs, and (**B**) Fruchtermann Reingold layout algorithm. Node colour represents the community (e.g., the biggest communities represented by cluster 3, 10 and 12 are in blue, purple and green colours, respectively), and node size symbolizes the centrality values. There are 20 ABP outliers (singletons). This figure has been created using the software starPep toolbox (version 0.8), available at http://mobiosd-hub.com/starpep; accessed on 17 March 2022.

Once a community structure is found, we rank nodes in decreasing order according to the community Harmonic centrality measure for retaining the top-*k* of the ranked list. Particularly, the top 10 exposes densely connected groups of nodes like cliques, which are defined to be complete subgraphs. These related sequences may be forming families in the chemical space of ABPs. These central peptides within each local leading community are given in SI1-B, and they may be representing sequence fragments or naturally occurring peptides that could be identified as starting structures for lead discovery. For instance, the peptide starpep_00000, starpep_05561, starpep_00361 are the most central nodes of the CSN (all in cluster 10). ABPs starpep_03668, starpep_04267, starpep_00004 and starpep_07895, starpep_12531, starpep_012529 are more central inside Communities 3 and 12, respectively (Figure 8 and SI1-C).

As can be observed in Table in SI1-C, some neighbor nodes within the communities may be representing a family of similar ABPs. Another example of closely related sequences can be seen in the 3 members of the Cluster 3 (see all ABPs in Community 3 in SI1-B). The peptides inside this cluster have the same length of 12 aas. So, it is expected that there are many ABPs with similar centrality values in the CSN, and it is advisable to extract some non-redundant ABPs from communities than just selecting the highest-ranked ones. To clearly extract central but non-redundant ABPs from each cluster (scaffold extraction, see Figure 7), we sort ABPs according to the decreasing order of their harmonic values. Then, the redundant sequences are removed at a given % of sequence identity. We have used an identity cutoff of > 35% to consider that a particular sequence is related to already-selected central ABPs and, as a consequence, removed from the CSN. Finally, the non-redundant 44 ABPs were ranked according to their decreasing values of Harmonic measure. The sorted list is given in SI1-D, and the top ranked peptides are those having relatively small similarity paths to all other nodes in the CSN.

This workflow allows the extraction of the most representative nodes/peptides describing the biologically-active chemical space (SI1-D). This representative subset can be used for multi-query similarity searches against peptide databases to retrieve all possible hits (Figure 9). The multi-query similarity search consists in using both the most central/representative nodes of the network communities and also the so-called singletons (isolated peptide nodes) as references/queries to retrieve the most similar peptides from databases by using local alignments. The best matches against the reference/query chemical space are determined by the maximum fusion rule by firstly ranking-down the similarity scores, to retrieve the best match between a query peptide and a target database and afterwards the best similarity scores are ranked for all reference peptides. Some studies have demonstrated that fusion by similarity scores and the maximum fusion rule are the best parameters for these models [126,127].

The integrated collection of 45 120 bioactive peptides registered in StarPepDB (http://mobiosd-hub.com/starpep/; accessed on 17 March 2022), that probably is the largest and most diverse bioactive peptide database to date, can be used for the discovering of central peptide nodes targeting an specific biological activity in the Chemical Space Networks (CSNs) and for taking advantage of them in multi-query similarity searches [33]. In this sense, Marrero-Ponce et al. explored different similarity networks of antiparasitic peptides (APPs) from StarPepDB to identify the most relevant and non-redundant APPs, that were later used as queries in similarity-based searches to identify potential APPs among non-labelled peptides as such in the StarPepDB. The proposed multi-query similarity search strategy outperformed state-of-the-art machine learning models aimed at APPs prediction like the AMPDiscover (https://biocom-ampdiscover.cicese.mx; accessed on 17 March 2022) and the AMPFun (http://fdblab.csie.ncu.edu.tw/AMPfun/index.html; accessed on 17 March 2022) webservers [30,31]. The methodology will also permit the design of new APPs by using the motifs found among the repurposed APPs [49]. More recently, a similar workflow using CSNs was applied to identify the most relevant tumor-homing peptides (THPs) within the StarPepDB. Such THPs were considered as queries (Qs) for multi-query similarity searches that apply a group fusion (MAX-SIM rule) model.

The resulting similarity searching models outperformed state-of-the-art tools for THPs detection, and the best one was applied to repurpose AMPs from the StarPepDB as THPs. Novel THP leads were identified as well as new motifs accounting for their TH activity [50].

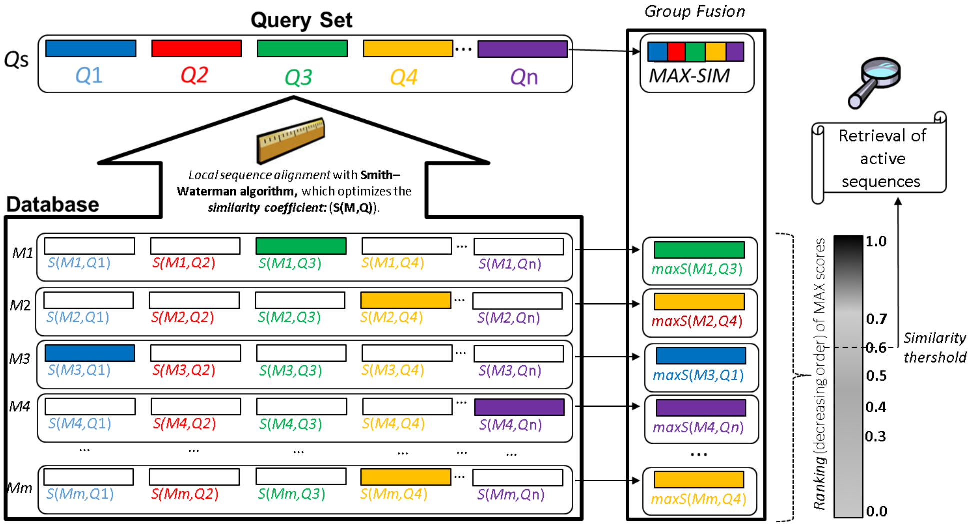

Figure 9. Schematic representation of the group fusion and similarity searching processes. Qi is a i peptide from a query/reference dataset, n is the number of peptides contained in a query dataset, S is identity coefficient between M and Q obtained by local alignment with Smith-Waterman algorithm, m is the number of peptides included in the target dataset. The similarity threshold is related to the percentage of identity.

5. Models of Sequence Evolution for the Design and Optimization of Bioactive Peptides

Several *in silico* computational approaches inspired in molecular evolution events have been applied to the design and optimization of a peptide with a promising biological activity, known in medicinal chemistry as a "leading compound". These algorithms are aimed to produce offspring peptides from a parent (hit peptide) until the "desired property" is meet according to selection criteria conducted either by ML prediction models or by biological assays (Figure 2). The offspring generation process can be iterated until reaching optimized peptidic scaffolds showing a trade-off between desirable/undesirable activities. The simulation process for generating offspring have evolved from inducing random mutations within the peptide sequence until guiding such aa substitutions under directed evolution concepts [41,128,129]. Although, algorithms inducing random mutations are commonly applied to generate sequence diversity in the peptide library, they could render unpredictable results that should be carefully analysed with selection algorithms. By contrast, computational algorithms inspired on directed mutagenesis have focused the design and optimization of "leading peptides" by guiding the generation of peptide offspring incorporating secondary structure features that influence positively on the antimicrobial activity such as amphipathic helices, kinked amphipathic helices, and other structures aimed to interact with lipid membranes [130].

Schneider et al. were the pioneers to apply simulated molecular evolution (SME) algorithms as a strategy for a rational peptide design by coupling the *in silico* generation of peptidases cleavage sites of 12 residues long to a selection mechanism represented by trained ANN [131,132]. The design was oriented to this region by generating offspring from a 12-residue sequence/peptide (parent sequence) which was iteratively mutated until meeting the best ANN quality classification metrics, used as a selection criterion of the design. The offspring sequence simulation was performed by introducing random mutations according to Gaussian-distributed probability values around the parent sequence. The mutation degree (small or large) is then conditioned by the estimation of position-specific mutability and the selected aa distance matrix [131,132]. As the position-specific mutability is averaged resulting the same for every position in the sequence; the aa mutation degree is determined by the aa substitution/scoring matrix type such the Grantham matrix [133], the Myata matrix [134], and the Risler matrix [135].

This SME approach was later applied by the same group to the optimization of anticancer peptides (ACP) aimed at improving their membranolytic activity and cell-type selectivity [51,136]. In [51], a known α-helical ACP served as the parent sequence for the generation of the offspring (ACP-derivatives). So that the generated offspring peptides retained similarity with the initial structural/property space and thus enabling a systematic optimization; the mutation function was controlled. This time the SME approach was accompanied with experimental measurements as a selection criteria or fitness objective within the optimization scheme. They used the half-effective concentration (EC_{50}) on the breast cancer cell line MCF7 and the secondary structure preferences by circular dichroism (CD) spectroscopy as experimental filters. A similar SME protocol was applied in [136] to optimize the cell-type selectivity of the highest-scored candidate toward non-cancer cells and human erythrocytes. This candidate termed AmphiArc2 peptide resulted from the screening of virtual libraries generated by more advanced algorithms incorporating secondary structures features (alpha and amphipathic helices) that influence positively on the membranolytic action [130]. AmphiArc2 was selected as a parent sequence in the SME algorithm in which the mutated sequences (offspring) are generated from it. The offspring was scored according to a fitness function, defined by the anticancer activity and selectivity with respect to non-transformed cells. The best offspring was selected as a parent for the following optimization iteration [136].

Although the SME approach and the generation of oriented libraries toward certain secondary structures, relevant for the interaction with lipid membranes, have represented a step forward in the design and optimization pipeline of AMPs and ACPs [130,136], there still room for improving the simulation of molecular evolution of the offspring peptides. In this sense, algorithms that traditionally have been used for simulating sequence evolution in the field of molecular phylogenetics were recently applied to provide more rationality to the peptide library generation [52]. These algorithms were initially developed to evaluate the accuracy of MSA and phylogenetic reconstruction tools by generating sets of related simulated protein sequences from known phylogenies. The most representative ones are: ROSE (Random Model of Sequence Evolution) [137], SIMPROT (Simulation Protein Evolution) [138], and INDELible (Insertions and Deletions Simulator) [139]. In general, they are controlled by several evolutionary parameters such as tree topology, evolutionary distance matrices, mutation rate, insertion and deletion probabilities to simulate the evolution of offspring from a parent sequence. Ruiz-Blanco et al. incorporated the ROSE algorithm into the *de novo* design pipeline of peptide inhibitors of *E. coli* ATP synthase [52,76]. As parent peptides, both the natural inhibitor (IF_1) of the mitochondrial ATP synthase and fragments of interfaces involved in protein—protein interactions between subunits of *E. coli* ATP synthase, were selected to generate peptide libraries. The residue conservation degree on these parent peptides was identified by MSAs within each class. A consensus parent peptide with its corresponding conservation scoring profile was estimated so different mutation rates to each position in the sequence could be assigned. This mutation probability vector together with a user-defined phylogenetic tree with a known topology and branch lengths guided the probabilistic function performing mutations, insertion and deletions on the parent peptide [52,76]. On the other hand, the sequence diversity of the offspring peptides in the library can be controlled by calibrating ROSE parameters against the pairwise identity [81]. A predefined binary phylogenetic tree with 1023 nodes and depth 9 implemented in ROSE was used in [52,76] for the generation of diversity-oriented libraries. The Figure 10 shows a schematic description of the ROSE algorithm.

Peptide libraries were screened by the PPI-Detect [77], an SVM-based model that predicts peptide interactions with both domains of the *E. coli* and human ATP synthases. As selection criterion, the high-scored interacting peptides with the *E. coli* ATP synthase but showing low values with the human's were subsequently evaluated by in vitro inhibition tests. At applying advanced SME algorithms involving more evolutionary models/parameters like ROSE makes easier subsequently screening steps to find lead peptides at high success rate.

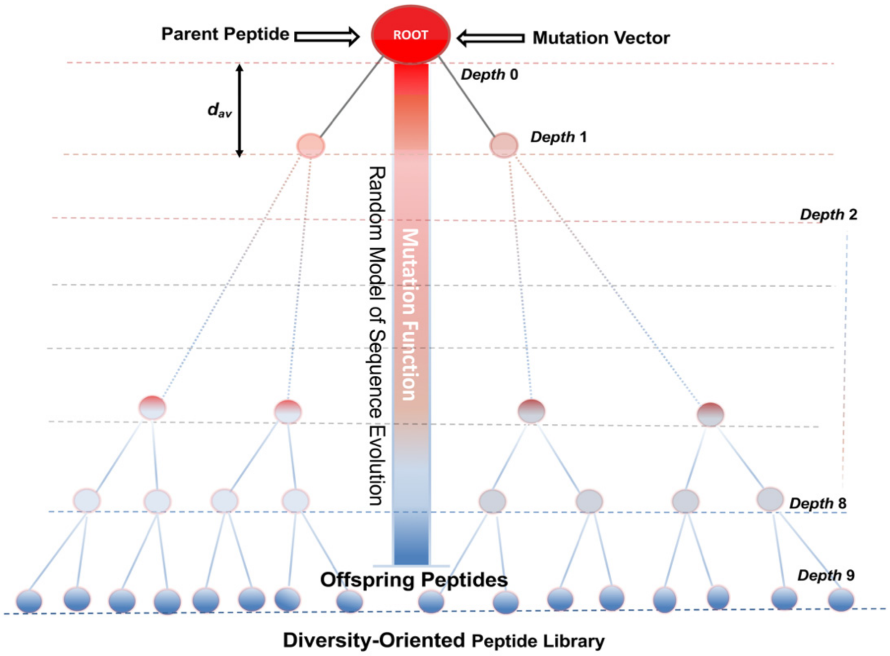

Figure 10. The binary mutation guide tree used by ROSE to mutate the parent/root peptide. The binary tree topology is determined by the number of nodes (1023), depth (9) and average distance (dav = 5–20 PAMs). Peptide library may be selected either from internal or terminal nodes of the tree. The identity percentage of the offspring peptides respect to the parent/root peptide is coloured-illustrated. Red colour means closely-related peptides to the parent while blue colour represents those distantly-related ones.

6. Considerations in the Workflow for the High-Throughput Discovering of Bioactive Peptides

6.1. Brief Comparisons between High-Throughput (HT) and Classical Methods

The classical approach for discovery of bioactive peptides has changed from analysing biological extracts/fluids to perform a wide-genomics and proteomics search. In this sense both next-generation sequencing (NGS) technologies and mass spectrometry (MS)–based proteomics combined with bioinformatic tools have provided suitable approaches for the large-scale identification of bioactive peptides outperforming the classical methods. These last ones usually include a purification step combined with bio-guided assays, which require higher amount of biomass from the subject organism. Although they can determine the biological activity of bioactive compounds relatively at high accuracy, are time-consuming and the yield of bioactive compounds is low as well as the coverage of the chemical space [140]. On the contrary, the HT analyses can be performed with around 1 cm^3 or 0.5–1 g of fresh or preserved tissues, for genomic/transcriptomic or proteomic purposes, respectively [53,141,142]. Generally, the HT methods allow covering the whole picture for potential bioactive compounds much faster. Despite HT methods usually require of powerful computational resources, both NGS and MS-based proteomics are becoming cheaper and their corresponding workflows are continuously optimized within the discovery process as well, resulting in a long-term sustainable approach [143,144]. Moreover, HT OMICs technologies yield a big amount of free public data, allowing the decentralization of the knowledge for the biodiscovery process.

Hence, the integration of OMICs approaches is more recommendable than the classical ones at the early stage of bioactive peptide discovery. However, bioassays-guided methods are still valid and complementary at advanced phases of the research [140,145].

6.2. Optimized Workflow for the Large-Scale AMPs Discovery from Profiled/Unexplored Organisms

Despite the advances in the discovery of bioactive peptides, improved protocols are still needed to increase the accuracy in both their large-scale identification and functional characterization, which is a major challenge, nowadays. Figure 11 illustrates the overall steps for the HT bioactive peptide discovery from model and unexplored organisms.

In order to analyze OMICs data released by NGS and MS-based HT proteomics, several computational/bioinformatic tools and platforms have been developed. Among them, for the *de novo* genome/transcriptome assembly we can mention, i.e., MIRA [146], Spades [147], CAP3 [148], OASES [149] and the Trinity package [150] including the *de novo* assembler and the TransDecoder for ORFs prediction (https://github.com/TransDecoder/TransDecoder/releases; accessed on 17 March 2022). Other ALL-IN-ONE licensed software like the toolbox CLC Genomic Workbench (CLC Bio-Qiagen, Aarhus, Denmark) [151] and OMICsBox (BioBam Bioinformatics, Valencia, Spain) [152], have integrated several tools for the complete workflow, including the *de novo* assemblers, custom/online/cloud functional annotation options with Blast+ [153], eggNOG [154], KEGG [155], providing as well as a set of functional analyses and statistical tests (i.e., Gene Ontology, deferential expression analyses and enrichment).

Among the NGS analyses, the RNA-seq has gained relevance because it can explore the coding regions of the genome by assembling, annotating and comparing expression profiles of the resulting transcripts [141,156]. Since elucidating the transcriptome demands lower computational cost than whole genome, and also provide useful information, its number has increasingly growth in databases. In this sense, transcriptomes from the same or related species are translated, usually with the TransDecoder or Six-Frame Translations Tool (S-FTT) (https://github.com/iracooke/protk; accessed on 17 March 2022), then annotated, and thus considered as reference database for improving protein identification in proteomics analyses from a target organism [157]. These are the grounds of proteogenomic analyses where genomic, transcriptomic and proteomic data are combined to assist the discovery of peptides from MS–based proteomic data, especially if they are not present in protein databases such as UniprotKB and other related ones (i.e, Swiss-Prot, TrEMBL and UniRef), the protein section of NCBI, Mendeley and ProteomExchange consortium [158]. On the other hand, the proteomic data can also be used to confirm gene expression [159].

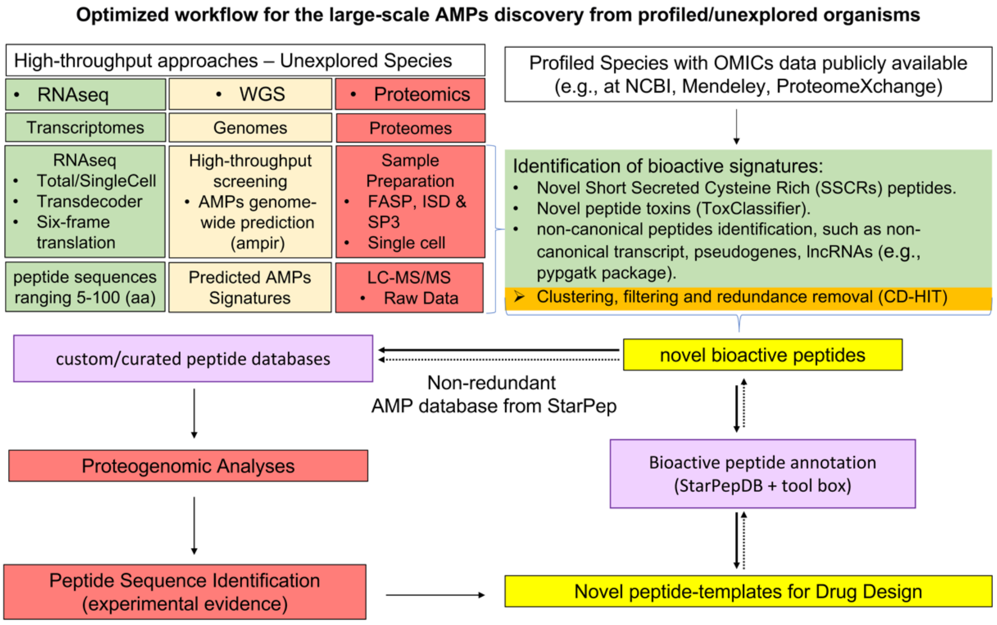
Optimized workflow for the large-scale AMPs discovery from profiled/unexplored organisms
High-throughput approaches – Unexplored Species
• RNAseq
• WGS
• Proteomics
Transcriptomes
Genomes
Proteomes
RNAseq
• Total/SingleCell
• Transdecoder
• Six-frame translation
High-throughput screening
• AMPs genome-wide prediction (ampir)
Sample Preparation
• FASP, ISD & SP3
• Single cell
peptide sequences ranging 5-100 (aa)
Predicted AMPs Signatures
LC-MS/MS
• Raw Data
Profiled Species with OMICs data publicly available (e.g., at NCBI, Mendeley, ProteomeXchange)
Identification of bioactive signatures:
• Novel Short Secreted Cysteine Rich (SSCRs) peptides.
• Novel peptide toxins (ToxClassifier).
• non-canonical peptides identification, such as non-canonical transcript, pseudogenes, lncRNAs (e.g., pypgatk package).
➢ Clustering, filtering and redundance removal (CD-HIT)
custom/curated peptide databases
novel bioactive peptides
Non-redundant
AMP database from StarPep
Proteogenomic Analyses
Bioactive peptide annotation (StarPepDB + tool box)
Peptide Sequence Identification (experimental evidence)
Novel peptide-templates for Drug Design

Figure 11. Optimized workflow for the high-throughput (HT) AMPs discovery from profiled and/or unexplored organisms. The figure summarizes the main phases in the AMPs discovery using genomic,

transcriptomic and proteomic data from profiled or underexplored organisms. The figure depicts the pipeline for de novo HT discovery from un(der)explored organisms using OMICs approaches (shown in the top-left panel), and from nucleotide and proteomic information available at public databases (top-right). Genomic information publicly available at NCBI (Genome database https://www.ncbi.nlm.nih.gov/genome/; accessed on 17 March 2022) and transcripts encoding protein sequences under 100 aa length provided by the Transcriptome Shotgun Assembly (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/; accessed on 17 March 2022), can be screened with the computational tool ampir for fast genome-wide prediction of AMPs [160]. Likewise, the remaining transcripts encoding peptides sequences ranging 5-100 aas length, usually discarded in transcriptomic analyses, can be translated with the six-frame translations tool (S-FTT) [157,161] after ORFs prediction with the TransDecoder. Considering bioactive peptides include animal toxins which are usually rich in cysteine, the aa sequences obtained with S-FFT can be either analyzed by the Proteomic toolkit (https://github.com/iracooke/protk; accessed on 17 March 2022) to identify cysteine-rich regions to discover novel Short Secreted Cysteine Rich (SSCRs) peptides, or by the Machine Learning (ML) tool ToxClassifier, that enables a simple and consistent discrimination of toxins from non-toxin sequences [162]. In addition, new tools like the pypgatk package [163] can recover a significant number of cryptic peptides of biomedical interest from pseudogenes, long non-coding RNAs (lncRNAs) and other non-canonical coding transcripts produced by alternative splicing. These filtering tools can be applied together CD-HIT [164] to screen nucleotide databases before custom and non-redundant peptide databases building for proteogenomic analyses or HT annotation. Finally, the StarPepDB with its associated tools [33] may have several roles within the presented workflow by providing non-redundant bioactive databases and also at reducing custom peptide databases with the identification of the most relevant peptides for proteogenomic analyses. Moreover, bioactive peptides detected in HT screening can be classified and clustered with StarPep in different categories according to their biomedical potential (e.g., AMPs, antitumor, antibacterial, antiparasitic, etc.).

In general, the overall proteomic approach for the discovery of bioactive peptides includes the following steps: (*i*) protein digestion, (*ii*) peptide separation, (*iii*) peptide fragmentation and MS spectra acquisition, (*iv*) peptide identification using MS spectra database by similarity searches or by *de novo* sequencing. In this sense, steps (*i*) and (*ii*) are addressed by several sample preparation protocols which selection determine the best yields/results. Specifically, for bioactive peptide discovery, it is advisable the solid-phase-enhanced sample-preparation (SP3) protocol [165] since it reaches a wider coverage of peptides than the filter-aided sample preparation (FASP) [166]; moreover, is less complicated and faster than the in-solution digestion (ISD) [167].

Besides to protocol improvements in sample processing [161], there have been advances in the peptide identification step by applying several computational strategies that have also refined their bioactivity prediction [159]. In addition to use transcriptomic data to increase peptide detection accuracy, the inclusion of custom databases is being applied to characterize the part of the proteome that remains unannotated. In this sense, composite databases have been explored for a deeper proteomic characterization of the salivary glands from *Octupus vulgaris* looking for revealing underexplored bioactive peptides/toxins from previous studies [54,157,161]. The composite database comprised data from the UniProtKB, built from *de novo* transcriptome assembly of Anterior (ASGs) and Posterior Salivary Glands (PSGs), combined with those retrieved from all transcriptomes available from the cephalopods' PSGs. In addition, a comprehensive non-redundant AMPs database [124] was also included to provide additional insights about bioactive compounds such as putative AMPs [54]. In a previous work the same AMP subset was also considered as custom database to characterize the Ascidian tunic proteome by shotgun proteomics [53]. The computational analysis of the raw data implied searches against the Uniprot database (Bacteria and Metazoan section) and the AMP database. The Ascidian tunic revealed the presence of AMPs from both eukaryotes and prokaryotes and the "Biosynthesis of antibiotics" pathway was among the most significant ones, which support this tissue as an interesting reservoir of bioactive peptides/toxins and its role on the interactions Ascidians and their associated organisms. The AMP subset integrated in these previous analyses was published by Aguilera-Mendoza and probably was the most comprehensive and non-redundant AMP database reported so far [124], that later was updated in the StarPepDB (http://mobiosd-hub.com/starpep/; accessed on 17 March 2022) [78], as mentioned above.

Other important handicaps in the workflow of proteogenomic analysis are the False Discover Rate generated at analysing large protein/peptide databases [168–170] and the probable loss of information represented by small size transcripts encoding protein fragments < 100 aas that could be discarded by the TransDecoder [54,157,161], the tool dedicated to identify candidate coding regions within transcripts generated by *de novo* RNA-Seq, and such small-sized fragments could account for bioactive peptides. In order to perform a wider proteome analysis looking for uncovered AMPs and peptide toxins in the PSGs of *O. vulgaris,* contigs discarded in previous proteogenomic analyses (<100 aas) were translated with the S-FTT and then included in the protein database [54]. To optimize further proteogenomic analyses (i.e., time of analyses, FDR), or peptide annotation, sequences redundancy should be reduced with the CD-HIT [164] since the S-FTT generates many peptides sharing high similarities that could affect the overall peptide identification when increasing the FDR [170].

Other filters within the computational pipeline to process proteomic data have been applied to refine the search of peptide toxins against both canonical and custom databases. For example, the search can be framed against those toxins/peptides having signal peptides, responsible for their transport and secretion. Signal peptides have shown to contain common features across all life kingdoms [171]. In addition, cysteine-rich secretory proteins (CRISPs), small toxins (<100 aas) commonly found within the secretions of animal venoms, can be extracted from protein databases, to enrich reference databases for increasing proteomic toxin peptides detection [172]. Besides, the custom protein/peptide database can also be screened with ML-based tools e.g., ToxClassifier, that enables simple and consistent discrimination of toxins from non-toxin sequences [162], allowing the discovery of novel toxin-like bioactive peptides. Moreover, the fast genome-wide prediction of AMPs, using the ampir R package [160] can be used in the pipeline to retrieve novel peptides with antimicrobial signatures from public nucleotide databases, *de novo* transcriptomes/genomes assemblies, or as a filtering step before using S-FTT. More recently, new tools for the creation of proteogenomic databases considering the translation of pseudogenes, long non-coding RNAs (lncRNAs) and other non-canonical coding transcripts produced by alternative splicing, have allowed the identification of a significant number of cryptic peptides that may show interesting biological activities [163].

7. Concluding Remarks

Protein features inspired on molecular descriptors from chemoinformatics have emerged as successful predictors for AMPs activities. Particularly, ProtDCal's descriptors have been recently incorporated in two RF-based webservers (AMPDiscover and ABPFinder) targeting AMPs predictions as well as their specific activities and putative bacterial targets. Moreover, ProtDCal's descriptors have been involved in the design of antibiotic peptides by predicting their interaction to druggable targets from *E. coli*.

Among the recent ML approaches, undoubtedly DNNs have been the algorithm of choice for AMPs prediction in emerging tools. However, recently it has been shown that deep learning models' performance in AMP prediction is comparable to the one of classical ML algorithms being their use mostly advisable when the performance gains justify the associated computational cost.

Currently, the network science implemented in StarPep is being applied as one of the top emerging approaches, regardless of ML, to assist the search and design of bioactive peptides through the identification of lead peptides within the known chemical space. On the other hand, methodologies that simulate sequence evolution in the phylogenetics field have been repurposed to assist the optimization of such peptide leads by generating diversity-oriented libraries which are strictly controlled by evolutionary parameters.

New considerations in analysing genomic, transcriptomic and proteomic data for AMPs discovery from either profiled or underexplored organisms are being also applied. Several filtering steps have been proposed to reduce the FDR in AMPs detection when custom databases are included, but at the same time, to encompass the highest number and diversity of peptides as possible.

8. Future Research Directions

Despite a great diversity of peptide features (classical and non-classical) that has been used in AMPs prediction/design, most of those features are sequences- or property- based; however, the 3D structural information of AMPs has not been deeply exploited for such aims [173–175]. Although experimental determinate 3D structures of AMPs are used in minor proportion than their sequences, the 3D structure prediction tools are becoming more accessible and less computational demanding when considering new advances in both software and computer architectures [176,177]. These facts will ease the gradually inclusion of 3D structural features in the prediction models.

Another alternative for the inclusion of higher structural information in AMPs encoding is the use of artificial representations, which have been commonly used in comparative analyses of DNA and proteins and in QSAR-type modelling [81]. The integration of peptide features from heterogeneous sources e.g., from pairwise alignments and peptide sequences into the same classifier could be another outlook for improving the classification rates of AMPs. The main problem is to figure out a framework to integrate them (the resulting features, not the source methodologies) into ML models training. As a clue for future research directions, the alignment- based and -free similarity measures were successfully integrated for training bigdata ML-based classifiers for orthologs detection [79]. Bigdata solutions applied to the prediction/optimization of AMPs have not been explored yet in spite of the fact that the number of AMPs has grown in databases as well as the number of features/descriptors that can be derived from them. Bigdata platforms could be applied when performing virtual screening of millions of peptides, especially if they are described with computationally demanding structural descriptors. As previously mentioned, it would be advisable that future ML models for the AMP prediction could consider the natural imbalance ratio between AMPs and non-AMPs as well as the existing one among the AMPs activities. Moreover, the prediction of AMPs activities should be addressed with fuzzy-based models since they generally show overlapping activities which are not evenly-distributed within the AMP population [23]. Therefore, the resulting predictions for AMPs activities may be scored with probability values and not only treated as a binary value. On the other hand, for peptide leads optimization, the offspring generation step is crucial for the overall process. This step generally is carried out by evolutionary algorithms that introduce structural diversity among child peptides somewhat randomly. Although these AI-based algorithms have been continuously evolving to guide such diversity in order to gain optimization efficiency; there is still room for improvements in this direction. Thus, the algorithms commonly used in phylogenetics for simulating sequence evolution could provide more rationality to the generation of offspring peptides since they have been designed with more evolutionary parameters that can be strictly controlled [52,76].

Finally, StarPep is probably the most promising methodology regardless of ML approaches, that has been reported so far. The complex network theory implemented in this tool has provided a different outlook to address several steps in peptide drug discovery process. StarPep bears particular analysis tools that have not previously reported for peptides, such as (*i*) the chemical space analysis of AMP databases by similarity networks, (*ii*) the identification of the most representative and non-redundant subset of AMPs from the original chemical space, (*iii*) the mapping of unlabelled peptide datasets on similarity networks built with the representative AMPs (*iv*) the multi-query similarity searches using representative peptides against target databases. Consequently, StarPep is becoming in a competing tool to the existing ML-based methods since it has being giving clues of improved classification rates [49,50], and because of its great potentialities for the identification and optimization of new peptide leads from either in silico generated peptide libraries or released data by the Omics techniques (Figure 11).

The effort of StarPepDB developers to gather all AMP databases in a non-redundant database [124] has shown a direct impact for the AMPs prediction tools [31]. However, the annotation quality for the reported AMPs must still be improved as well as the information on their biological or molecular targets. It is urgent that AMPs activity evaluations can be harmonized under the same protocols to construct more reliable benchmark datasets for the accuracy sake of the computational analysis tools. The diverse computational methods available for AMPs discovery are a powerful tool for the accurate design of peptide drugs. The growing availability of 3D structural descriptors and scoring functions will allow developing more effective in silico peptide drug design technologies. The assembling of ML methods with peptide-protein docking and molecular dynamics seems to be an effective alternative as well [178]. If all these aspects were considered for the computational-assisted search/design of peptide drugs, the next-generation of AMP leads will be more valuable for developing therapeutic agents to face challenging health problems such as cancer, infectious diseases and more recently, COVID-19.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/antibiotics11070936/s1. File S1—Additional data relative to the analysis of the Chemical Space Network built with Anti-Biofilm Peptides (ABPs) from StarPepDB. File SI1-A—Metadata associated to ABPs. SI1-B—Node properties of the 174 ABPs embedded in the similarity network representing the ABPs chemical space. SI1-C—More central ABPs according to network centrality measures. SI1-D—ABPs selected as queries for multi-query similarity searches.

Author Contributions: G.A.-C. and A.A. designed and structured the overall review as well as drafted the initial manuscript. D.G.-C. analyzed emerging ML algorithms for the AMP prediction. D.D.-P. gathered the new considerations in OMICs analysis/workflow for the biodiscovery of AMPs. Y.M.-P. worked on the conceptualizing of the complex networks and similarity searching methods applied to AMPs discovery. G.P.-M. and M.T. looked for the non-standard peptide features used for AMPs prediction. G.A-C. reviewed the application of algorithms simulating sequence evolution to peptide drug design. G.A.-C., D.G.-C. and A.A. participated in the review-edition of the manuscript All authors have read and agreed to the published version of the manuscript.

Funding: Yovani Marero-Ponce was supported by the USFQ Collaboration Grant (Project ID16885). This research was supported in part by the Strategic Funding UIDB/04423/2020 and UIDP/04423/2020 through national funds provided by FCT and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program—COMPETE 2020 and by National Funds through the FCT under the project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Murray, C.J.; Ikuta, K.S.; Sharara, F.; Swetschinski, L.; Aguilar, G.R.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; et al. Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet* **2022**, *399*, 629–655. [CrossRef]
2. Fair, R.J.; Tor, Y. Antibiotics and Bacterial Resistance in the 21st Century. *Perspect. Med. Chem.* **2014**, *6*, S14459. [CrossRef] [PubMed]
3. Yeaman, M.R.; Yount, N.Y. Mechanisms of Antimicrobial Peptide Action and Resistance. *Pharmacol. Rev.* **2003**, *55*, 27–55. [CrossRef] [PubMed]
4. Guevara Agudelo, A.; Muñoz Molina, M.; Navarrete Ospina, J.; Salazar Pulido, L.; Castro-Cardozo, B. New Horizons to Survive in a Post-Antibiotics Era. *J. Trop Med. Health* **2018**, *10*, JTMH-130. [CrossRef]

5. Breijyeh, Z.; Jubeh, B.; Karaman, R. Resistance of Gram-Negative Bacteria to Current Antibacterial Agents and Approaches to Resolve It. *Molecules* **2020**, *25*, 1340. [CrossRef]
6. Gohel, V.; Kamal, A. Peptides as Potential Anticancer Agents. *Curr. Top. Med. Chem.* **2019**, *19*, 1491–1511. [CrossRef]
7. Schütz, D.; Ruiz-Blanco, Y.B.; Münch, J.; Kirchhoff, F.; Sanchez-Garcia, E.; Müller, J.A. Peptide and peptide-based inhibitors of SARS-CoV-2 entry. *Adv. Drug Deliv. Rev.* **2020**, *167*, 47–65. [CrossRef]
8. Zhang, L.-J.; Gallo, R.L. Antimicrobial peptides. *Curr. Biol.* **2016**, *26*, R14–R19. [CrossRef]
9. Porto, W.F.; Pires, A.S.; Franco, O.L. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnol. Adv.* **2017**, *35*, 337–349. [CrossRef]
10. Sundararajan, V.S.; Gabere, M.N.; Pretorius, A.; Adam, S.; Christoffels, A.; Lehväslaiho, M.; Archer, J.A.C.; Bajic, V.B. DAMPD: A manually curated antimicrobial peptide database. *Nucleic Acids Res.* **2011**, *40*, D1108–D1112. [CrossRef]
11. Waghu, F.H.; Barai, R.S.; Gurung, P.; Idicula-Thomas, S. CAMP R3: A database on sequences, structures and signatures of antimicrobial peptides: Table 1. *Nucleic Acids Res.* **2016**, *44*, D1094–D1097. Available online: http://www.ncbi.nlm.nih.gov/pubmed/26467475 (accessed on 23 January 2019). [CrossRef] [PubMed]
12. Zhao, X.; Wu, H.; Lu, H.; Li, G.; Huang, Q. LAMP: A Database Linking Antimicrobial Peptides. *PLoS ONE* **2013**, *8*, e66557. [CrossRef] [PubMed]
13. Fan, L.; Sun, J.; Zhou, M.; Zhou, J.; Lao, X.; Zheng, H.; Xu, H. DRAMP: A comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **2016**, *6*, 24482. [CrossRef] [PubMed]
14. Lee, H.-T.; Lee, C.-C.; Yang, J.-R.; Lai, J.Z.C.; Chang, K.Y. A Large-Scale Structural Classification of Antimicrobial Peptides. *BioMed Res. Int.* **2015**, *2015*, 1–6. [CrossRef]
15. 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*, D288–D297. [CrossRef]
16. Meher, P.K.; Sahu, T.K.; Saini, V.; Rao, A.R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **2017**, *7*, srep42362. [CrossRef]
17. Spänig, S.; Heider, D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Min.* **2019**, *12*, 7. [CrossRef]
18. Xu, J.; Li, F.; Leier, A.; Xiang, D.; Shen, H.-H.; Lago, T.T.M.; Li, J.; Yu, D.-J.; Song, J. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings Bioinform.* **2021**, *22*, bbab083. [CrossRef]
19. Veltri, D.; Kamath, U.; Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **2018**, *34*, 2740–2747. [CrossRef]
20. Yi, H.-C.; You, Z.-H.; Zhou, X.; Cheng, L.; Li, X.; Jiang, T.-H.; Chen, Z.-H. ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. *Mol. Ther.-Nucleic Acids* **2019**, *17*, 1–9. [CrossRef]
21. Chen, J.; Cheong, H.H.; Siu, S.W.I. xDeep-AcPEP: Deep Learning Method for Anticancer Peptide Activity Prediction Based on Convolutional Neural Network and Multitask Learning. *J. Chem. Inf. Model.* **2021**, *61*, 3789–3803. [CrossRef] [PubMed]
22. Boone, K.; Camarda, K.; Spencer, P.; Tamerler, C. Antimicrobial peptide similarity and classification through rough set theory using physicochemical boundaries. *BMC Bioinform.* **2018**, *19*, 1–10. [CrossRef] [PubMed]
23. Chharia, A.; Upadhyay, R.; Kumar, V. Novel fuzzy approach to Antimicrobial Peptide Activity Prediction: A tale of limited and imbalanced data that models won't hear; 2021. In Proceedings of the NeurIPS 2021 AI for Science Workshop, Vancouver, BC, Canada, 13 December 2021.
24. Wang, G.; Li, X.; Wang, Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087–D1093. [CrossRef]
25. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct. Funct. Bioinform.* **2001**, *43*, 246–255. [CrossRef]
26. Xiao, X.; Wang, P.; Lin, W.-Z.; Jia, J.-H.; Chou, K.-C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177. [CrossRef] [PubMed]
27. Lin, W.; Xu, D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. *Bioinformatics* **2016**, *32*, 3745–3752. [CrossRef]
28. Gull, S.; Shamim, N.; Minhas, F. AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Comput. Biol. Med.* **2019**, *107*, 172–181. [CrossRef]
29. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
30. Chung, C.-R.; Kuo, T.-R.; Wu, L.-C.; Lee, T.-Y.; Horng, J.-T. Characterization and identification of antimicrobial peptides with different functional activities. *Brief. Bioinform.* **2019**, *21*, 1098–1114. [CrossRef]
31. Pinacho-Castellanos, S.A.; García-Jacas, C.R.; Gilson, M.K.; Brizuela, C.A. Alignment-Free Antimicrobial Peptide Predictors: Improving Performance by a Thorough Analysis of the Largest Available Data Set. *J. Chem. Inf. Model.* **2021**, *61*, 3141–3157. [CrossRef]
32. Ruiz-Blanco, Y.B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinform.* **2015**, *16*, 162. [CrossRef]

33. Aguilera-Mendoza, L.; Marrero-Ponce, Y.; García-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*, 1–23. [CrossRef]
34. Kavousi, K.; Bagheri, M.; Behrouzi, S.; Vafadar, S.; Atanaki, F.F.; Lotfabadi, B.T.; Ariaeenejad, S.; Shockravi, A.; Moosavi-Movahedi, A.A. IAMPE: NMR-Assisted Computational Prediction of Antimicrobial Peptides. *J. Chem. Inf. Model.* **2020**, *60*, 4691–4701. [CrossRef] [PubMed]
35. Joseph, S.; Karnik, S.; Nilawe, P.; Jayaraman, V.K.; Idicula-Thomas, S. ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1535–1538. [CrossRef]
36. Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S.W.I. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* **2018**, *8*, 1–10. [CrossRef]
37. Lawrence, T.J.; Carper, D.L.; Spangler, M.K.; Carrell, A.A.; Rush, T.A.; Minter, S.J.; Weston, D.J.; Labbe, J.L. amPEPpy 1.0: A portable and accurate antimicrobial peptide prediction tool. *Bioinformatics* **2021**, *37*, 2058–2060. [CrossRef] [PubMed]
38. Veltri, D.P. A Computational and Statistical Framework for Screening Novel Antimicrobial Peptides. Ph.D. Thesis, George Mason University, Fairfax County, VA, USA, 2015.
39. García-Jacas, C.R.; Pinacho-Castellanos, S.A.; García-González, L.A.; Brizuela, C.A. Do deep learning models make a difference in the identification of antimicrobial peptides? *Brief. Bioinform.* **2022**, *23*, bbac094. [CrossRef]
40. Wong, K.-C. Evolutionary algorithms: Concepts, designs, and applications in bioinformatics. In *Nature-Inspired Computing: Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2017; pp. 111–137.
41. Bozovičar, K.; Bratkovič, T. Evolving a Peptide: Library Platforms and Diversification Strategies. *Int. J. Mol. Sci.* **2019**, *21*, 215. [CrossRef]
42. Yoshida, M.; Hinkley, T.; Tsuda, S.; Abul-Haija, Y.; McBurney, R.T.; Kulikov, V.; Mathieson, J.S.; Reyes, S.G.; Castro, M.D.; Cronin, L. Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* **2018**, *4*, 533–543. [CrossRef]
43. Barigye, S.J.; Garcia de la Vega, J.M.; Perez-Castillo, Y.; Castillo-Garit, J.A. Evolutionary algorithm-based generation of optimum peptide sequences with dengue virus inhibitory activity. *Future Med. Chem.* **2021**, *13*, 993–1000. [CrossRef]
44. Fjell, C.D.; Jenssen, H.; Cheung, W.; Hancock, R.; Cherkasov, A. Optimization of Antibacterial Peptides by Genetic Algorithms and Cheminformatics. *Chem. Biol. Drug Des.* **2010**, *77*, 48–56. [CrossRef]
45. Fjell, C.D.; Hiss, J.A.; Hancock, R.E.W.; Schneider, G. Designing antimicrobial peptides: Form follows function. *Nat. Rev. Drug Discov.* **2011**, *11*, 37–51. [CrossRef]
46. Aronica, P.G.; Reid, L.M.; Desai, N.; Li, J.; Fox, S.J.; Yadahalli, S.; Essex, J.W.; Verma, C.S. Computational Methods and Tools in Antimicrobial Peptide Research. *J. Chem. Inf. Model.* **2021**, *61*, 3172–3196. [CrossRef]
47. Ng, X.Y.; Rosdi, B.A.; Shahrudin, S. Prediction of Antimicrobial Peptides Based on Sequence Alignment and Support Vector Machine-Pairwise Algorithm Utilizing LZ-Complexity. *BioMed Res. Int.* **2015**, *2015*, 1–13. [CrossRef]
48. Boone, K.; Wisdom, C.; Camarda, K.; Spencer, P.; Tamerler, C. Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides. *BMC Bioinform.* **2021**, *22*, 1–17. [CrossRef]
49. Ayala-Ruano, S.; Marrero-Ponce, Y.; Aguilera-Mendoza, L.; Pérez, N.; Agüero-Chapin, G.; Antunes, A.; Aguilar, A.C. Exploring the Chemical Space of Antiparasitic Peptides and Discovery of New Promising Leads through a Novel Approach based on Network Science and Similarity Searching. *ChemRxiv* **2021**. [CrossRef]
50. Romero, M.; Marrero-Ponce, Y.; Rodríguez, H.; Agüero-Chapin, G.; Antunes, A.; Aguilera-Mendoza, L.; Martinez-Rios, F. A Novel Network Science and Similarity-Searching-Based Approach for Discovering Potential Tumor-Homing Peptides from Antimicrobials. *Antibiotics* **2022**, *11*, 401. [CrossRef]
51. Neuhaus, C.S.; Gabernet, G.; Steuer, C.; Root, K.; Hiss, J.A.; Zenobi, R.; Schneider, G. Simulated Molecular Evolution for Anticancer Peptide Design. *Angew. Chem. Int. Ed.* **2018**, *58*, 1674–1678. [CrossRef]
52. Ruiz-Blanco, Y.B.; Ávila-Barrientos, L.P.; Hernández-García, E.; Antunes, A.; Agüero-Chapin, G.; García-Hernández, E. Engineering protein fragments via evolutionary and protein–protein interaction algorithms: De novo design of peptide inhibitors for $F_{O}F_{1}$-ATP synthase. *FEBS Lett.* **2020**, *595*, 183–194. [CrossRef]
53. Matos, A.; Domínguez-Pérez, D.; Almeida, D.; Agüero-Chapin, G.; Campos, A.; Osório, 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*, 362. [CrossRef] [PubMed]
54. Almeida, D.; Domínguez-Pérez, D.; Matos, A.; Agüero-Chapin, G.; Osório, 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* **2020**, *9*, 757. [CrossRef]
55. Agüero-Chapin, G.; Pérez-Machado, G.; Molina-Ruiz, R.; Pérez-Castillo, Y.; Morales-Helguera, A.; Vasconcelos, V.; Antunes, A. TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. *Amino Acids* **2010**, *40*, 431–442. [CrossRef] [PubMed]
56. Speck-Planche, A.; Kleandrova, V.V.; Ruso, J.M.; Cordeiro, M.N.D.S. First Multitarget Chemo-Bioinformatic Model To Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *J. Chem. Inf. Model.* **2016**, *56*, 588–598. [CrossRef] [PubMed]

57. De Armas, R.R.; Díaz, H.G.; Molina, R.; González, M.P.; Uriarte, E. Stochastic-based descriptors studying peptides biological properties: Modeling the bitter tasting threshold of dipeptides. *Bioorganic Med. Chem.* **2004**, *12*, 4815–4822. [CrossRef] [PubMed]
58. Kleandrova, V.V.; Ruso, J.M.; Speck-Planche, A.; Cordeiro, M.N.D.S. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb. Sci.* **2016**, *18*, 490–498. [CrossRef]
59. Estrada, E. Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844–849. [CrossRef]
60. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **2006**, *56*, 237–248.
61. Agüero-Chapin, G.; Molina-Ruiz, R.; Pérez-Machado, G.; Vasconcelos, V.; Rodríguez-Negrin, Z.; Antunes, A. TI2BioP—Topological Indices to BioPolymers. A Graphical–Numerical Approach for Bioinformatics. In *Recent Advances in Biopolymers*; IntechOpen: Zagreb, Croatia, 2016.
62. González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E.; González-Díaz, H. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750–778. [CrossRef]
63. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20. [CrossRef]
64. Randić, M. Graph theoretical approach to structure-activity studies: Search for optimal antitumor compounds. *Prog. Clin. Biol. Res.* **1985**, *172*, 309–318.
65. Moreau, G.; Broto, P. The Autocorrelation of a topological structure. A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359–360.
66. Balaban, A.T.; Beteringhe, A.; Constantinescu, T.; Filip, P.A.; Ivanciuc, O. Four New Topological Indices Based on the Molecular Path Code. *J. Chem. Inf. Model.* **2007**, *47*, 716–731. [CrossRef]
67. Hall, P.R.; Malone, L.; Sillerud, L.O.; Ye, C.; Hjelle, B.L.; Larson, R.S. Characterization and NMR Solution Structure of a Novel Cyclic Pentapeptide Inhibitor of Pathogenic Hantaviruses. *Chem. Biol. Drug Des.* **2007**, *69*, 180–190. [CrossRef]
68. Kier, L.B.; Hall, L.H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *07*, 801–807. [CrossRef]
69. Ivanciuc, O. Building–Block Computation of the Ivanciuc–Balaban Indices for the Virtual Screening of Combinatorial Libraries. *Internet Electron. J. Mol. Des.* **2002**, *1*, 1–9.
70. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, 1st ed.; Wiley-VCH: Mannheim, Germany, 2000; Volume 1, p. 667.
71. Estrada, E. Characterization of the folding degree of proteins. *Bioinformatics* **2002**, *18*, 697–704. [CrossRef]
72. Estrada, E. Characterization of the amino acid contribution to the folding degree of proteins. *Proteins: Struct. Funct. Bioinform.* **2004**, *54*, 727–737. [CrossRef]
73. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, A.M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481–2491. [CrossRef]
74. Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: Protein domains identifier. *Nucleic Acids Res.* **2005**, *33*, W116–W120. [CrossRef]
75. Molina, R.; Agüero-Chapin, G.; Pérez-González, M. *TI2BioP (Topological Indices to BioPolymers) Version 2.0*; Molecular Simulation and Drug Design (MSDD): Chemical Bioactives Center, Central University of Las Villas, Santa Clara, Cuba, 2011.
76. Avila-Barrientos, L.P.; Cofas-Vargas, L.F.; Agüero-Chapin, G.; Hernández-García, E.; Ruiz-Carmona, S.; Valdez-Cruz, N.A.; Trujillo-Roldán, M.; Weber, J.; Ruiz-Blanco, Y.B.; Barril, X.; et al. Computational Design of Inhibitors Targeting the Catalytic β Subunit of Escherichia coli FOF1-ATP Synthase. *Antibiotics* **2022**, *11*, 557. [CrossRef]
77. Romero-Molina, S.; Ruiz-Blanco, Y.B.; Harms, M.; Münch, J.; Sanchez-Garcia, E. PPI-Detect: A support vector machine model for sequence-based prediction of protein-protein interactions. *J. Comput. Chem.* **2019**, *40*, 1233–1242. [CrossRef]
78. 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*, 4739–4747. [CrossRef]
79. Galpert, D.; Fernández, A.; Herrera, F.; Antunes, A.; Molina-Ruiz, R.; Agüero-Chapin, G. Surveying alignment-free features for Ortholog detection in related yeast proteomes by using supervised big data classifiers. *BMC Bioinform.* **2018**, *19*, 166. [CrossRef]
80. Agüero-Chapin, G.; Molina-Ruiz, R.; Maldonado, E.; de la Riva, G.; Sánchez-Rodríguez, A.; Vasconcelos, V.; Antunes, A. Exploring the adenylation domain repertoire of nonribosomal peptide synthetases using an ensemble of sequence-search methods. *PLoS ONE* **2013**, *8*, e65926. [CrossRef]
81. Agüero-Chapin, G.; Galpert, D.; Molina-Ruiz, R.; Ancede-Gallardo, E.; Pérez-Machado, G.; De la Riva, G.A.; Antunes, A. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* **2019**, *10*, 26. [CrossRef]
82. Borozan, I.; Watt, S.; Ferretti, V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics* **2015**, *31*, 1396–1404. [CrossRef]
83. Empel, A.; Ziv, J. On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. [CrossRef]
84. Wang, P.; Hu, L.; Liu, G.; Jiang, N.; Chen, X.; Xu, J.; Zheng, W.; Li, L.; Tan, M.; Chen, Z.; et al. Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods. *PLoS ONE* **2011**, *6*, e18476. [CrossRef]
85. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
86. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]

87. Lin, Y.; Cai, Y.; Liu, J.; Lin, C.; Liu, X. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinform.* **2019**, *20*, 1–10. [CrossRef]
88. 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*, 1085–1095. [CrossRef]
89. Lertampaiporn, S.; Vorapreeda, T.; Hongsthong, A.; Thammarongtham, C. Ensemble-AMPPred: Robust AMP Prediction and Recognition Using the Ensemble Learning Method with a New Hybrid Feature for Differentiating AMPs. *Genes* **2021**, *12*, 137. [CrossRef]
90. Yu, Q.; Dong, Z.; Fan, X.; Zong, L.; Li, Y. HMD-AMP: Protein Language-Powered Hierarchical Multi-label Deep Forest for Annotating Antimicrobial Peptides. *bioRxiv* **2021**. [CrossRef]
91. Chen, X.; Li, C.; Bernards, M.T.; Shi, Y.; Shao, Q.; He, Y. Sequence-based peptide identification, generation, and property prediction with deep learning: A review. *Mol. Syst. Des. Eng.* **2021**, *6*, 406–428. [CrossRef]
92. Wan, F.; Kontogiorgos-Heintz, D.; de la Fuente-Nunez, C. Deep generative models for peptide design. *Digit. Discov.* **2022**, *1*, 195–208. [CrossRef]
93. Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; dos Santos, C.; Chen, P.-Y.; et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **2021**, *5*, 613–623. [CrossRef]
94. Van Oort, C.M.; Ferrell, J.B.; Remington, J.M.; Wshah, S.; Li, J. AMPGAN v2: Machine Learning-Guided Design of Antimicrobial Peptides. *J. Chem. Inf. Model.* **2021**, *61*, 2198–2207. [CrossRef]
95. Das, P.; Wadhawan, K.; Chang, O.; Sercu, T.; Santos, C.N.D.; Riemer, M.; Padhi, I.; Chenthamarakshan, V.; Mojsilovic, A. PepCVAE: Semi-Supervised Targeted Design of Antimicrobial Peptide Sequences. *arXiv* **2018**, arXiv:1810.07743.
96. Dean, S.N. Variational Autoencoder for the Generation of New Antimicrobial Peptides. *ACS Omega* **2021**, *5*, 20746–20754. [CrossRef]
97. Witten, J.; Witten, Z. Deep learning regression model for antimicrobial peptide design. *bioRxiv* **2019**. [CrossRef]
98. Lee, B.; Shin, M.K.; Hwang, I.-W.; Jung, J.; Shim, Y.J.; Kim, G.W.; Kim, S.T.; Jang, W.; Sung, J.-S. A Deep Learning Approach with Data Augmentation to Predict Novel Spider Neurotoxic Peptides. *Int. J. Mol. Sci.* **2021**, *22*, 12291. [CrossRef]
99. Wang, C.; Garlick, S.; Zloh, M. Deep Learning for Novel Antimicrobial Peptide Design. *Biomolecules* **2021**, *11*, 471. [CrossRef] [PubMed]
100. Bin Hafeez, A.; Jiang, X.; Bergen, P.J.; Zhu, Y. Antimicrobial Peptides: An Update on Classifications and Databases. *Int. J. Mol. Sci.* **2021**, *22*, 11691. [CrossRef] [PubMed]
101. Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H.K.; Wong, K.H.; Siu, S.W. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther.-Nucleic Acids* **2020**, *20*, 882–894. [CrossRef] [PubMed]
102. Babgi, B.A.; Alsayari, J.H.; Davaasuren, B.; Emwas, A.-H.; Jaremko, M.; Abdellattif, M.H.; Hussien, M.A. Synthesis, structural studies, and anticancer properties of [CuBr (PPh3) 2 (4,6-dimethyl-2-thiopyrimidine-S]. *Crystals* **2021**, *11*, 688. [CrossRef]
103. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
104. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **1990**, *183*, 63–98. [CrossRef]
105. Hammami, R.; Zouhir, A.; Ben Hamida, J.; Fliss, I. BACTIBASE: A new web-accessible database for bacteriocin characterization. *BMC Microbiol.* **2007**, *7*, 89. [CrossRef]
106. De Jong, A.; Van Hijum, S.A.F.T.; Bijlsma, J.J.E.; Kok, J.; Kuipers, O.P. BAGEL: A web-based bacteriocin genome mining tool. *Nucleic Acids Res.* **2006**, *34*, W273–W279. [CrossRef]
107. Mulvenna, J.; Mylne, J.; Bharathi, R.; Burton, R.; Shirley, N.; Fincher, G.B.; Anderson, M.; Craik, D.J. Discovery of Cyclotide-Like Protein Sequences in Graminaceous Crop Plants: Ancestral Precursors of Circular Proteins? *Plant Cell* **2006**, *18*, 2134–2144. [CrossRef]
108. Porto, W.F.; Silva, O.N.; Franco, O.L. Prediction and rational design of antimicrobial peptides. In *Protein Structure*; IntechOpen: Zagreb, Croatia, 2012.
109. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763. [CrossRef]
110. Thompson, K. Programming Techniques: Regular expression search algorithm. *Commun. ACM* **1968**, *11*, 419–422. [CrossRef]
111. Jonassen, I. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* **1997**, *13*, 509–522. [CrossRef]
112. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37. [CrossRef]
113. Sigrist, C.J.A.; de Castro, E.; Cerutti, L.; Cuche, B.A.; Hulo, N.; Bridge, A.; Bougueleret, L.; Xenarios, I. New and continuing developments at PROSITE. *Nucleic Acids Res.* **2012**, *41*, D344–D347. [CrossRef]
114. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [CrossRef]
115. Hammami, R.; Zouhir, A.; Le Lay, C.; Ben Hamida, J.; Fliss, I. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiol.* **2010**, *10*, 22. [CrossRef]
116. Fjell, C.D.; Hancock, R.E.W.; Cherkasov, A. AMPer: A database and an automated discovery tool for antimicrobial peptides. *Bioinformatics* **2007**, *23*, 1148–1155. [CrossRef]

117. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef]
118. Gille, C.; Goede, A.; Preißner, R.; Rother, K.; Frömmel, C. Conservation of substructures in proteins: Interfaces of secondary structural elements in proteasomal subunits. *J. Mol. Biol.* **2000**, *299*, 1147–1154. [CrossRef]
119. Lee, J.; Wu, S.; Zhang, Y. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*; Rigden, D.J., Ed.; Springer: Dordrecht, The Netherlands, 2009; pp. 3–25.
120. Eswar, N.; Webb, B.; Marti-Renom, M.A.; Madhusudhan, M.; Eramian, D.; Shen, M.y.; Pieper, U.; Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinform.* **2006**, *15*, 5–6. [CrossRef]
121. Hammami, R.; Fliss, I. Current trends in antimicrobial agent research: Chemo- and bioinformatics approaches. *Drug Discov. Today* **2010**, *15*, 540–546. [CrossRef]
122. Torrent, M.; Di Tommaso, P.; Pulido, D.; Nogués, M.V.; Notredame, C.; Boix, E.; Andreu, D. AMPA: An automated web server for prediction of protein antimicrobial regions. *Bioinformatics* **2011**, *28*, 130–131. [CrossRef]
123. Notredame, C.; Higgins, D.; Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217. [CrossRef]
124. 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*, 2553–2559. [CrossRef]
125. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
126. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053. [CrossRef]
127. Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185. [CrossRef]
128. Marasco, D.; Perretta, G.; Sabatella, M.; Ruvo, M. Past and future perspectives of synthetic peptide libraries. *Curr. Protein Pept. Sci.* **2008**, *9*, 447–467. [CrossRef]
129. Irving, M.B.; Pan, O.; Scott, J.K. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. *Curr. Opin. Chem. Biol.* **2001**, *5*, 314–324. [CrossRef]
130. Müller, A.; Gabernet, G.; Hiss, J.A.; Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **2017**, *33*, 2753–2755. [CrossRef] [PubMed]
131. Schneider, G.; Wrede, P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: De novo design of an idealized leader peptidase cleavage site. *Biophys. J.* **1994**, *66*, 335–344. [CrossRef]
132. Schneider, G.; Schuchhardt, J.; Wrede, P. Peptide design in machina: Development of artificial mitochondrial protein precursor cleavage sites by simulated molecular evolution. *Biophys. J.* **1995**, *68*, 434–447. [CrossRef]
133. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862–864. [CrossRef]
134. Miyata, T.; Miyazawa, S.; Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **1979**, *12*, 219–236. [CrossRef] [PubMed]
135. Risler, J.; Delorme, M.; Delacroix, H.; Henaut, A. Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix. *J. Mol. Biol.* **1988**, *204*, 1019–1029. [CrossRef]
136. Gabernet, G.; Gautschi, D.; Müller, A.T.; Neuhaus, C.S.; Armbrecht, L.; Dittrich, P.S.; Hiss, J.A.; Schneider, G. In silico design and optimization of selective membranolytic anticancer peptides. *Sci. Rep.* **2019**, *9*, 1–11. [CrossRef]
137. Stoye, J.; Evers, D.; Meyer, F. Rose: Generating sequence families. *Bioinformatics* **1998**, *14*, 157–163. [CrossRef]
138. Pang, A.; Smith, A.D.; Nuin, P.A.; Tillier, E.R. SIMPROT: Using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinform.* **2005**, *6*, 236. [CrossRef]
139. Fletcher, W.; Yang, Z. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Mol. Biol. Evol.* **2009**, *26*, 1879–1888. [CrossRef]
140. Bosso, M.; Ständker, L.; Kirchhoff, F.; Münch, J. Exploiting the human peptidome for novel antimicrobial and anticancer agents. *Bioorganic Med. Chem.* **2018**, *26*, 2719–2726. [CrossRef] [PubMed]
141. Domínguez-Pérez, D.; Durban, J.; Agüero-Chapin, G.; López, J.T.; Molina-Ruiz, R.; Almeida, D.; Calvete, J.J.; Vasconcelos, V.; Antunes, A. The Harderian gland transcriptomes of Caraiba andreae, Cubophis cantherigerus and Tretanorhinus variabilis, three colubroid snakes from Cuba. *Genomics* **2018**, *111*, 1720–1727. [CrossRef] [PubMed]
142. Mayr, L.M.; Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580–588. [CrossRef] [PubMed]
143. Prentis, P.J.; Pavasovic, A.; Norton, R.S. Sea Anemones: Quiet Achievers in the Field of Peptide Toxins. *Toxins* **2018**, *10*, 36. [CrossRef] [PubMed]
144. Holford, M.; Daly, M.; King, G.F.; Norton, R.S. Venoms to the rescue. *Science* **2018**, *361*, 842–844. [CrossRef]
145. Rodríguez, A.A.; Otero-González, A.; Ghattas, M.; Ständker, L. Discovery, Optimization, and Clinical Application of Natural Antimicrobial Peptides. *Biomedicines* **2021**, *9*, 1381. [CrossRef]
146. Chevreux, B. MIRA: An Automated Genome and EST Assembler. Ph.D. Thesis, Ruprecht-Karls University, Heidelberg, Germany, 2007.
147. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]

148. Huang, X.; Madan, A. CAP3: A DNA Sequence Assembly Program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef]
149. Schulz, M.H.; Zerbino, D.R.; Vingron, M.; Birney, E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **2012**, *28*, 1086–1092. [CrossRef]
150. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.D.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
151. Sequencing, H. CLC Genomics Workbench. 2011. Available online: https://research.ncsu.edu/gsl/bioinformatic-resources/clc/ (accessed on 17 March 2022).
152. Bioinformatics, B.; Valencia, S. OmicsBox-Bioinformatics made easy. *March* **2019**, *3*, 2019.
153. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
154. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.V.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2018**, *47*, D309–D314. [CrossRef] [PubMed]
155. Mitchell, A.L.; Attwood, T.K.; Babbitt, P.C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S.D.; Chang, H.-Y.; El-Gebali, S.; Fraser, M.I.; et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **2019**, *47*, D351–D360. [CrossRef] [PubMed]
156. Domínguez-Pérez, D.; Martins, J.C.; Almeida, D.; Costa, P.R.; Vasconcelos, V.; Campos, A. Transcriptomic Profile of the Cockle *Cerastoderma edule* Exposed to Seasonal Diarrhetic Shellfish Toxin Contamination. *Toxins* **2021**, *13*, 784. [CrossRef] [PubMed]
157. Fingerhut, L.C.H.W.; Strugnell, J.M.; Faou, P.; Labiaga, 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*, 3866–3876. [CrossRef]
158. Deutsch, E.W.; Bandeira, N.; Sharma, V.; Perez-Riverol, Y.; Carver, J.J.; Kundu, D.J.; García-Seisdedos, D.; Jarnuczak, A.F.; Hewapathirana, S.; Pullman, B.S.; et al. The ProteomeXchange consortium in 2020: Enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **2020**, *48*, D1145–D1152. [CrossRef]
159. Nesvizhskii, A. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114–1125. [CrossRef]
160. Fingerhut, L.C.H.W.; Miller, D.J.; Strugnell, J.M.; Daly, N.L.; Cooke, I.R. ampir: An R package for fast genome-wide prediction of antimicrobial peptides. *Bioinformatics* **2020**, *36*, 5262–5263. [CrossRef]
161. 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*, 110. [CrossRef]
162. Gacesa, R.; Barlow, D.; Long, P.F. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Comput. Sci.* **2016**, *2*, e90. [CrossRef]
163. Umer, H.M.; Audain, E.; Zhu, Y.; Pfeuffer, J.; Sachsenberg, T.; Lehtiö, J.; Branca, R.M.; Perez-Riverol, Y. Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. *Bioinformatics* **2022**, *38*, 1470–1472. [CrossRef] [PubMed]
164. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
165. Hughes, C.S.; Moggridge, S.; Müller, T.; Sorensen, P.H.; Morin, G.B.; Krijgsveld, J. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **2019**, *14*, 68–85. [CrossRef]
166. Wiśniewski, J.R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6*, 359–362. [CrossRef] [PubMed]
167. León, I.R.; Schwämmle, V.; Jensen, O.N.; Sprenger, R.R. Quantitative Assessment of In-solution Digestion Efficiency Identifies Optimal Protocols for Unbiased Protein Analysis. *Mol. Cell. Proteom.* **2013**, *12*, 2992–3005. [CrossRef] [PubMed]
168. Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinform.* **2012**, *13*, S2. [CrossRef]
169. The, M.; MacCoss, M.J.; Noble, W.S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1719–1727. [CrossRef]
170. Käll, L.; Storey, J.D.; MacCoss, M.J.; Noble, W.S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *J. Proteome Res.* **2007**, *7*, 40–44. [CrossRef]
171. Bhandari, B.K.; Gardner, P.P.; Lim, C.S. Razor: Annotation of signal peptides from toxins. *bioRxiv* **2021**. [CrossRef]
172. Maxwell, M.; Undheim, E.A.B.; Mobli, M. Secreted Cysteine-Rich Repeat Proteins "SCREPs": A Novel Multi-Domain Architecture. *Front. Pharmacol.* **2018**, *9*, 1333. [CrossRef]
173. Liu, S.; Bao, J.; Lao, X.; Zheng, H. Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides. *Sci. Rep.* **2018**, *8*, 1–12. [CrossRef] [PubMed]
174. Kumar, V.; Kumar, R.; Agrawal, P.; Patiyal, S.; Raghava, G.P. A Method for Predicting Hemolytic Potency of Chemically Modified Peptides From Its Structure. *Front. Pharmacol.* **2020**, *11*, 54. [CrossRef]
175. Zhao, Y.; Wang, S.; Fei, W.; Feng, Y.; Shen, L.; Yang, X.; Wang, M.; Wu, M. Prediction of Anticancer Peptides with High Efficacy and Low Toxicity by Hybrid Model Based on 3D Structure of Peptides. *Int. J. Mol. Sci.* **2021**, *22*, 5630. [CrossRef] [PubMed]

176. Zhong, B.; Su, X.; Wen, M.; Zuo, S.; Hong, L.; Lin, J. Parafold: Paralleling alphafold for large-scale predictions. In Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region Workshops, Kobe, Japan & Online, 12–14 January 2022; pp. 1–9.
177. Contreras-Torres, E.; Marrero-Ponce, Y.; Terán, J.E.; García-Jacas, C.R.; Brizuela, C.A.; Sánchez-Rodríguez, J.C. *MuLiMs-MCoMPAs*: A Novel Multiplatform Framework to Compute Tensor Algebra-Based Three-Dimensional Protein Descriptors. *J. Chem. Inf. Model.* **2019**, *60*, 1042–1059. [CrossRef] [PubMed]
178. Torres, M.D.T.; de la Fuente-Nunez, C. Reprogramming biological peptides to combat infectious diseases. *Chem. Commun.* **2019**, *55*, 15020–15032. [CrossRef]