OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf065
https://doi.org/10.1093/bib/bbaf065
Review

A review of neural networks for metagenomic binning

Jair Herazo-Álvarez [1,2,‡], Marco Mora [2,3,*,‡], Sara Cuadros-Orellana [2,4,‡], Karina Vilches-Ponce [2], Ruber Hernández-García [2,3]

[1]Doctorado en Modelamiento Matemático Aplicado, Universidad Católica del Maule, Talca, Maule 3480564, Chile
[2]Laboratory of Technological Research in Pattern Recognition (LITRP), Universidad Católica del Maule, Talca, Maule 3480564, Chile
[3]Departamento de Computación e Industrias, Facultad de Ciencias de la Ingeniería, Universidad Católica del Maule, Talca, Maule 3480564, Chile
[4]Centro de Biotecnología de los Recursos Naturales (CENBio), Universidad Católica del Maule, Talca, Maule 3480564, Chile
*Corresponding author. Departamento de Computación e Industrias, Facultad de Ciencias de la Ingeniería, Universidad Católica del Maule, Talca, Maule 3480564, Chile. E-mail: mmora@ucm.cl
‡Jair Herazo-Álvarez, Marco Mora and Sara Cuadros-Orellana authors contributed equally to this work.

Abstract

One of the main goals of metagenomic studies is to describe the taxonomic diversity of microbial communities. A crucial step in metagenomic analysis is metagenomic binning, which involves the (supervised) classification or (unsupervised) clustering of metagenomic sequences. Various machine learning models have been applied to address this task. In this review, the contributions of artificial neural networks (ANN) in the context of metagenomic binning are detailed, addressing both supervised, unsupervised, and semi-supervised approaches. 34 ANN-based binning tools are systematically compared, detailing their architectures, input features, datasets, advantages, disadvantages, and other relevant aspects. The findings reveal that deep learning approaches, such as convolutional neural networks and autoencoders, achieve higher accuracy and scalability than traditional methods. Gaps in benchmarking practices are highlighted, and future directions are proposed, including standardized datasets and optimization of architectures, for third-generation sequencing. This review provides support to researchers in identifying trends and selecting suitable tools for the metagenomic binning problem.

Keywords: metagenomics; binning; neural networks; deep learning; classification; clustering

Introduction

Every microorganism possesses its own genome, which is the complete sequence of DNA within its cells [1]. A metagenome consists of genetic material extracted from the assemblage of microorganisms present in a defined environment [2]. Metagenomics is the study of metagenomes of microbial communities, which are largely uncultivable and largely unknown [3]. This has enabled a deeper revelation and understanding of microbial diversity in diverse environments, such as gut microbiomes, soils, water bodies, and extreme habitats [4–8]. These insights have driven significant advances in fields such as medicine, environmental care and remediation, as well as ecology and agriculture [9–12].

Metagenomics research has expanded considerably in the past two decades, due to the advent of high-throughput sequencing methods and the concurrent development of bioinformatic algorithms, which made the fast and reproducible processing of an unprecedented amount of data possible. According to the Integrated Microbial Genomes with Microbiome Samples (IMG/M: https://img.jgi.doe.gov/m/) system, there are currently over 46,000 available metagenomic datasets covering engineered (9%), host-associated (18%) and environmental (73%) samples.

Shotgun metagenomic sequencing is the main strategy used to obtain metagenomic data [13]. The resulting reads must be curated to ensure quality and can be assembled into larger contiguous sequences, called contigs, prior to biological interpretation [2, 14, 15]. Some of the most popular downstream analyses are functional annotation (i.e. assigning functions to genome elements by associating them to biological processes), taxonomic profiling (i.e. reporting of the relative abundances of taxa within a dataset without classifying individual reads), taxonomic binning (i.e. classification of individual sequence reads to reference taxa), and sequence binning (i.e. clustering of DNA sequences, such as reads or contigs, into discrete groups based on their intrinsic characteristics or their similarity to each other). Although the term binning is used more often in the literature to refer to contig clustering, it is not limited to this task. Binning is regarded by several authors, by the Critical Assessment of Metagenome Interpretation (CAMI) [16–18], and also herein, as 'the grouping of data into clusters or classes' (Fig. 1). In fact, early literature used the term 'binning' in metagenomics and 'metagenomic classification' interchangeably [16], establishing the wide scope of the term 'binning', assuming that the assignment of labels that are shared between different sequences goes beyond the identification of individual sequences in a dataset, allowing also their categorization into subsets or 'bins' and the assessment of the number of observations in each bin. As a consequence, the area of metagenomic binning currently covers a wide range of supervised, semi-supervised, and unsupervised methods, as will be described in more detail in this manuscript.

Computational metagenomics [19–21] has enabled a deeper understanding of microbial diversity, by allowing the automatic execution of complex tasks, such as clustering and classification of large amounts of sequencing data [4–8, 22]. Binning, particularly, has played a pivotal role in recovering individual genome

Received: June 17, 2024. **Revised:** January 2, 2025. **Accepted:** March 7, 2025

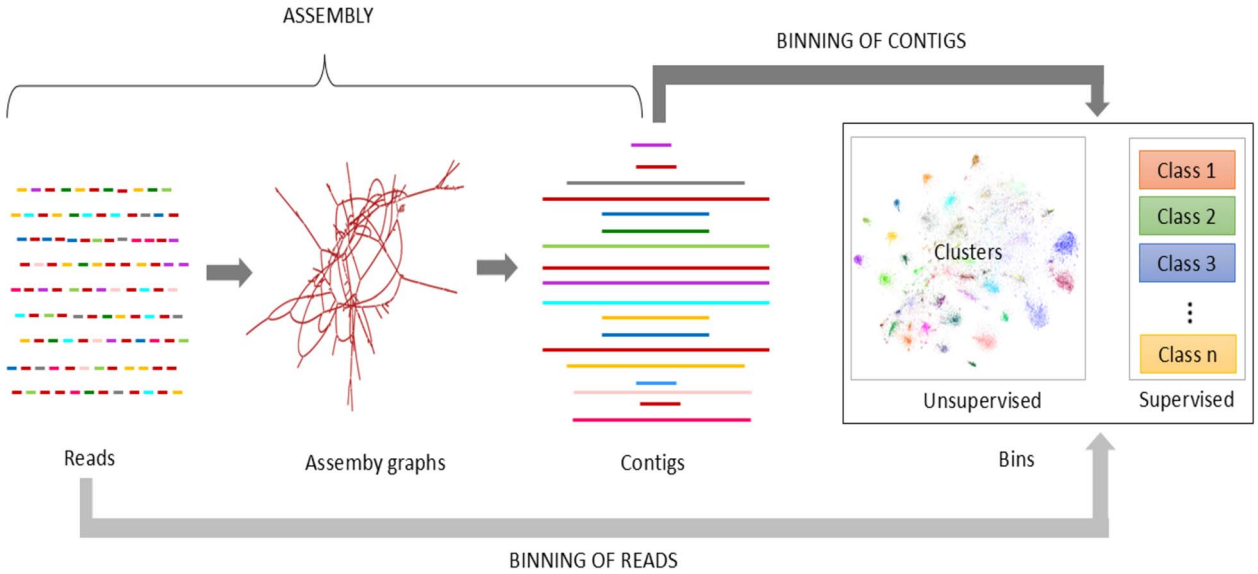


Figure 1. Two approaches for metagenomic binning are illustrated. The first approach directly classifies or clusters reads (bottom arrow). The second approach involves the classification or clustering of contigs (top arrow). In both cases, the resulting bins correspond to clusters (unsupervised methods) or predefined classes (supervised methods).

information from metagenomic data, revealing uncharacterized microbial diversity and complex ecological relations involving uncultured microorganisms [23], which otherwise would have remained unknown [24, 25].

Several binning tools have been proposed and standardized methods for metagenome-assembled genomes (MAGs) generation and description were established [26]. Two crucial aspects in binning are accuracy (ACC), as mis-binning can lead to erroneous biological and ecological inferences [27], and computational efficiency, especially in an era of terabase-scale metagenomics projects [28].

Binning can be performed post-assembly, taking contigs as inputs, or through an assembly-free approach, using reads as inputs [29] (Fig. 1). Each approach can follow different workflows. Taxonomy-dependent (supervised) binning uses information from existing databases or labeled sequences to train algorithms that classify new sequences, often based on alignment methods or features like composition [16]. Taxonomy-independent (unsupervised) binning clusters sequences based on intrinsic information like composition and abundance [17]. Finally, semi-supervised binning combines the advantages of the former methods, leveraging limited labeled data and larger sets of unlabeled data [30].

In the classical literature, the premises, methodologies, advantages, and limitations of binning tools were analyzed, focusing on taxonomy-dependent methods ([16]). State-of-the-art reviews ([17, 31]) include taxonomy-independent binning and read classification and metagenomic assembly strategies. In addition, the use of binning to reconstruct genomes from metagenomes was reviewed ([32]), bringing attention to binners and bin-refinement tools ([33]), including DAS Tool [34] and MetaWRAP [35].

Microbiome research raised challenges in data processing and classification tasks. Artificial neural networks (ANNs) appear as powerful tools for improving ACC in classification tasks [36]. ANNs are computational models inspired by biology and mathematics, used to tackle various tasks in machine learning (ML) [37]. They involve learning methods [38], as shown in Table 1, and downstream and upstream steps (Fig. 2).

The rapid development of binning tools makes it challenging for researchers to keep up with the state-of-the art literature.

Table 1. Key features of learning methods

Methodology	Key Features
Supervised learning	Requires labeled data. High ACC. Regression and classification tasks.
Unsupervised learning	No labeled data needed. Useful for discovering hidden structures in data. Clustering and dimensionality reduction tasks.
Semi-supervised learning	Combines labeled and unlabeled data. Useful when labeling is expensive or time-consuming.

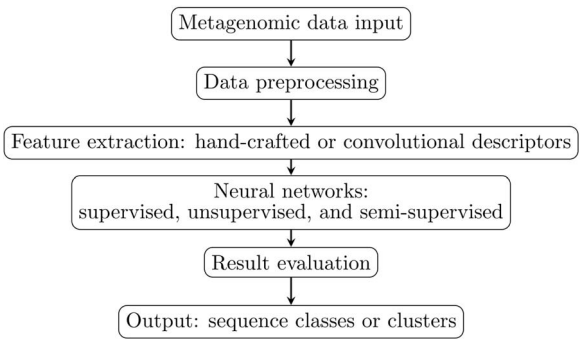


Figure 2. Flowchart of the metagenomic binning process with Neural Networks, from metagenomic data input to the output of sequence.

As the use of ANNs in metagenomic binning has rarely been reviewed, we aimed to assess and analyze the contributions of ANNs in this area. We performed a literature search on Scopus, Web of Science and PubMed covering articles in English published in journals between 2020 and 2023. Our search covered the Scopus, Web of Science, and PubMed databases. Only papers written in English and published in journals from 2020 to 2023 were considered. The key terms used were (binning OR classification OR clustering) AND (metagenom*) AND ('deep learning' OR 'neural networ*'), where * stands for autocompletion and quotation marks for searching the exact phrases used. These terms were only searched in the title, abstract, and

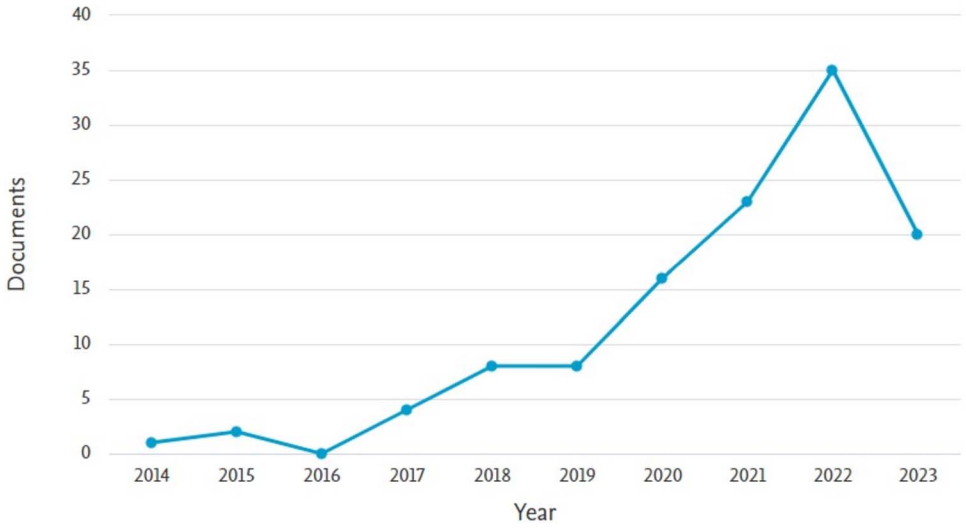


Figure 3. Number of Scopus papers related to ANNs and metagenomic binning in the past decade.

keyword sections of the articles. The number of documents obtained in each database were: 110 in Scopus, 80 in WOS and 69 in PubMEd (total=259). These articles were exported to the RefWorks tool where an information cross-check was made and redundancy across databases was eliminated. As a result, 131 documents were obtained. To further filter these, the following exclusion criteria were used: Metagenomics is not mentioned in the main parts of the paper; The training or text data do not correspond to shotgun data; The paper does not report the binning (either classification or clustering) of metagenomic sequences; The paper does not present binning tools or their variants using neural networks or deep learning. As a result, only 34 unique manuscripts remained, which are described in sections 3, 4, and 5.

Our search revealed a trend of polynomial growth in the number of articles each year in all databases, reaching a peak in 2022 (Fig. 3), suggesting that this field of research is not only prolific but also holds significant potential for future advancements.

Here, we present a systematic analysis of the available resources (tools and databases) that have supported these developments, and discuss their strengths and weaknesses, as pointed out by their own developers. This manuscript is organized as follows. Section 2 describes the basics of metagenomic binning and the types of metagenomic binning. Sections 3, 4, and 5 present the different ANNs used in supervised, unsupervised and semi-supervised binning, respectively. Finally, in Section 6, a discussion is provided.

Fundamentals of metagenomic binning

Supervised binning

Supervised binning is a method used for the detection, characterization, taxonomic classification or abundance profiling of microbial samples [39]. The advantage of this approach is that it can help identify the composition and the relative abundance of organisms in communities [40], detect pathogens [41], and classify viruses [42].

Supervised binning requires the availability known reference sequences and their corresponding taxonomic labels for training. Algorithms based on alignment or on trained models use this information to classify metagenomic sequences, so that diversity profiling can be performed based on the results of the taxonomic assignment [43]. The mapping step is often based on reads or

Table 2. Metrics for evaluation metagenomic binning. Here TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative

Metric	Mathematical definition
ACC	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
F1-score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$

k-mers alignment, using either nucleotide or amino acid database searches [31]. Performance evaluation of classification tools typically uses metrics such as ACC, Precision, Recall, and F1-score (Table 2) and, to a lesser extent, time and memory space used [44].

An advantage of these approaches is that they can be and usually are used for read-level analysis, avoiding assembly-associated biases, allowing the direct exploration and analysis of sequencing data [39]. Also, these tools are well suited to small datasets, as they are generally less time-consuming, and their ACC is usually high [45]. Such strengths also present drawbacks, though. For instance, the ACC of these tools depend on the complexity of the sample, the length of the fragment produced by different sequencing technologies, the methods and algorithms applied, and, in the case of classification, on the amount of data used for training [46]. They are highly dependent on good quality labeled databases and on the number of known species present.

Unsupervised binning

Unsupervised binning can explore the composition of metagenomic data without prior knowledge on the number of organisms or their genetic or taxonomic information [47]. Therefore, this approach is not limited by prior knowledge of species or by the availability of labeled databases. It can capture microbial community information using intrinsic genomic features that serve to find similarities and differences of fragments belonging to different species [48].

In unsupervised binning, a first group of tools use sequence composition patterns, such as GC content and k-mers frequency [49–51] or their combinations, based on the rationale that the frequency of occurrences of such features are maintained locally in the genomes and are different among them. However, composition-based tools can misclassify species with similar

compositional features [52]. A second group of tools incorporates abundance or coverage information [53, 54], and may also misclassify species with similar abundance. Thirdly, hybrid tools combine composition and coverage information [55–58], in order to reduce misclassification. The latter are currently the most frequently used approaches and have shown good performance [18]. Additional features, such as the use of marker genes, codons, assembly and alignment information, can be taken into account to improve binning [32].

Unsupervised binning is used to produce MAGs. A MAG represents a putative microbial genome and is formed by one or more sequences with similar characteristics derived from metagenomic assembly. MAGs allow to study individual genomes in a culture-independent way, and to understand their possible functions in an ecosystem [32]. A workflow for MAG recovery can be found in [59], together with a compilation of tools to tackle each step. The power of these tools is reflected in the increasing numbers of compendiums or catalogs of MAGs [60–63]. Generally, complete workflows and classic tools, such as metagenomic assembly with MEGAHIT [64] and quality control of MAGs with CheckM [65], are used before and after binning, respectively.

Semi-supervised binning

Semi-supervised learning is a ML approach that combines the two previous approaches [66]. It is used because usually at the time of training there is a small amount of labelled data - either because labelling is difficult or costly - and a large amount of unlabelled data [67]. In the field of metagenomics, this can be a very useful approach to understand the 'microbial dark matter' [30, 68]. The proportion of unknown microorganisms compared to known microorganisms is often large, and known information is useful in this type of learning. In fact, as mentioned in [69], the number of reference (labelled) phage genomes is small and the number of new (unlabelled) phage contigs derived from metagenomic studies increases rapidly.

Neural networks for supervised binning

In this section, 25 supervised binning tools are described that employ ANNs. Their initial categorization is based on the types of DNA sequence they are designed to work with at the time of their proposal: reads (Table 3), contigs (Table 4), or general sequences (Table 5). The description of each tool includes the biological context in which it is used, the type of ANN employed, the data sources, the advantages and disadvantages, and other relevant information. In addition, a brief explanation of a workflow used in ANN-based binning tools (Fig. 4(a)) and the the main ANN-models used by these tools is given (Fig. 4(b), (c), and (d)).

Supervised binning on reads

MultiLayer perceptrons (MLPs) are a class of fully connected and feed-forward ANNs [115]. They consist of three types of node layers: an input layer, one or more hidden layers and an output layer (Fig. 4(b)). Application of an MLP for the classification of reads and the estimation of viral taxonomic distribution at the order level is addressed in [70]. *k*-mers distributions, internucleotide distances and DNA sequence motifs are tested as input features, but settled on the combination of the first two for better performance. 2332 viral genome sequences from the National Center for Biotechnology Information (NCBI) database [71] were used as training (70%), validation (15%), and test (15%) sets (available at URL (https://github.com/klausjung-hannover/taxaEstimator)). The algorithm was also tested on real sample of a harbour seal [72]. The ANN achieves the highest ACC rates on the test data, although it depends on the specific classification problem.

Convolutional neural networks (CNNs) are deep learning models [116] that consist of feature extraction and inference zones (Fig. 4(c)). CHEER [73] was designed for taxonomic classification of reads, utilizing hierarchically structured CNNs organized by genus, family, and order. It is capable of tagging new species in viral metagenomic data and is designed to handle the difficult cases of assigning taxonomic labels to species reads that have not been observed before. The features used are the one-hot encoding and the *k*-mers (words) frequency embedding. All viruses used for training were obtained from the NCBI RefSeq virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/∖#/find-data/virus). For testing, a synthetic dataset was generated with the WgSim readout simulation tool [77], and for the real data case, NCBI SRR1930021 and SRR10714026 were used. The results show that CHEER competes favourably with state-of-the-art tools.

A CNN architecture was applied in DL-TODA [78]. This tool was trained on more than 3000 bacterial species. A total of 9,859 bacterial genomes were selected from the Genome Taxonomy Database (GTDB), version 95, and the NCBI RefSeq database. 7405 and 2454 genomes were assigned for training and testing, respectively. The features used from the reads were the 12-mers. DL-TODA achieved an ACC of 97% at the species level, higher than 0.93 for Kraken2 and 0.85 for Centrifuge on the same test set. Compared to other tools, DL-TODA predicted different relative abundance rankings and is less biased towards a single taxon.

SquiggleNet [82] was introduced for real-time classification of reads generated with Nanopore technology. The input to SquiggleNet is the electrical signal generated by a nucleic acid as it transits through a nanopore in Oxford Nanopore's MinION sequencer. These electrical signals are recorded as 'squiggles', which are a one-dimensional time series of electrical current values. The architecture consists of a 1D-CNN based on ResNet. SquiggleNet uses 1D-ResNet style bottleneck blocks with increasing numbers of filters and pooling averaging, followed by a fully connected layer with softmax activation to classify the electrical signals of the molecules. The SquiggleNet tool was evaluated on a human respiratory metagenome dataset published in [117]. The tool achieved an overall ACC of over 90%.

The long short-term memory (LSTM) are a type of recurrent neural network that contain a memory cell that can maintain information over long periods.[118]. Each LSTM unit contains input, forget, and output gates that regulate the flow of information (Fig. 4(d)). DeePaC [85] was presented as a tool that used LSTMs and CNNs for pathogenic classification. It includes a flexible framework that allows easy evaluation of neural architectures with inverse complement parameter exchange. CNNs and LSTM are used separately and combined in different runs, demonstrating good performance against alignment approaches. The sequences in this case are one-hot encoding form. CNNs, LSTMs and the combination of both outperforms the state-of-the-art methods in simulated and real data, accurately predicting potential pathogens from isolated NGS reads.

DeepMicrobes [92] was presented as a deep learning-based computational framework for the taxonomic classification. This tool was trained on genomes reconstructed from gut microbiomes and facilitates efficient investigation of uncharacterised functions of metagenomic species. To train the CNN and LSTM networks used, two types of features are used, namely *k*-mers embedding and one-hot encoding, were used. Better performance is obtained with the former. 2505 representative genomes of human

Table 3. Summary of tools that use neural networks in the supervised binning of reads, and their main features

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
TaxaEstimator [70] https://github.com/klausjung-hannover/taxaEstimator	MLP / k-mers and Inter-nucleotide distances.	On test data: TPR (56.6%), TNR (96.6%), Positive Predictive Value (PPV) (53.6%), Negative Predictive Value (NPV)(94.5%), and ACC (41.0%). / LDA and SVM.	Viral (order level) / Training, validation, and test data: reads (150 bp) of 2332 viral genome sequences from NCBI database [71]. Validation on real case: Harbour seal [72].	Outperformed a mapping approach. Introduced a novel correction approach to improve the taxonomic frequency distribution estimation. / A large number of misclassifications. The large intra-order and small inter-order heterogeneity of viral sequences makes classification difficult.
CHEER [73] https://github.com/KennthShang/CHEER	CNN / k-mers and one-hot.	ACC of 96% at order level. AUROC of 99.76% in the rejection layer for Ebolavirus. / Naïve Bayes classifier (NBC) model [74], CNN-based classifier called metagenomicDC [75], and alignment-based method VirusSeeker [76].	Viral (from order to genus) / Training data: reads (250 bp) of virus genomes from the NCBI RefSeq virus database(https://www.ncbi.nlm.nih.gov/labs/virus/vssi/∖#/find-data/virus). Test data: a synthetic dataset generated with WgSim [77]. For the real data case: NCBI SRR1930021 and SRR10714026.	Robust to sequencing errors and fast, when compared to VirusSeeker. Good for labeling new species in viral metagenomic data. / Classification ACC decreases throughout the taxonomic ranks. Model training requires heavy computing resources.
DL-TODA [78] https://github.com/zhanglab/dl-toda	CNN / k-mers	Micro and macro averages for precision (98% and 91%, respectively), recall (97% and 76%, respectively), and F1-score (98% and 80%, respectively), were observed in the testing set. ACC was 97% or greater at all levels. / Alignment-based method, Kraken2 [79], and pseudo alignment-based method, Centrifuge [80], in test and real data.	Bacterial (from species to phylum) / Training and validation used complete genomes from the Genome Taxonomy Database (GTDB) and NCBI RefSeq database, from which paired-end reads (PE250) were generated. Training used 7405 genomes from 3313 species, and testing used 2454 genomes from 639 species. Real data from the human oral cavity [81] and cropland soil (NCBI ERR5004682, ERR5003895, ERR5003204, ERR5001925, ERR4995171) were used.	Strong performance when applied to metagenomes from diverse environments, such as the human oral cavity and cropland soil. High ACC at all taxonomic ranks tested, outperforming alignment-based methods. / Certain species yielded low precision, and the application of a decision threshold can reduce the number of classified reads. Requires extensive computational power, e.g. multiple GPUs and significant memory resources.
SquiggleNet [82] https://github.com/welch-lab/SquiggleNet	CNN / Electrical signals	ACC of 90%, TPR and TNR around 90%, and precision of approximately 90% (for HeLa&Zymo dataset). The AUROC was also around 90%. / Guppy+Minimap2 [83], an alignment-based method and UNCALLED [84], a mapping tool.	Bacterial / The training and test data: HeLa&Zymo dataset, which contains millions of reads of human and bacterial DNA (NCBI SRA SRX9818341). Real data: a human respiratory metagenome sample (NCBI SRA SRX9818342)	Requires little computational resources (304 KB of RAM), has minimal processing delays compared to alignment-based methods. Strong performance in classifying species from Nanopore signals, / Weaker precision in highly imbalanced datasets, such as the Human&Zymo_b56. Limited ACC when distinguishing certain bacterial species, especially Gram-positive ones. Dependency on preprocessing, that is, effectiveness relies on proper signal normalization and preprocessing, such as adaptors and barcodes removal.

(*Continued*)

Table 3. Continued

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
DeePaC [85] https://gitlab.com/rki_bioinformatics/DeePaC	CNN + LSTM / One-hot	ACC of 99.2% (temporal paired-end read data); PPV of 100% (real and temporal paired-end read data); TPR of 0.984 (real data); TNR of 0.93 (BacPaCS dataset); Balanced ACC of 84% (BacPaCS dataset); AUROC of 0.948 and AUPRC of 0.989 (paired-end read data). / BLAST [86], an alignment-based method. Model-based methods PaPrBaG [87], BacPaCS [88], and [89].	Pathogens / Training data: reads both the single-end and the paired-end from IMG/M database, at read-level and species-level until April 2018, Refer to source [90]; Temporal hold-out data, re-accessing the IMG/M database in January 2019; Real data case study, extracted from [91]; and, BacPaCS dataset, the data used in the source [88].	Reverse-complement networks allows to consider both strands of DNA simultaneously. Can be used for different types of DNA sequences, including novel, unrecognized, and synthetic sequences. / Shows an ACC gap between validation and test data. Limited ability to determine pathogenicity in complex multi-factorial scenarios.
DeepMicrobes [92] https://github.com/MicrobeLab/DeepMicrobes	CNN + LSTM / *k*-mers	Precision higher than 95% and recall of 42.8% (Read-level on the gut-derived MAGs dataset). Classification rate average of 89.40% (at genus-level on mock datasets). The L2 distance for abundance estimation performed better on genus quantification. / All are alignment-based or pseudo-alignment methods: Kraken [93], Centrifuge [80], CLARK [94], CLARK-S [95], Kaiju [96], DIAMOND-MEGAN and BLAST-MEGAN [97].	Taxonomic classification (species and genus) / Trained on 2505 representative genomes of human gut microbiome [98]. Benchmark data were obtained from the European Nucleotide Archive (ENA ERP108418, ERP105624, ERP012217), and NCBI (PRJNA348753).	Superior genus and species classification in complex microbial communities. Good for exploring novel MAGs and reliable for estimating the relative abundances of microbial taxa. / Memory and computational resource demanding, it relies on GPUs for training and testing, and adding new species requires retraining. The performance, particularly the recall, is affected by the similarity between the tested genomes and the training set.
DETIRE [99] https://github.com/crazyinter/DETIRE	GCN + CNN + BiLSTM / *k*-mers	ACC of 90.30%, recall of 90.40%, precision of 89.57%, and F1-score of 89.65% (the CAMI dataset). Good time consumption for all datasets. / All tools are model-based methods: DeepVirFinder [100], PPR-Meta [101], and CHEER [73].	Viral / Training and test data: millions of reads (500 bp) of NCBI Virus RefSeq genomes (NCBI accession numbers can be found on GitHub of DETIRE) and prokaryotic host genomes, as used in VirFinder [102]. Real test data: 2nd CAMI Challenge Marine Dataset (in https://data.cami-challenge.org/participate) and Human Gut Metagenome form NCBI SRA (Accession: SRA052203)	Good in recognizing viral sequences shorter than 1000 bp and robustness in identifying viral sequences from different environments. Efficient processing for large-scale metagenomic studies. / Performance decreases as the length of the input sequences increases. Limited phage identification, it shows slightly lower ACC in phage identification.
RdRpBin [103] https://github.com/HubertTang/RdRpBin	CNN + GCN / One-hot	Precision around 93.1% (real dataset). Recall around 84.9% and F-score around 89.1% (simulated dataset). / Alignment-based methods: DIAMOND [104], MMseqs2 [105], and Kraken2 [79]. Pseudo-alignment-based method: Kaiju [96].	Viral / A reference database called NCBI-RdRp was built for training. Sequence simulation for testing was based on Pfam [106] and all NCBI-RdRps that could be aligned with profile hidden Markov models. It was tested on a simulated viral marine metagenomic dataset extracted from [107] and NCBI SRA ERR2185279. A real metagenomic data from NCBI (SRA SRR9216246) was used.	Superior balance between recall and precision in scenarios where the query RNA viruses have low similarity with known viruses. Useful for classifying new and divergent RNA viruses, combining alignment-based strategies and graph-based learning models. / Limitations on imbalanced data, where reads from rare orders may be misclassified. Affected by limited training samples, particularly at lower taxonomic ranks.

(*Continued*)

Table 3. Continued

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
MetaTransformer [108] https://zenodo.org/badge/latestdoi/584424427	Transformer / *k*-mers	Precision of 98.2% (Benchmark-Mock dataset). Recall of 90.4% (Benchmark-Gut dataset). Training time the within 9-16h. / Alignment-based methods: Kraken2 [79] and CLARK [94]. Pseudo alignment-based method: Centrifuge [80]. Model-based method: DeepMicrobes [92].	Taxonomic classification (species and genus) / Training and validation data: PE150 reads from 1952 unclassified metagenomic species and 553 species from the human gut reference http://ftp.ebi.ac.uk/pub/databases/metagenomics/umgs_analyses/. Evaluation data: from The European Nucleotide Archive (ENA), Human gut benchmark data (ERP108418), Mock community data (ERP105624 and ERP01221), and Species absent from reference data (RJNA348753). Real-world data from ENA (PRJNA398089).	It outpaces DeepMicrobes 2 Lower precision and reduced performance for absent species detection compared to traditional alignment-based methods. Larger *k*-mer sizes generally yield better results, but increase memory consumption.
ViBE [109] https://github.com/DMnBI/ViBE	BERT / *k*-mers	MCC of 0.88 for (read-level validation). F1-score of 0.962 (domain-level classification). Precision of 0.965. Recall (0.9599). AUROC (0.98). AUPRC (0.99). / Alignment-based methods: VirSorter [110] MetaPhlAn3 [111], and Kraken2 [79]. Model-based methods: CHEER [73], DeepVirfinder [100].	Viral / Training and validation data (151 bp and 251 bp paired-end reads) were simulated from 2293 and 2733 genomes for eukaryotic DNA and RNA viruses, respectively. Experimental datasets from NCBI SRA (SRR7969779, SRR7969850, SRR14403295, SRR10419849, and ERR2681947). Real data: six vaginal virome samples from NCBI SRA (SRR12213162, SRR12213197, SRR12213424, SRR12213928, SRR12214082, and SRR12214159).	Adaptability to novel virus subtypes, even when trained without reference genomes. It outperformed state-of-the-art alignment-free methods for all simulated test cases. When compared to alignment-based methods, it showed higher recall rates. / A notable percentage of reads from small eukaryotes like *Candida albicans* and *Saccharomyces cerevisiae* were misclassified as viruses. Performance is affected by the unbalanced availability of reference genomes.
TaxonomicClassification-NGS-NN [112] https://github.com/CBMI-HTW/TaxonomicClassification-NGS-NN	BERT / Proteins encoded by the reads	The ROC curve and error matrix revealed that the frame classification model performed well with an ACC of 98% and few misclassifications. Precision of 0.91 (taxonomic classification model). / Model-based methods: CNN-MGP [113] and FragGeneScan [114].	Pathogens (domain level) / Test, validation, and training data: For taxonomic classification, UniProtKB/Swiss-Prot database, manually curated for bacterial, viral, and human (mammalian) sequences (Available at https://zenodo.org/record/4306240). For frame classification, randomly selected viral, bacterial, and human genome sequences from RefSeq (available at https://zenodo.org/record/4306248). Real data: human skin metagenomic study (SRA ID: SRR7188139) and Swine feces metagenome (SRA ID: ERR3013343).	Good for prediction the taxonomic classification of sequences that can not be classified by comparison to a reference database. It achieved an ACC of 98% in frame classification, demonstrating robustness in determining the correct reading frame of a sequence. / Only works with reads that are within coding sequences, which could restrict its applicability to a subset of the data. Difficulty to distinguish between viral and human sequences, particularly due to the presence of retroviral sequences in the human genome.

Table 4. Summary of tools that use neural networks in the supervised binning of contigs, and their main features

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
PredicTF [122] https://github.com/mdsufz/PredicTF	MLP/ Sequence similarity to known TFs	The number of TFs: 792 TFs (Anaerobic ammonium oxidation microbial community). ACC was 98.69%, reached in *Pseudomonas fluorescens*./ An alignment-based method, Prokka [123].	Protein/ Training data: BacTFDB, 11 961 TFs distributed in 99 types of TF families, available at https://github.com/mdsufz/PredicTF. Testing data: an anaerobic ammonium-oxidising microbiome (NCBI PRJNA511011).	It can be used with no previous knowledge regarding TFs. It is fast, and it requires low memory when compared to Blast-based annotation. As a user-friendly tool with high precision, could be considered over traditional methods./ While the BacTFDB database is regularly updated, it may still face limitations in terms of diversity and sequence inclusion. It was not able to predict new TFs for all organisms.
Tiara [124] https://github.com/ibe-uw/tiara	CNN/ *k*-mers	ACC of 98.93% for prokaryotic genomes and 98.83% for nuclear eukaryotic genomes (test dataset). ACC was 99.60% for plastidial and 98.86% for mitochondrial genomes./ EukRep [125], a model-based tool.	Classification of eukaryotic nuclear/ Training dataset consists of 8,220 genomic sequences: Eukaryotic genomes (4,381 sequences), Bacterial genomes (1860 sequences), and Archaeal genomes (1979 sequences). Test dataset consists of 550 genomes from the three domains of life (165 Eukarya, 306 Bacteria and 79 Archaea), which were not present in the training dataset. Real test data: metagenome of *Pseudoblepharisma tenue* [126] and Tara Oceans Dataset (NCBI SRA: ERR1726574, ERR1726673, ERR868402).	When using parallel execution, making it highly efficient for large-scale datasets. The only available tool that correctly classifies organellar sequences (mitochondrial and plastidial genomes) in metagenomic datasets./ The classification ACC decreases with shorter sequence lengths. It may misclassify NUMTs (nuclear mitochondrial DNA fragments) and NUPTs (nuclear plastid DNA fragments), assigning them as organellar sequences.
DeepVirFinder [100] https://github.com/jessieren/DeepVirFinder	CNN/ One hot	AUROC values were observed for 3000-bp sequences, with a mean value of 0.98. AUPRC, for contigs of length greater than 500 bp, being of 0.9296./ VirFinder [127], which is a model-based tool.	Viral/ A total of 2314 reference genomes of prokaryotic viruses were obtained from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/genome/browse) for use in the training, validation, and test phases. The genomes were selected based on the date of discovery and were split into separate sets to avoid overlap. Real data: human gut metagenomic samples from patients with colorectal carcinoma (ENA accession number ERP005534).	It shows robustness and ability to handle short viral contigs (such as 300 and 500 bp) effectively. Ability to identify underrepresented viral groups, improved the prediction ACC for viral groups that are underrepresented./ Potential for misidentification with metagenomic samples contain eukaryotic contamination. Sensitivity to viral groups with high representation, that is, when the model was trained using the enlarged dataset, the AUROCs for the two most abundant viral types were decreased.

(*Continued*)

Table 4. Continued

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
EdeepVPP and EdeepVPP-hybrid [128] *NO-URL*	CNN + LSTM/ One-hot	AUCROC achieved was 0.9976 (the 2014_G1 dataset), and AUCPR was 0.9971 (the 2014_G1 dataset)./ Model-based method: ViraMiner [129]. Furthermore, the results were contrasted with those (theorical) obtained from five additional tools.	Viral/ Training, validation, and test data: consists of human metagenomic contig experiments derived from human samples. 19 different metagenomic contig experiments collected from NGS studies are included. They contain 300-bp contigs. The details of these datasets are described in ViraMiner [129].	The interpretable capability of the model allows it to extract underlying patterns that are biologically relevant, which enhances the explainability of the results. Ability to generalize well across different datasets and to recommend viral sequences that were not included in the training set./ The computational cost of training deep learning models is significantly high. The requirement of a large number of labeled sequences for training limits the tool's applicability, making it less effective in scenarios with insufficient training data.
MLR-OOD [130] https://github.com/xinbaiusc/MLR-OOD	LSTM/ One-hot	AUROC and AUPRC were consistently above 0.9 for contigs of at least 1000 bp across all datasets./ Model-based methods: Likelihood Ratio method [131] and max-LL (proposed in this paper).	Taxonomic classification/ Training and test data: contigs from two subsets consisted of bacterial sequences: Test2016, which was constructed in [131], and Test2018, which was generated using the same procedure with which the former was constructed. The third subset consisted of 1295 prokaryotic viral genomes and 818 genomes of plasmids. All the data used are available at https://drive.google.com/drive/folders/1Kz0kQ_D1VWYqA-GDld783O7H8AzNuHkC	It eliminates the need for tuning model parameters and achieves remarkably higher prediction ACC across all tested datasets. Highly robust to GC content./ High computational resource. The improvement in prediction ACC on viral and plasmid datasets is limited.
PhaMer [132] https://github.com/KennthShang/PhaMer	Transformer/ Protein composition + the proteins' positions	On the short contig test set, the precision is around 0.8. On the IMG/VR v3 dataset, the recall is around 0.85. On the mock metagenomic dataset, PhaMer achieved an F1-score that is around 0.6./ Alignment-based method: VirSorter [110] and four learning-based methods: Seeker [133], DeepVirfinder [100], VirFinder [127] and PPR-meta [101].	Viral/ 2126 bacteriophage genomes from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/) released before December 2018. Test data: after December 2018 were used as the test set, including 2284 bacteriophage genomes. Real data: includes mock community sequencing data from the ENA (BioProject PRJEB19901), as well as viral contigs extracted from IMG/VR v3 [134].	More robust performance than others on short contigs. Better generalization to real metagenomic data, addressing the challenge of local similarities between phages and their hosts,/ High computational resource demands. Trade-off between recall and precision, where the E-value cutoff in protein matching affects performance.

Table 5. Summary of neural networks for supervised binning of sequences, and their main features

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
VIBRANT [144] https://github.com/AnantharamanLab/VIBRANT	MLP/ Protein signatures + v-score metric	Recall was 98.43% (viral fragment dataset from NCBI RefSeq and GenBank databases). Precision was 99.87% (the viral dataset). ACC of 99.15% across the datasets. Specificity reached 99.90% for bacterial/archaeal fragments and 98.90% for plasmid fragments. F1-score was 0.991 on viral fragment data. MCC of 0.983, in distinguishing viral from non-viral sequences./ Model-based methods: VirSorter [110], VirFinder [127], and MARVEL [102].	Viral/ Training and test data: includes sequences from bacteria/archaea (181), plasmids (1452), and viruses (15 238), downloaded from RefSeq and GenBank databases (NCBI), accessed in July 2019. All sequences were split into non-overlapping, non-redundant fragments between 3 and 15 kb. Real data: metagenomic reads of Crohn's disease metagenomes from [145].	Not limited by sequence motifs or databases, meaning it can recover both free viruses and proviruses from metagenomic data. Superior performance in specificity and recall, making fewer false identifications compared to other tools./ Designed to exclude short scaffolds. This restriction could limit its ability to identify viruses in assemblies where short fragments are common. Focus on bacterial and archaeal viruses, it may not be as effective in recovering viral genomes from other domains.
NLP-MeTaxa [146] https://github.com/padriba/NLP_MeTaxa/	MLP/ Word embeddings	Precision: 83.59% (CAMI-Medium), recall: 62.37% (CAMI-High), and F1-score: 65.89% (CAMI-High)./ Alignment-based method: Kraken [79]. Model-based methods: PhyloPythiaS+ [147] and Taxator-tk [148].	Taxonomic classification/ Training data: trained using over 14 000 microbial genomes from the NCBI RefSeq database. Test data included simulated datasets from the First CAMI challenge [149], with three sub datasets: low, medium, and high complexity. Real data consists of metagenome samples from two obese human twins' guts, used in previous microbiome studies [150].	It outperforms other tools, especially in classifying low-ranking taxonomic classes such as species and genus. It offers a reduced processing time per nucleotide, achieving an average of 2.8 ms/nucl, which is faster than traditional methods like Kraken./ The ACC is potentially limited by the availability and comprehensiveness of reference genomes in public databases. Model complexity in training and require significant computational resources. The complexity of the model being trained can necessitate the utilisation of substantial computational resources.
Predict MICs [151] *No-URL reported*	MLP + LSTM/ *k*-mers + Single-nucleotide polymorphism	ACC of 91.9% using k-mers, Coefficient of determination, was 0.836 using k-mers, and Root Mean Square Error of 1.955 using k-mers./ XGBoost, a model-based method presented here.	Antimicrobial/ - Training data: metagenomic sequences of Klebsiella pneumoniae genomes. The study includes 110 genome samples. The sequence data can be accessed via the NCBI BioProject access numbers PRJNA376414, PRJNA386693, and PRJNA396774 (NCBI SRA access number available in supplementary materials). Test and training sets were split with an 80:20 ratio. HS112861 was selected to be reference genome for SNP calling (NCBI SRA access number available in supplementary materials from [152]).	The use of a top 40 feature-based model allowed to focus on the most important genomic characteristics, improving the prediction model's efficiency while avoiding overfitting. The DNN model consistently achieved higher ACC, particularly in predicting MIC values./ Small sample size limits the model's performance, this could restrict the generalizability of the results to larger datasets. DNN model required significant computational resources for training compared to XGBoost.

(*Continued*)

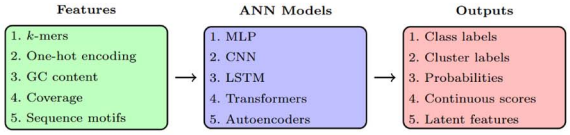
Table 5. Continued

Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
CNN-RAI [153] https://github.com/mrymaltin/CNN-RAI	CNN/ *k*-mers	ACC was around 98% (for sequences with 10 000 bp read length). Training time and test time were measured for each Model-based methods: RAIphy [154] and CNN-COk [75].	Taxonomic classification/ Trained and tested with 10-fold cross validation. Three different datasets representing three different scenarios. Whole Genome Sequencing, simulated using Illumina technology with different read lengths from 757 microbial species and 61 genera; 16S rRNA, simulated metagenomic fragments gathered from the RDP database, containing 100 genera, each represented by 10 species. Long-Read Metagenomic Sequencing, real data obtained from the Oxford Nanopore MinION sequencing technology with six genera (NCBI SRA: ERR1898312, ERR1474981, SRR5117441, SRR5277601, SRR5344355)	Robust learning ability. It was particularly effective in handling noisy data, as observed with Oxford Nanopore MinION reads. The model benefits from GPU acceleration, making it suitable for real-time pathogen detection using real-time sequencing technologies./ Overfitting on short reads, such as the 250-bp reads in the sr-16S dataset. The model demands higher memory resources during training, especially for larger *k*-mer sizes.
VirionFinder [155] https://github.com/zhenchengfang/VirionFinder	CNN/ One-hot	Sensitivity of 94.50% was recorded in the full-length protein test dataset, specificity was 90.07% in the same dataset./ Model-based methods: iVIREONS [156], PVPred [157], PVP-SVM [158], PVPred-SCM [159], and Meta-iPVP [160].	Viral/ The training and test data: were constructed using complete prokaryotic virus genomes obtained from the NCBI RefSeq viral database (in November 28, 2019, from ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/). Genomes released before 2018 were used for training. Test data, containing genomes released after 2018. 10-fold cross validation was performed. Real virome data: metagenomic data from lung virome samples (NCBI SRA: SRR5224158.1) and 22 virome samples of healthy human gut from [161].	Better performance than the existing tools in identifying PVVPs. Robust performance on real data and ability to handle incomplete genes./ Potential false positives, training set did not include bacterial proteins, the tool may struggle to distinguish between PVVPs and non-PVVPs when host contamination are present. Limited to prokaryotic viruses, its current training set excludes eukaryotic viruses.
Seeker [133] https://github.com/gussow/seeker	LSTM/ One-hot + Integers	TPR was 0.90 (on environmental sequences from phage families). Balanced ACC of 84% (on short sequences mimicking metagenomic data). Precision of around 0.85 (on highly divergent phage sequences)./ Model-based methods: VirSorter [110], VirFinder [127], DeepVirFinder [100], PPR-Meta [101], and VIBRANT [144].	Viral/ Training and test data: thousand of phage and bacterial genomes from NCBI https://www.ncbi.nlm.nih.gov/ and https://www.ebi.ac.uk/genomes/phage.html were used (with the accessions available in Supplementary Table S1). datasets from the IMG/VR https://img.jgi.doe.gov/cgi-bin/vr/main.cgi, and also downloaded short phage and bacterial sequences (1K–5K in length) from NCBI. Real data: including the human gut and sheep rumen microbiomes (NCBI projects: PRJEB22623, PRJEB25190, PRJNA504765, and PRJNA577476).	The number of parameters used is relatively small, precluding memorization and overfitting and it does not require substantive computational resources, compared to other approaches. The model is trained on segments and thus performs better on shorter sequences, which is critical for application to metagenomic data./ It does not perfectly distinguish phages from bacterial genomes, and some bacteria and phages are misclassified. Seeker was not trained to identify prophages within bacterial genomes and its ability to do so has not been tested.

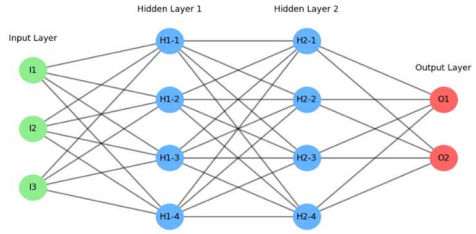
(*Continued*)

Table 5. Continued

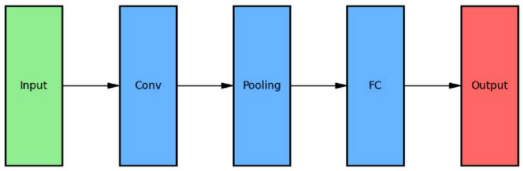
Tool name/URL	ANN model/Feature	Metric (%)/Compared to	Context/Datasets	Advantages/Disadvantages
Virtifier [162] https://github.com/crazyinter/Seq2Vec	LSTM/ One-hot	AUROC was 0.9354 (sequences of 500 bp in the test dataset). AUPRC was 0.6286 (the highly imbalanced CAMI Challenge Dataset 3 CAMI_high dataset). Virtifier outperformed other tools in terms of precision, recall, and F1-score, but exact values for these metrics appear in supplementary tables./ Model-based methods: VirFinder [127], PPR-Meta [101], and DeepVirFinder [100].	Viral/ Training and test data:The Viral and host RefSeq genomes used are available at https://dx.doi.org/10.1186/s40168-017-0283-5, it was divided into non-overlapping sequences of various lengths (300, 500, 1000, 3000, 5000, 10 000 bp). The CAMI Challenge Dataset 3 CAMI_high dataset is available at https://data.cami-challenge.org/participate. Real data: human gut metagenomes are available at https://dx.doi.org/10.1101/gr.142315.112.	It excels in identifying short viral sequences (<500 bp) from metagenomes. This is a significant issue, as a considerable number of existing tools are unable to effectively process short sequences. Robustness to sequencing errors (e.g. base substitutions and insertions/deletions)./ Limited memory when dealing with longer sequences (>500 bp), as the long input chain of the LSTM network can lead to memory impairment. Low identification rate for long viral sequences compared to other tools like PPR-Meta.
BERTax [163] https://github.com/f-kretschmer/bertax	BERT/ Tokens	ACC was 98.62% for superkingdom prediction and ACC of 95.10% for phylum classification./ Alignment-based methods: Kraken2 [79], sourmash [164], MMseqs2 [105], and minimap2 [165]. Model-based method: DeepMicrobes [92].	Taxonomic classification (levels of superkingdom, phylum, and genus)/ 1 million fragments of length 1500 nt of archaean and eukaryotic genomes from NCBI us ing ncbi-genome-download (https://github.com/kblin/ncbi-genome-download/, version 0.2.12); viral and bacterial genomes from NCBI manually. The list of genomes, is provided in supplementary information, Dataset S1.	Versatile classifier that can handle any genome region across all superkingdoms, without being restricted to coding regions or requiring similar sequences in a database. Combination with database methods, like MMseqs2, BERTax increases recall while preserving precision./ ACC decreases at the genus level compared to higher taxonomic ranks. Its performance is slightly affected by shared homology between sequences.



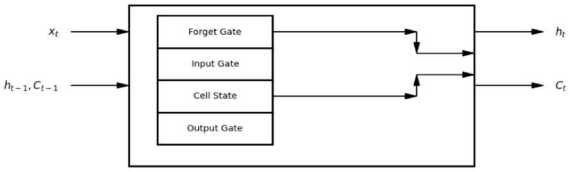
(a) Workflow in ANN-based binning tools



(b) MLP Architecture



(c) CNN Architecture



(d) LSTM Architecture

Figure 4. The figure shows the main ANN architectures used by the supervised binning tools analyzed in this review. (a) The diagram illustrates the three main components in a workflow used in ANN-based binning tools. Features represent the input data for training the ANN models (i.e. k-mers and GC content). ANN Models process these features to extract patterns. Finally, the Outputs represent the results produced by the models, which are used to bin (i.e. class labels and cluster labels). (b) MLP: In the input layer, each neuron (I-i) represents an input feature. In the hidden layers, each neuron (H1-j) and (H2-k) performs a linear combination of the inputs received from the previous layer, followed by the application of a non-linear activation function. In the output layer, the neurons (O-n) produce the final predictions. (c) CNN: The Conv layer applies convolutional filters to the input, extracting local features. The pooling layer reduces the dimensionality through operations such as max-pooling. The fully connected layer takes the extracted features and combines them into a final representation for classification or regression. (d) x_t represents the current input. LSTM: h_{t-1}, C_{t-1} are the previous states. The Forget Gate decides how much information to forget from the previous state C_{t-1}, the Input Gate regulates how much new information to store, the Cell State is the updated long-term memory, and the *Output Gate* controls how much of the updated cell state, C_t, is used as output, h_t, to the next cell.

gut microbiome found in [98] were used for training. Benchmark data were obtained from the European Nucleotide Archive (ENA) ERP108418, ERP105624, and ERP012217, as well as from NCBI PRJNA348753. The DeepMicrobes algorithm and all the data are available at URL (https://github.com/MicrobeLab/DeepMicrobes-data). The results show that DeepMicrobes outperforms state-of-the-art taxonomic classification tools in species and genus identification.

For the detection of viral fragments in metagenomes, DETIRE [99] was presented. This tool uses three types of ANNs. The first, a GCN is used for embedding of k-mers frequency ($k = 3$). Then, the output of the former network is fed into a CNN and a Bidirectional LSTM (BiLSTM) to learn their spatial and sequential features, respectively, for classification. The training, validation, and test datasets were obtained from NCBI Virus RefSeq genome Virus, downloaded on 11 October 2022. The results show that DETIRE outperforms other methods in identifying short sequences.

RdRpBin [103] was introduced as a viral read classification tool based on marker genes, specifically RNA-dependent RNA polymerase (RdRp) genes. For classification, RdRpBin can take either DNA or protein sequences, as it combines an alignment-based strategy with ML models to fully exploit the sequence properties of RdRp. RdRpBin uses a CNN to process the sequences (one-hot encoding) and generate the embedding vector for each node of the graph and a convolutional graph network (CGN) for the final classification. The tool can maintain a higher F1-score than the other tools, in particular, when query RNA viruses share low sequence similarity with known viruses.

MetaTransformer [108] is presented for the classification of reads at the genus or species taxonomic level. Aiming to improve tasks with regard to classification speed and prediction ACC, they employ a Transformer network with a multi-head self-attenuation module. The six datasets used for training and testing are described, and their sources are shown, in Table 1 of the paper. The used features of the reads are embedding or k-mers tokens. Compared to DeepMicrobes [92], MetaTransformer is similar in genus-level prediction and outperforms it in species-level prediction.

ViBE [109] was proposed a Transformer-based deep learning model for identifying and classifying viruses using metagenomic sequencing data. Taking as a reference other tools proposed that use the BERT model as a basis, ViBE also uses a hierarchical BERT model at the domain and order level. It divides the input sequences into a list of tokens, which are (k - 1)-overlapping k-mers with k=4 and adds the classification token and the separation token. Embedding is done for the tokens representing each of them and which are the input to the network. Input reads are first classified at the domain level, and those classified as eukaryotic DNA viruses or RNA viruses are classified at the order level. 10 119 viral genomes from the NCBI's RefSeq database were used for training and testing at genomic and reads level. Evaluation on real data was performed on two sets: the first one composed of virus and bacterial sequences downloaded from NCBI SRA SRR7969779, SRR7969850, SRR14403295, SRR10419849, and ERR2681947; and the second one composed of six virome sequences SRR12213162, SRR12213197, SRR12213424, SRR12213928, SRR12214082, and SRR12214159. Validation with genomic and metagenomic data showed better ability to find new viruses than existing methods.

In [112], a pipeline is proposed for detecting new pathogens in metagenomic datasets by classifying (segments of) proteins encoded by sequences within these datasets. It takes reads as inputs, translates them with the Biopython tool [119], and ends in the classification of frames and taxonomic classification of each sequence into virus, bacterial, and mammalian taxonomic groups. A pre-trained language model called ProtBert, based in

BERT, is implemented. They used UniProt amino acid sequences [120] for the taxonomic classification model and RefSeq reference sequences [121] for the frame classification model. The pipeline was tested on reads from two metagenomic sequencing projects available at NCBI SRA SRR7188139 and ERR3013343, and it achieved a taxonomic assignment ACC of 91%.

Most of the 11 tools described in this subsection (e.g. CNN, LSTM and BERT) use deep learning and process reads based on *k*-mers frequency through one-hot encoding (Table 5). Most were validated using viral sequences, and all of them show good performance compared to state-of-the-art tools regarding alignment (e.g. CHEER and ViBE). However, their performance is affected by read length (e.g. DETIRE and DeepMicrobes) and by the existence of low abundant taxa (e.g. DL-TODA and RdRpBin).

Supervised binning on contigs

PredicTF [122] was presented as a platform supporting the prediction and classification of novel bacterial transcription factors (TFs) in individual species and complex microbial communities. PredicTF is based on a deep learning algorithm found in [135], which is an MLP consisting of four dense hidden layers of 2000, 1000, 500, and 100 neurons that propagate the bit score distribution to dense and abstract features. To train PredicTF, the BacTFDB database, which has a total of 11 961 TFs distributed in 99 types of TF families, was used. This database is the result of joining and filtering items from the CollecTF [136] and UniProtKB [137] databases. To test PredicTF in a complex microbial community, an anaerobic ammonium-oxidising microbiome (NCBI PRJNA511011) was used. Translated contigs were used as the input. With these metagenomic data, PredicTF was able to predict 792 TFs from 27 different families.

Tiara [124] is a deep learning-based approach for the identification of eukaryotic sequences in the metagenomic datasets. The initial step involves classification into six different categories: three for prokaryotes, two for eukaryotes (dividing organellar categories into plastidial and mitochondrial classes), and one for unknown. This workflow uses an MLP with two hidden layers, and it takes contig k-mer frequency as input. For training and testing, 8,220 and 550 genomes from the NCBI database [138] and Joint Genome Institute database [139]. The metagenomic sequences of [125] and the metagenome of *Pseudoblepharisma tenue* [126] were used as test datasets. Furthermore, it was tested on real data from the Tara Oceans project [140]. According to different *k*-mers lengths, Tiara achieved an average prediction ACC between 98.65% and 98.93% for prokaryotic genomes, and between 95.94%-98.83% for eukaryotic genomes.

DeepVirFinder [100] is proposed to identify viral sequences in metagenomic data using deep learning. CNNs are employed on contigs represented by one-hot encoding and the output is a score between 0 and 1 that indicates the likelihood of being a viral sequence. 2314 reference genomes of prokaryotic viruses were obtained from the NCBI RefSeq database and used and split, according to discovery date, for training, validation and testing steps. When tested on sequences of different lengths, DeepVirFinder achieved an AUROC greater than or equal to 0.93 for each sequence. Also, by applying DeepVirFinder to real human gut metagenomic samples, 51 138 viral sequences belonging to 175 bins were identified.

In [128] two deep learning models are introduced. The first is EdeepVPP, an interpretable CNN for pattern (motif) extraction, which predicts true and pseudo viral sequences. This model consists of a stack of convolution and pooling layers that take one-hot encoding DNA sequences as input and generate probabilities for the classification of true and false viral sequences as output. The EdeepVPP module performs two tasks: prediction of novel viruses and interpretability. The second model, EdeepVPP-hybrid, consists of CNN and LSTM layers to efficiently identify viral genomes. The EdeepVPP-hybrid predictor outperforms ViraMiner [129] and other existing methods by achieving an average AUROC of 0.992 and AUPRC of 0.990 on 19 datasets from human metagenomic contig experiments by 10-fold cross-validation, with the data extracted from [141] and [129].

MLR-OOD [130] is a method based on a deep generative model. The probability of a test sequence belonging to the out-of-distribution (OOD) set of sequences is measured by means of the likelihood ratio of the maximum of the conditional likelihoods of the in-distribution (ID) class and the likelihood of the Markov chain of the test sequence that measures the complexity of the sequence. LSTM is used to classify metagenomic sequences. Three NCBI datasets consisting of bacterial, viral and plasmid sequences were used in training and testing for OOD detection. Two subsets consisted of bacterial sequences: Test2016, which was constructed in [131], and Test2018, which was generated using the same procedure with which the former was constructed. The third subset consisted of 1295 prokaryotic viral genomes and 818 genomes of plasmids. All the data used are available at (https://drive.google.com/drive/folders/1Kz0kQ_D1VWYqA-GDld783O7H8AzNuHkC). It is shown that MLR-OOD achieves good performance for the different types of microbial data.

PhaMer [132] was presented as a Transformer ANN to perform phage contig identification. By constructing a vocabulary of protein groups, contigs are considered as defined phrases in a phage vocabulary and both the composition and protein position of each contig are fed into Transformer networks. Transformer can learn protein organisation and associations using the self-attenuation mechanism and predicts the label for the test contigs. Several datasets were used: a training and a test dataset composed of phages published on RefSeq before and after December 2018, respectively; a short contig test set, produced by randomly cutting the test phage genomes into segments of different lengths: 1, 2, 3, 5, 10, and 15 kbp; a simulated metagenomic dataset, using CAMISIM [142], containing six common bacteria of the human gut; a mock metagenomic dataset, composed of nine metagenomic shotgun sequencing replicates of a simulated [143] community retrieved from the European Nucleotide Archive (BioProject PRJEB19901); and an IMG/VR dataset, extracted from IMG/VR v3 [134]. The results demonstrate that PhaMer can offer improved performance in the identification of new phages.

In this subsection, six tools are described (Table 4), most of which use deep learning models, such as CNN and LSTM, and their validation is mainly performed using viral sequences. One-hot encoding is the preferred method to represent contigs, and their performance compares well with other state-of-the art methods. EdeepVPP addresses the unbalanced data set problem very well. Also noteworthy is the fact that classification ACC is affected by contig length (MLR-OOD and EdeepVPP) and by sequencing errors or viral mutations (DeepVirFinder).

Supervised binning on sequences

VIBRANT [144] was introduced as a workflow that employs a hybrid ML and protein similarity approach, bypassing sequence features, for automated virus retrieval and annotation. It assesses genome quality and integrity, and characterizes viral community functions from metagenomic assemblies. VIBRANT uses a newly developed MLP on protein signatures and a v-score metric that circumvents traditional boundaries to maximise the identification

of lytic viral genomes and integrated proviruses, including highly diverse viruses. To generate the training and test datasets, representative sequences of bacteria, archaea, plasmids, and viruses were downloaded from NCBI's RefSeq and Genbank databases. The final datasets consisted of 400 291 fragments for bacteria/archaea, 14 739 for plasmids and 111 963 for viruses. When applied to 120 834 viral sequences derived from metagenomes representing various human and natural environments, VIBRANT recovered an average of 94% of viruses, outperforming other benchmarked tools.

NLP-MeTaxa [146] was proposed as a method for taxonomic classification at different taxonomic ranks, including species, genus, and other levels. It consists of two main steps. Firstly, the vector representation of DNA fragments is constructed using a Natural Language Processing (NLP) method. Then, these vectors are used as inputs to classify an MLP. The training was done on a dataset of over 14 000 genomes extracted from NCBI RefSeq. For the evaluation, data from the CAMI project [149] and a real dataset, a metagenome sample from the guts of two obese human twins [150], were used. The proposed tool outperforms the others in almost all CAMI data ranges and allocates more reads at the lowest level. Furthermore, in the real case, NLP-MeTaxa was able to predict species that were not predicted by the other tools.

In [151], a model that used single-nucleotide polymorphism (SNP) and *k*-mers frequency based on metagenomic data were presented to predict the Minimum Inhibitory Concentration (MIC) of meropenem against *Klebsiella pneumoniae*. Two ANNs were introduced: 1) a regression deep neural network (DNN) using metagenomic sequencing data to predict the minimum inhibitory concentration of meropenem; and 2) a classification DNN utilizing the same data to determine whether a strain of *Klebsiella pneumoniae* is sensitive or resistant to meropenem. Features from 110 sequenced *K. pneumoniae* genomes were combined and modelled with the EXtreme Gradient Boosting (XGBoost) algorithm. A dense model (an MLP) was used when features were considered separately (SNP or *k*-mer count), and a dense model and an LSTM were used when features were combined. The data used were taken from NCBI BioProjects PRJNA376414, PRJNA386693, and PRJNA396774. DNN regressions for *k*-mers, SNPs, and *k*-mers+SNPs had prediction ACC of 91.89%, 87.05%, and 91.77%, respectively. DNNs perform better in predicting MIC values with improved overall ACC compared to XGBoost models.

CNN-RAI [153] was proposed as a CNN approach based on *k*-mers representation for the metagenomic fragment classification problem. This tool generates a *k*-mers frequency-based DNA representation with Relative Abundance Index (RAI), and then performs a metagenomic fragment classification with a CNN. The effects of different parameters, such as *k*-mers length and read length, on the generated models are examined. Three datasets were used to observe the performance. One was simulated using Illumina technology and contains a total of 757 microbial species originating from 61 genera. The second dataset uses the metagenomic fragments belonging to regions of the 16S rRNA gene to make taxonomic calls available from a Web repository (http://tblab.pa.icar.cnr.it/public/BMC-CIBB_suppl/datasets/). The third database uses long-read metagenomic sequencing with Nanopore technology to detect microbial pathogens in a low complexity community. These were extracted from NCBI SRA ERR1898312, ERR1474981, SRR5117441, SRR5277601, SRR5344355. On account of parallel implementations of deep learning frameworks that can run on GPU hardware, CNN-RAI can run faster with lower memory resource requirements.

VirionFinder [155] was a tool proposed to identify prokaryote virus virion proteins (PVVPs) using the sequence and biochemical properties of amino acids based on a deep learning technique. VirionFinder takes a complete or partial prokaryotic virus protein as input and evaluates whether the given protein is a PVVP. In its process it uses CNNs on the protein sequences that are one-hot encoding and their biochemical properties. For training and testing, all prokaryotic viruses from the NCBI viral database RefSeq (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) were used. The tool was also evaluated on real data of a virome downloaded from NCBI SRA with accession code SRR5224158.1. In the first set of tests, in all cases, VirionFinder performed much better than the other tools. In the real case, VirionFinder identified 76.47% of PVVP.

Taking advantage of recent advances in deep learning for the task of bacteriophage detection, Seeker [133] was presented. It employs models of LSTMs and does not rely on predetermined sequence features. As input, sequences are encoded in two ways, one-hot encoding and integer sequences ('A' = 1, 'T' = 2, 'C' = 3, and 'G' = 4). This tool was trained in two steps. In the first one, all publicly available phage and bacterial genome sequences downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) were used; and, in the second one, all annotated full-genome phages from a comprehensive database search on NCBI and the WEB (https://www.ebi.ac.uk/genomes/phage.html) were used. These sets were divided into training and test sets. The latter set showed an AUROC greater than 0.9. This tool is compared with other tools on datasets from the IMG/VR (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi), and also on short phage and bacterial sequences (1K–5K in length) obtained from NCBI. The metrics used show a competitive True Positive Rate (TPR) when compared to the other tools. Finally, Seeker was used for the discovery of new bacteriophages, with promising results.

Virtifier [162] was proposed as a deep learning-based viral identifier for sequences from metagenomic data. Nucleotide sequences are converted into codon sequences and one-hot encoding. The tool uses Seq2Vec to efficiently extract codon relationships. Then, an LSTM further analyzes codon relationships and sift out the parts that contribute to the final features. All NCBI RefSeq genomes (https://dx.doi.org/10.1186/s40168-017-0283-5) were used for training and testing. Another set of tests was downloaded from the CAMI project (https://data.cami-challenge.org/participate). In addition, real human gut metagenomic sample (https://dx.doi.org/10.1101/gr.142315.112) was used for testing. When applied to the CAMI dataset and real human gut metagenomic sample, Virtifier demonstrates superior performance in identifying viral sequences compared to VirFinder, DeepVirFinder, and PPR-Meta.

BERTax [163] is introduced as a DNN based on NLP for the classification of DNA sequences into three different taxonomic levels: superkingdom, phylum, and genus. The tool assumes that DNA is a 'language' and classifies taxonomic origin based on an understanding of this language rather than using local similarity to known genomes in a database. The genomes of archaea, viruses and bacteria used for training, validation, and testing were downloaded from NCBI and are available from a GitHub repository of BERTax. Although the method has higher average Precision than other comparable methods, the authors show that a combination of approaches, such as MMseqs2 and BERTax, can further improve the prediction results.

The main deep learning models used in the eight tools described in this subsection are LSTM and MLP (Table 5). One-hot encoding is widely used to represent sequences, and the main

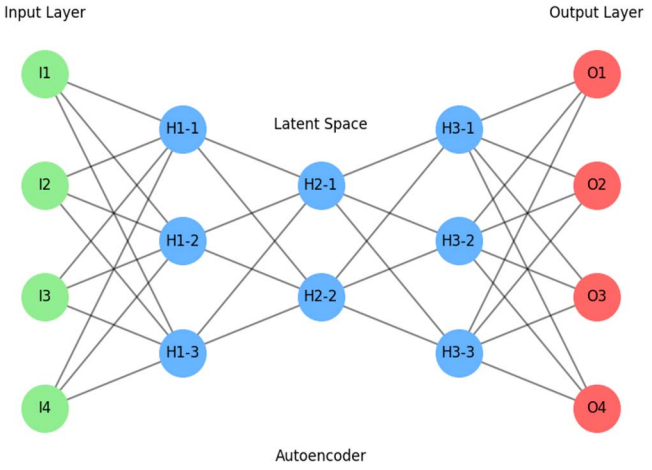


Figure 5. The input layer receives original data. Subsequent encoding layers reduce the dimensionality of data through a series of non-linear transformations, extracting the most relevant features and creating a compressed representation of data (latent space). Decoding layers apply non-linear transformations in an attempt to reconstruct the original input. Finally, the output layer produces the reconstruction of original data.

applications are viral and taxonomic classification. These binning tools have high ACC. Some perform well without substantial computational resouces (e.g. Seeker, CNN-RAI) and some achieve a solid performance in large-scale viral data provided host contamination is absent (e.g. VirionFinder). A decrease in the predictive ability at lower taxonomic ranks was also reported (e.g. BERTax). Training of some model is on small data sets, this may generate overfitting and affect its ability for future predictions (e.g. Predict MICs).

Neural networks for unsupervised binning

Seven tools using ANNs for the unsupervised binning on reads and contigs are presented in this section (Table 6). Tools are described according to the type of ANN, the clustering method, the input features, the comparison models, and the performance metrics. Data sources and algorithms used are also provided.

Unsupervised binning on reads

Autoencoders are deep learning models used to learn efficient representations of data, typically for dimensionality reduction [199]. They compress the input into a latent representation (encoding), and reconstruct the original input from this representation (decoding) (Fig. 5). MetaDEC [166] uses Autoencoder to learn a latent space representation of metagenomic data. It works with compositional features (4-mers) and uses k-means to perform the clustering in the latent space. MetaDEC divides sequences into groups of overlapping sequences and classify them into clusters using adversarial deep clustering technique. Critic network is used to improve the quality of latent representation, and Discriminator network is used to ensure that the generated data points are realistic and that the latent representation is of high quality. The proposed method was tested on 25 simulated datasets [167] and on a real acid mine drainage metagenome [168]. The data are publicly available (https://bioinfolab.fit.hcmute.edu.vn/MetaDEC). In most cases, the algorithm presented performed well and outperformed the others.

Long-read sequencing platforms generate reads that can be longer than the contigs resulting from the assembly of short NGS reads. While most tools were designed to deal with short reads and contigs, the LRBinner algorithm [186] combines composition and coverage information from long read datasets. Composition and coverage vectors are concatenated to form a single 64-dimensional vector. A Variational Autoencoder (VAE) is trained using read batches of size 10 240 for 200 epochs with LeakyRELU activation to reconstruct the entries. VAE weights are learned by backpropagation and stochastic gradient descent. The latent representations obtained by the encoder, on which the clustering is executed, are of lower dimension. A clustering algorithm based on distance histograms is used to extract clusters with different sizes. Four sets were generated using SimLoRD [187] for long reads, and for the case of real data, they were downloaded from NCBI BioProjects ERR3152364, PRJNA546278, and PRJNA680590. Metrics used to compare binning performance show that LRBinner is better.

Unsupervised binning on contigs

AAMB [171] is a tool for contig clustering that uses an Adversarial Autoencoder (AAE) to obtain continuous and discrete latent representations. The contig representation used is a concatenated vector of co-abundances and tetranucleotide frequencies. Clustering in the latent space was performed using iterative medoid clustering. For training and validation, the CAMI [149] and MetaHIT (https://db.cngb.org/search/project/CNPhis0002808/) datasets were used. Furthermore, data from Almeida *et al.* [98] and Price *et al.* [172] were used for addictive evaluations. A higher number of high-quality genomes were recovered using AAMB, as compared to VAMB [173] or other binners. However, since AAMB and VAMB complemented each other well, their combination resulted in the proposed AVAMB tool, which outperforms each of the two separately and requires only slightly more computational power.

An unsupervised binning method based on a Convolutional Autoencoder (CAE) is described in [176]. This neural network is used to extract and reduce features from each contig, which are mathematically represented as vectors that combine in a certain way the tetranucleotide frequency (4-mers), the hexanucleotide frequency (6-mers) and the GC content. In the latent representation of the contigs, clustering is performed with the k-means algorithm. The data used for this study were downloaded from the MyCC section of the SOURCEFORGE website (https://sourceforge.net/projects/sb2nhri/files/MyCC/Data/). Three datasets were chosen: 10s, containing 10 species; 25s, containing 25 species; Sharon, containing 32 species. From the results it can be seen that the proposed tool in the metrics Silhouette index and Rand index achieves a value of 0.5 and 0.8, respectively.

GraphMB [182] uses Graph Neural Networks (GNNs) to incorporate information from the assembly graph and improve binning. A VAE is used for the embedding of contigs, which are then input along with the information from the assembly graph to a GNN and finally run the Iterative k-medoids algorithm. The contigs based on long reads have their composition (k-mers frequency) and coverage extracted. Experiments were performed on a simulated dataset, six wastewater treatment plant datasets [200] and one soil dataset [201]. GraphMB generates on average 17.5% more high-quality bins compared with the state-of-the-art tools.

iDeLUCS [189] is a standalone software tool that exploits the capabilities of deep learning to cluster genomic and metagenomic sequences. iDeLUCS is an enhancement of the DeLUCS tool [190] that employs a deep network. It assigns a cluster identifier to each DNA sequence present in a dataset, while incorporating several visualisation tools. The features extracted from the sequences are the k-mers frequency, and data are clustered using Invariant Information Clustering (IIC). As shown in this study, iDeLUCS

Table 6. Summary of tools that use neural networks for the unsupervised and semi-supervised binning, and the main features

Tool name/URL	Binning type/ANN model/Clustering	Sequence type/Feature/Datasets	Metric (%)/Compared to	Advantages/Disadvantages
MetaDEC [166] https://bioinfolab.fit.hcmute.edu.vn/MetaDEC	Unsupervised/ Autoencoder, Critic, and Discriminator/ *k*-means	Reads/ Composition and Coverage/ Test data: 25 simulated datasets (S1–S10 for short reads and L1–L6 for species with different abundance levels, and R1–R9 for long reads) from [167]. These datasets consist of Illumina short reads (80 bp), Roche 454 single-end long reads. Real data: Acid Mine Drainage from [168], the species include Ferroplasma sp., Leptospirillum sp., *Ferroplasmaacidarmanus*, and others.	Precision was around 98% (dataset S5). Recall was around 92% (dataset R3). F-measure was around 95% (dataset L5)./ Non-deep learning-based methods: BiMeta [167], MetaCluster2 [169], AbundanceBin [52], and MetaCluster5 [170].	Better precision compared to other methods on both short and long read datasets. Performs well on datasets with low-abundance species. Architecture size affects performance and requires careful tuning to avoid over- or under-tuning. High computational resources and long running times.
AAMB [171] https://github.com/RasmussenLab/vamb/tree/master/workflow_avamb	Unsupervised/ AAE/ *k*-medoids	Contigs/ Composition and Co-abundance/ For training and validation, the CAMI project's databases [149] and MetaHIThttps://db.cngb.org/search/project/CNPhis0002808/ were used. In addition, data from Almeida *et al.* [98] and Price *et al.* [172] were used for addictive evaluations.	Performance was evaluated based on the number of reconstructed genomes with precision above 0.9 and recall above 0.9. Number of near-complete (NC): 5077 NC genomes (Almeida dataset). Jaccard index was of 0.64 across all datasets./ VAMB [173], SemiBin [174], SemiBin2 [30] and MetaDecoder [175]	Increased genome recovery, AAMB recovers 7% more NC genomes across synthetic and real datasets. Complementary integration, AAMB is designed to work synergistically with VAMB, yields superior performance. Increased computational cost, AAMB requires significantly more compute power than VAMB. AAMB's bin quality was only slightly better.
MyCC[176] https://sourceforge.net/projects/sb2nhri/files/MyCC/	Unsupervised/ CAE/ *k*-means	Contigs/ Composition/ Test data: Three datasets were chosen: 10s, containing 10 species; 25s, containing 25 species from the MyCC section of the SOURCEFORGE website https://sourceforge.net/projects/sb2nhri/files/MyCC/Data/. Real data: Sharon dataset [177], containing 32 species.	ACC was 86%, Rand Index of 0.95 and a Silhouette Index of 0.60 for the 10s dataset. Precision, Recall, and F1-Score were all around 0.86./ Non-deep learning-baseds: CoMet [178] MetaConClust [179] MetaBAT [180] and MaxBin [181].	The tool achieves promising results with unsupervised methods, performing at par with MetaBAT and MaxBin. The tool provides a novel convolutional autoencoder-based deep clustering technique. 1 Requirement of extensive data for training and expensive GPUs. Organisms with a low contig contribution in the dataset cannot be grouped.

(*Continued*)

Table 6. Continued

Tool name/URL	Binning type/ANN model/Clustering	Sequence type/Feature/Datasets	Metric (%)/Compared to	Advantages/Disadvantages
GraphMB [182] https://github.com/MicrobialDarkMatter/GraphMB	Unsupervised/ VAE + GNN/ *k*-medoids	Contigs/ Composition and Coverage/ Reads generated from 100 microbial strains, corresponding to 50 species, were simulated using the badread tool [183]. Real data: six Wastewater Treatment Plant datasets (PRJNA629478) and a soil sample dataset (PRJEB50688). All data available at https://zenodo.org/records/6122610.	F1-score (0.852), average purity (0.967), and the average completeness (0.762), all on the Strong100 dataset./ Non-deep learning-based method: MetaBAT2 [56], MaxBin2 [55]. Graph-based method: SemiBin [184], VAMB [173], and GraphBin [185].	Integration of assembly graph and contig-specific features. Improved binning of High Quality (HQ) MAGs. Dependent on the quality of the assembly graph. Suboptimal performance on short-read datasets.
LRBinner [186] https://github.com/anuradhawick/LRBinner	Unsupervised/ VAE/ Distance-histogram-based	Reads/ Composition and Coverage/ Test data: four sets (contain 8, 20, 50, and 100 species) were generated using the SimLoRD [187]. Real data: from the NCBI BioProject (numbers ERR3152364, PRJNA546278, and PRJNA680590.)	Precision (99.14%), recall (99.14%), F1-score (99.14%), and Number of bins detected (eight bins), all for the Sim-8 dataset./ Non-deep learning-based method: MetaBCC-LR [188]	LRBinner handles large datasets without any sub-sampling. Better binning results on both simulated and real datasets and it reduces the computational resources required for assembly, Difficulty distinguishing similar species. Seed point dependency in clustering.
iDeLUC [189] https://github.com/Kari-Genomics-Lab/iDeLUCS	Unsupervised/ MLP/ IIC	Contigs/ Composition/ First analyzed 14 real datasets [190]: nine datasets of mitochondrial DNA, two Bacteria data sets, and three viral datasets. In addition, one dataset of microbial genomes [186], and 12 synthetic datasets [191].	ACC (around 98.5%), on synthetic datasets. Davies–Bouldin Index (0.39) and Silhouette Coefficient (0.81) on the Protists dataset. Homogeneity (0.82) and Completeness (0.83), on the Insects dataset./ Non-deep learning-based methods: *k*-means++, Gaussian Mixture Model (GMM), MeShClust v.3.0 [191]. Deep learning-based method: DeLUCS [190].	Accurate and scalable clustering method. Dynamic visualization of the underlying training process. Limitations with large numbers of clusters. It is not tested on datasets with short sequences.

(*Continued*)

Table 6. Continued

Tool name/URL	Binning type/ANN model/Clustering	Sequence type/Feature/Datasets	Metric (%)/Compared to	Advantages/Disadvantages
CoCoNet [192] https://github.com/Puumanamana/CoCoNet	Unsupervised/ MLP + CNN/ Leiden algorithm	Contigs/ Composition and Coverage/ Test data: Synthetic metagenomes generated using genomes from the NCBI RefSeq viral database. These include 9,316 were longer than 3 kb (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/; 12 September 2019). Real data: the Station ALOHA metagenomic dataset. Illumina short reads (NCBI-SRA: SRR10378148, SRR8811962, SRR8811963).	The ACC, F1-score, and AUROC (values above 95%) across the simulated viral datasets. The ARI (a median value of 0.617), homogeneity (0.944), and completeness (0.942) on the simulated data with 2000 genomes./ non-deep learning-based methods: CONCOCT [57] and Metabat2 [56].	Outperforms existing binning methods and robust to coverage variability, particularly in binning viral contigs. Requires fewer resources, computationally efficient. Limited to contigs longer than 2048 bp. The conservative clustering approach used emphasizes bin homogeneity over completeness.
SemiBin [174] https://github.com/BigDataBiology/SemiBin/	Semi-supervised/ Autoencoder/ Infomap algorithm	Contigs/ Composition and Coverage/ Test Data: Simulated CAMI datasets [18, 149]. Real data: German human gut dataset (PRJEB27928, PRJNA290729). Dog gut dataset (PRJEB20308). Marine dataset (PRJEB1787, PRJEB1788, PRJEB4352, and PRJEB4419). Soil dataset (see Supplementary Data 2). non-Westernized African human gut dataset (PRJNA504891).	Completeness (around 98%), purity (around 99%), and F1-score (around 98%), all on the CAMI I low-complexity dataset./ Non-deep learning-based methods: Metabat2 [56], MaxBin2 [55], SolidBin [193], and COCACOLA [194]. deep learning-based method: VAMB [173],	Ability to recover more high-quality bins with larger taxonomic diversity. Can be applied across multiple samples and environments. Computationally expensive, requiring significant time and memory resources. Cannot-link constraints derived from taxonomic annotations introduce dependency on reference genomes.
PhaGCN [69] https://github.com/KennthShang/PhaGCN	Semi-supervised/ MLP + CNN/ Markov clustering algorithm	Contigs/ One-hot/ Test data: one set was randomly sampled contigs from the reference genome (NCBI RefSeq database) and the other were reads simulated with the ART-Illumina [195]. Real data: Caudovirales (SRR1294999983 and SRR13132427).	Macro-ACC and macro-precision values were around 95% in simulated datasets. The macro-recall was around 0.92./ Non-deep learning-based methods: vConTACT 2.0 [196], POGs [197], and ClassiPhage [198]	Superior performance in classifying novel phages. Combining of sequence features and protein similarity. Challenges with short contigs and dependence on reference databases. Computational resource demand.

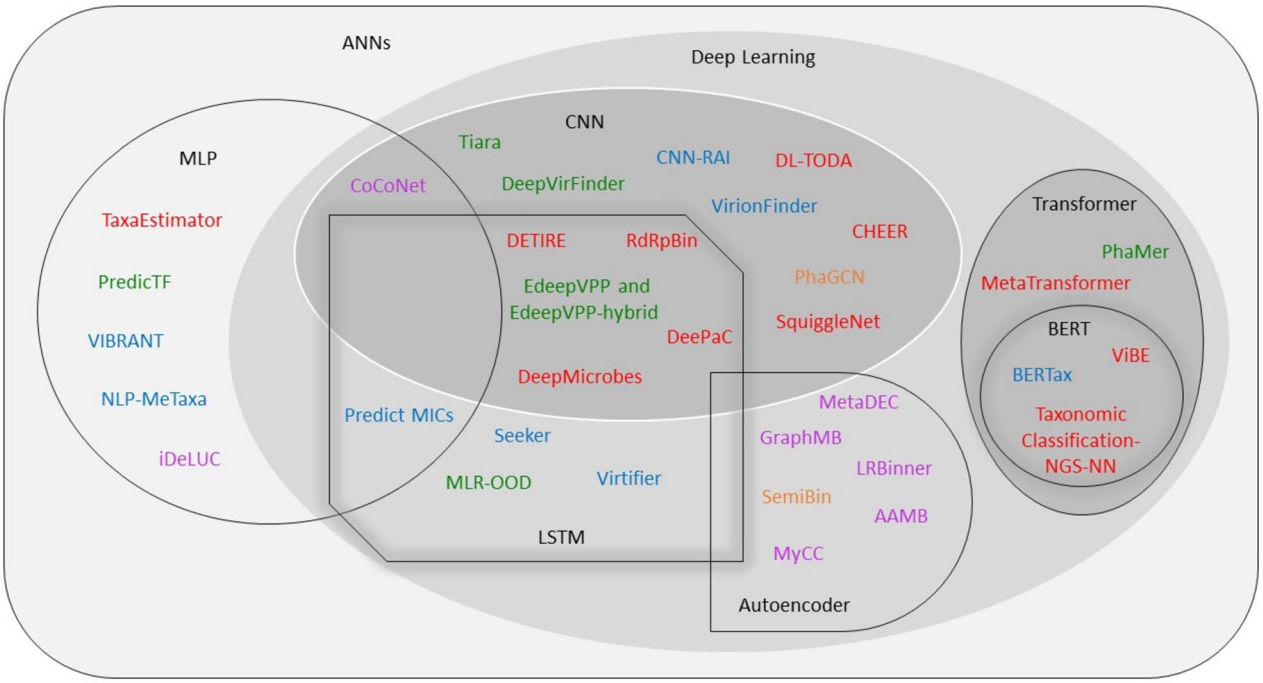


Figure 6. The distribution and complexity of metagenomic binning tools according to the ANNs used. Supervised tools represented in red, green, and blue for reads, contigs, and general sequences, respectively. Unsupervised tools are represented in purple and semi-supervised tools in orange.

outperforms other algorithms in clustering large datasets of unlabelled DNA sequences.

Deep learning combined with clustering was used to group contigs into homogeneous clusters representing species. This approach was implemented in CoCoNet [192], whose two-phase workflow begins with contig splitting into 1024-bp fragments spaced with a 128-bp step. Using these partitions, a DNN is trained to estimate the probability of two fragments belonging to the same genome, given their composition and coverage information. In the second phase, a computationally tractable heuristic is used to cluster contigs using the co-occurrence probabilities inferred in the previous phase. Finally, the Leiden clustering algorithm is used. For the simulations, four synthetic metagenomes generated with CAMISIM [142] were used and, for the experiments, three viral metagenomes collected at the ALOHA station in the North Pacific Subtropical Gyre [202] were used. The results show that CoCoNet outperforms the other tools on both simulated and real viral datasets.

Among the seven unsupervised binning tools described in this subsection (Table 6), the most used model is the autoencoder and its variants, while the most used features are *k*-mers frequency and coverage information. The *k*-means algorithm is a state-of-the-art clustering method. Generally, unsupervised binning tools show good performance. Most tools do not automatically estimate the number of clusters, and few of them seek an enhanced interpretability by assigning confidence scores (iDeLUCS) based on assembly graph information (GraphMB). A single tool takes long reads as an input (LRBinner). Low abundant taxa likely affect the classification performance of some tools (e.g. MyCC).

Neural networks for semi-supervised binning

SemiBin [174] uses deep siamese ANNs to implement a semi-supervised approach, i.e. it exploits information from reference genomes while retaining the ability to reconstruct high-quality bins outside the reference dataset. It first uses a deep siamese ANN (a network consisting of two identical sub-networks with the same weights and processed in parallel) for embedding of features of contigs (with autoencoder) and then the Infomap algorithm [203] for clustering. The features used were *k*-mers frequency and coverage information. Synthetic data from the CAMI project [18, 149] and six real datasets extracted from [63, 204–208] were used to evaluate the performance. The results show that SemiBin outperforms other state-of-the-art methods and highlight the advantages of using supervised knowledge, such as contig annotation and the use of single-copy marker genes, in metagenomic binning.

In light of the considerable viral diversity and the vast quantity of unlabelled phages, PhaGCN [69] emerges as a tool that employs a semi-supervised learning model to perform the taxonomic classification of phage contigs. In this learning model, a knowledge graph is constructed by combining the skip-gram embedding of DNA sequence features, which have been learned by a CNN, with the protein sequence similarity obtained from the gene exchange network. Subsequently, a GCN is employed to leverage both labelled and unlabelled samples in the training process, thereby enhancing the learning capacity. PhaGCN was evaluated on a simulated dataset comprised of randomly sampled contigs from a reference genome and contigs generated using ART-Illumina [195], and on a real dataset (NCBI SRA SRR1294999983 and SRR13132427). Additionally, an original dataset was utilized. The findings demonstrate that the proposed method is comparable to other available phage classification tools.

Discussion

Metagenomics is a promising field in scientific research, and binning, in particular, faces several challenges, such as the complexity and heterogeneity of metagenomic data, making the selection

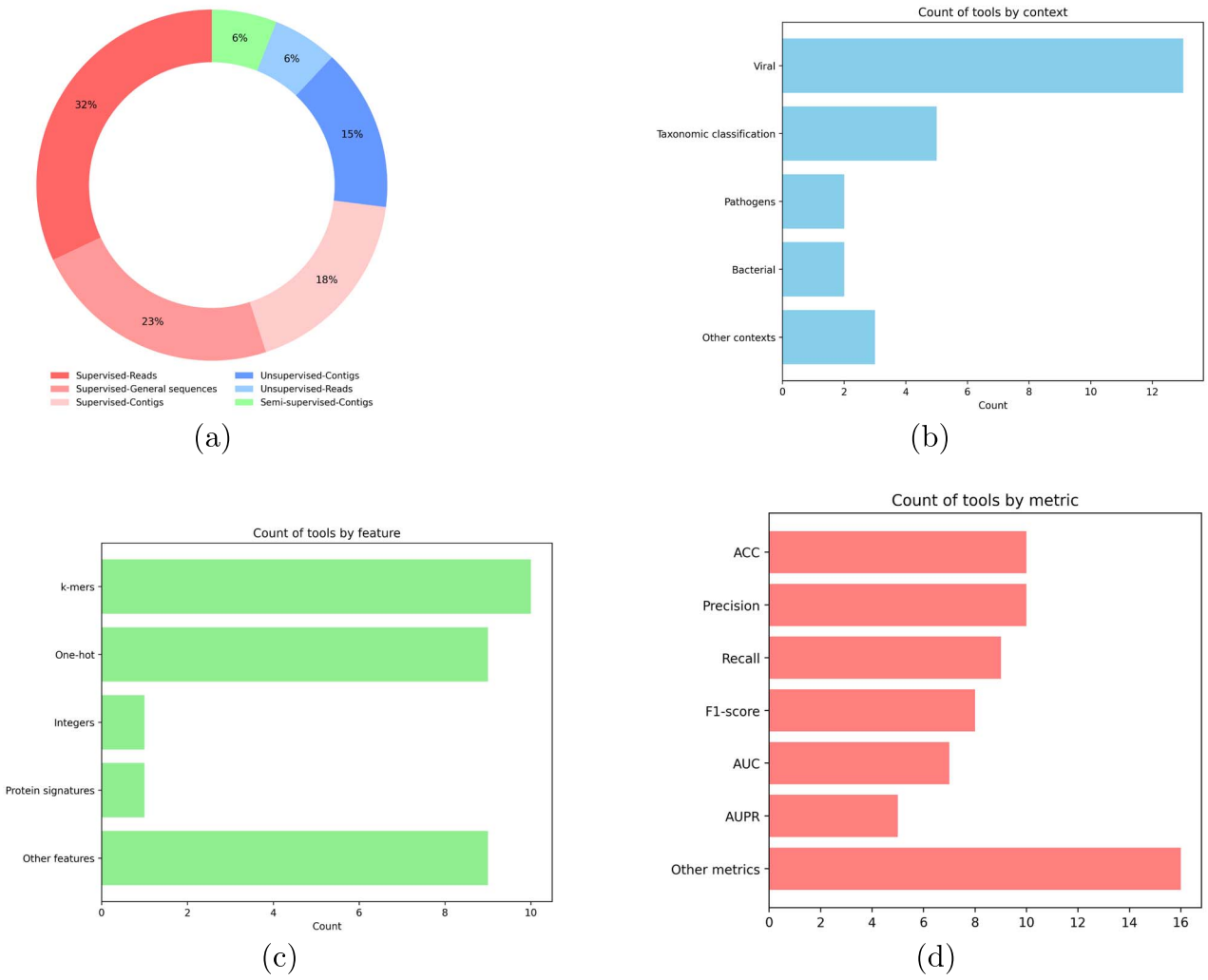


Figure 7. (a) Distribution of ANNs found in this review for different types of binning and types of sequences. For the metagenomic binning tools described in this review (Sections 3, 4, and 5): (b) The frequency with which contexts are encountered for supervised tools (4 main contexts plus 'Other contexts', including 3 additional contexts); (c) Frequency of occurrence of features (4 main features plus 'Other features', including 9 additional features); and (d) Frequency of occurrence of metrics (6 main metrics plus 'Other metrics', including 16 additional metrics).

of appropriate tools or methods (Table 7) a challenging process. The features representing the sequences and sequence length can be a determining factor in the ACC of binning tools. With the above challenges in mind, we review 34 tools that use ANNs in metagenomic binning and summarize their specific characteristics, with the aim of assisting the reader in identifying the most suitable tool for their specific research objectives (Sections 3, 4, and 5). All codes are freely available and most were validated with real data.

Deep learning is the new trend in ANNs for metagenomic binning. It develops a low-dimensional representation or embedding of sequence features in order to generate more accurate bins. In supervised contexts, CNN and LSTM stand out as the most widely used, while in unsupervised, AEs predominate (Fig. 6). In the semi-supervised case, only a couple of models were identified, one of which also utilizes AEs (Section 5). In comparison to other techniques, deep learning has emerged as a significant contender in the domain of metagenomics. However, metagenomic data impose several constraints that DL models may have trouble handling. Their heterogeneity and the errors that can be introduced in downstream processes, including sample collection, sequencing and assembly, can affect binning ACC [209]. Additionally, computational requirements of deep learning are often expensive, especially when hybrid models are used (Table 7).

Most of the tools analysed (73%) perform supervised binning (Fig. 7(a)) and show high performance, though their results depend on the quality and quantity of data with which they are trained. Generally, most tools take contigs as input, but within the supervised tools, most take reads (Fig. 7(a)). Supervised tools are often (48%) applied to the viral context, neglecting other contexts, e.g. bacterial and pathogenic ones (Fig. 7(b)). Among the supervised tools, the most commonly used input types are *k*-mers frequency vectors or one-hot encoding (Fig. 7(c)) and the most frequently employed metrics are ACC, Precision, Recall, and F1-score. (Fig. 7(d)).

Regarding datasets, we observed that there are no standardized sources of data for neural network training in the different metagenomic contexts reviewed. Most tools are trained and tested using unique datasets generated by their own developers. The lack of standardization makes it difficult to compare tools in an objective way. Nevertheless, relying on data from regularly updated datasets, such as NCBI RefSeq Virus, is also challenging, as the database change over time. Thus, initiatives such as the

Table 7. Summary of some relevant characteristics of the different ANN models most used by the metagenomic binning tools described

Characteristics	MLP	CNN	LSTM	Autoencoder
Strengths	Simple model. Vector data.	Extraction of spatial features. Matrix data.	Extraction of sequential features. Vector and matrix data.	Dimensionality reduction. Vector and matrix data.
Weaknesses	Overfitting	Computationally expensive	Slow training	Interpretability
Labeled data	Yes	Yes	Yes	No
ACC	Good	High	High	–
Computational cost	Medium	High	Very high	High
Metagenomic sequence types most used	Supervised-General sequences	Supervised-Reads	Supervised-Varied	Unsupervised-Contigs
Metagenomic contexts employed	Varied	Mostly viral	Mostly viral	MGS reconstruction
Optimization algorithms used by the described tools	Stochastic gradient descent	Adam optimizer and stochastic gradient descent	Adam optimizer	Adam optimizer

CAMI contain robust datasets with the corresponding ground truth, which could facilitate the standardization in metagenomic binning methods development.

The main problems reported by the developers of ANN-based tools refer to the amount of data and computational power required for training. On the other hand, a model with high generalization power can be used to analyze real world datasets with high accuracy. Regarding method performance, each tool presents advantages and shortcomings, which are pointed out systematically in Tables 3, 4, 5, and 6.

The main aspect highlighted as an advantage refers to accuracy, and generally ANN-based tools outperform traditional alignment-based tools. Also, a well-known aspect of ANNs is the ability to deal with large amounts of data in reduced time.

Metagenomics has greatly benefited from ANNs, and some future directions in metagenomic binning research using ANNs are listed below:

- To further explore and adapt emerging architectures, such as BERT, in supervised binning.
- To optimise and develop new variants of autoencoders that can better handle the complexity and diversity of metagenomic data in unsupervised binning, but also explore new alternatives.
- To create standardised datasets and perform more objective benchmarking (e.g. CAMI project [18]).
- To develop binning tools focused on third-generation reads.
- To consider the effect of sequencing, assembly errors, and low abundant taxa.

Key Points

- Papers on neural networks for metagenomic binning are revised.
- Supervised, unsupervised and semi-supervised binning methods are considered.
- Type of sequences for metagenomic binning are addressed.
- The source code of the tools, data source and metrics used for metagenomic binning are provided.
- Research trends on metagenomic binning are presented.

Funding

J.H.-A. was funded by the National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2022-21221825 and by the 'Doctorado en Modelamiento Matemático Aplicado' Program. M.M. acknowledges support from ANID Fondecyt 1200810 grant. S.C.-O. acknowledges support from ANID Fondecyt 1201692, 1211515, 1240871, and 1181773, and Fondequip EQM210185 grants.

Author contributions

J.H.-A., M.M., and S.C.-O. conceived the paper; J.H.-A. wrote the draft paper; M.M., S.C.-O., K.V., and R.H. reviewed the manuscript; and J.H.-A. wrote the final version of the paper.

References

1. Ogbe RJ, Ochalefu DO, Olaniru OB. Bioinformatics advances in genomics-a review. *Int J Curr Res Rev* 2016;**8**:5.
2. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome* 2015;**3**:31. https://doi.org/10.1186/s40168-015-0094-5
3. Hugenholtz P, Tyson GW. Metagenomics. *Nature* 2008;**455**:481–3. https://doi.org/10.1038/455481a
4. Yen S, Johnson JS. Metagenomics: a path to understanding the gut microbiome. *Mamm Genome* 2021;**32**:282–96. https://doi.org/10.1007/s00335-021-09889-x
5. Vester JK, Glaring MA, Stougaard P. Improved cultivation and metagenomics as new tools for bioprospecting in cold environments. *Extremophiles* 2015;**19**:17–29. https://doi.org/10.1007/s00792-014-0704-3
6. Thippeswamy M, Rajasrerlatha V, Shubha D. *et al.* Metagenomics and future perspectives in discovering pollutant degrading enzymes from soil microbial communities. In: Viswanath B. (ed.), *Recent Developments in Applied Microbiology and Biochemistry*. Academic Press; 2021, pp. 257–67.
7. Sousa J, Silvério SC, Costa AMA. *et al.* Metagenomic approaches as a tool to unravel promising biocatalysts from natural resources: soil and water. *Catalysts* 2022;**12**:385. https://doi.org/10.3390/catal12040385
8. Kamble P, Vavilala SL. Discovering novel enzymes from marine ecosystems: a metagenomic approach. *Bot Mar* 2018;**61**:161–75. https://doi.org/10.1515/bot-2017-0075

9. de Fátima L, Alves CA, Westmann GL. *et al.* Metagenomic approaches for understanding new concepts in microbial science. *Int J Genom* 2018;**2018**:1–15. https://doi.org/10.1155/2018/2312987
10. Taş N, de Jong AEE, Li Y. *et al.* Metagenomic tools in microbial ecology research. *Curr Opin Biotechnol* 2021;**67**:184–91. https://doi.org/10.1016/j.copbio.2021.01.019
11. Handel AS, Muller WJ, Planet PJ. Metagenomic next-generation sequencing (mNGS): SARS-CoV-2 as an example of the technology's potential pediatric infectious disease applications. *J Pediatric Infect Dis Soc* 2021;**10**:S69–70. https://doi.org/10.1093/jpids/piab108
12. Wei G, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu Rev Pathol* 2019;**14**:319–38. https://doi.org/10.1146/annurev-pathmechdis-012418-012751
13. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. *Brief Bioinform* 2021;**22**:178–93. https://doi.org/10.1093/bib/bbz155
14. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* 2014;**5**:209. https://doi.org/10.3389/fpls.2014.00209
15. Dröge J, McHardy AC. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform* 2012;**13**:646–55. https://doi.org/10.1093/bib/bbs031
16. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform* 2012;**13**:669–81. https://doi.org/10.1093/bib/bbs054
17. Sedlar K, Kupkova K, Provaznik I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* 2017;**15**:48–55. https://doi.org/10.1016/j.csbj.2016.11.005
18. Meyer F, Fritz A, Deng Z-L. *et al.* Critical assessment of metagenome interpretation: the second round of challenges. *Nat Methods* 2022;**19**:429–40. https://doi.org/10.1038/s41592-022-01431-4
19. Escobar-Zepeda A, de León AV-P, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front Genet* 2015;**6**:348. https://doi.org/10.3389/fgene.2015.00348
20. Hiraoka S, Yang C-c, Iwasaki W. Metagenomics and bioinformatics in microbial ecology: current status and beyond. *Microbes Environ* 2016;**31**:204–12. https://doi.org/10.1264/jsme2.ME16024
21. Jandaik S, Gautam HR, Sheela J. *et al.* An overview of metagenomics: the current era of bio-analysis. *Octa J Biosci* 2022;**10**:13–26.
22. Nam NN, Do HDK, Trinh KTL. *et al.* Metagenomics: an effective approach for exploring microbial diversity and functions. *Foods* 2023;**12**:2140. https://doi.org/10.3390/foods12112140
23. Wang Y, Yulin Zhang YH, Liu L. *et al.* Genome-centric metagenomics reveals the host-driven dynamics and ecological role of cpr bacteria in an activated sludge system. *Microbiome* 2023;**11**:56.
24. Gounot J-S, Chia M, Bertrand D. *et al.* Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in southeast asians. *Nat Commun* 2022;**13**:6044.
25. Zhu Q, Mai U, Pfeiffer W. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat Commun* 2019;**10**:5477. https://doi.org/10.1038/s41467-019-13443-4
26. Bowers RM, Kyrpides NC, Stepanauskas R. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31. https://doi.org/10.1038/nbt.3893
27. Orakov A, Fullam A, Coelho LP. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* 2021;**22**:1–19.
28. Riley R, Bowers RM, Camargo AP. *et al.* Terabase-scale coassembly of a tropical soil microbiome. *Microbiol Spectr* 2023;**11**:e00200–23. https://doi.org/10.1128/spectrum.00200-23
29. Gao B, Chi L, Zhu Y. *et al.* An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. *Biomolecules* 2021;**11**:530. https://doi.org/10.3390/biom11040530
30. Pan S, Zhao X-M, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short-and long-read sequencing. *Bioinformatics* 2023;**39**:21–29. https://doi.org/10.1093/bioinformatics/btad209
31. Breitwieser FP, Jennifer L, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2019;**20**:1125–36. https://doi.org/10.1093/bib/bbx120
32. Yang C, Chowdhury D, Zhang Z. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J* 2021;**19**:6301–14. https://doi.org/10.1016/j.csbj.2021.11.028
33. Zhou Y, Liu M, Yang J. Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: methods, applications, challenges, and opportunities. *Microbiol Res* 2022;**260**:127023. https://doi.org/10.1016/j.micres.2022.127023
34. Sieber CMK, Probst AJ, Sharrar A. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;**3**:836–43. https://doi.org/10.1038/s41564-018-0171-1
35. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018;**6**:1–13.
36. Mathieu A, Leclercq M, Sanabria M. *et al.* Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. *Front Microbiol* 2022;**13**:811495. https://doi.org/10.3389/fmicb.2022.811495
37. Yang GR, Wang X-J. Artificial neural networks for neuroscientists: a primer. *Neuron* 2020;**107**:1048–70. https://doi.org/10.1016/j.neuron.2020.09.005
38. Choi RY, Coyner AS, Kalpathy-Cramer J. *et al.* Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol* 2020;**9**:14–4.
39. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial genomics* 2020;**6**:000409. https://doi.org/10.1099/mgen.0.000409
40. Gwak H-J, Lee SJ, Rho M. Application of computational approaches to analyze metagenomic data. *J Microbiol* 2021;**59**:233–41. https://doi.org/10.1007/s12275-021-0632-8
41. Miller S, Naccache SN, Samayoa E. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res* 2019;**29**:831–42. https://doi.org/10.1101/gr.238170.118
42. Skewes-Cox P, Sharpton TJ, Pollard KS. *et al.* Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PloS One* 2014;**9**:e105067. https://doi.org/10.1371/journal.pone.0105067

43. Soueidan H, Nikolski M. Machine learning for metagenomics: methods and tools. *Metagenomics* 2016;**1**:1–19. https://doi.org/10.1515/metgen-2016-0001
44. Ye Simon H, Siddle KJ, Park DJ. *et al.* Benchmarking metagenomics tools for taxonomic classification. *Cell* 2019;**178**:779–94. https://doi.org/10.1016/j.cell.2019.07.010
45. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep* 2016;**6**:19233. https://doi.org/10.1038/srep19233
46. McHardy AC, Rigoutsos I. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol* 2007;**10**:499–503.
47. Siegel K, Altenburger K, Hon Y-S. *et al.* Puzzlecluster: a novel unsupervised clustering algorithm for binning DNA fragments in metagenomics. *Current Bioinformatics* 2015;**10**:231–52. https://doi.org/10.2174/1574893610021505181507 16
48. Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol* 2007;**5**:e82. https://doi.org/10.1371/journal.pbio.0050082
49. Saeed I, Tang S-L, Halgamuge SK. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* 2012;**40**:e34–4. https://doi.org/10.1093/nar/gkr1204
50. Strous M, Kraft B, Bisdorf R. *et al.* The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* 2012;**3**:410. https://doi.org/10.3389/fmicb.2012.00410
51. Laczny CC, Sternal T, Plugaru V. *et al.* Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* 2015;**3**:1–7. https://doi.org/10.1186/s40168-014-0066-1
52. Wu Y-W, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* 2011;**18**:523–34. https://doi.org/10.1089/cmb.2010.0245
53. Bjørn Nielsen H, Almeida M, Juncker AS. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;**32**:822–8. https://doi.org/10.1038/nbt.2939
54. Wang Y, Haiyan H, Li X. MBBC: an efficient approach for metagenomic binning based on clustering. *BMC Bioinform* 2015;**16**:1–11.
55. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;**32**:605–7. https://doi.org/10.1093/bioinformatics/btv638
56. Kang DD, Li F, Kirton E. *et al.* Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. https://doi.org/10.7717/peerj.7359
57. Alneberg J, Bjarnason BS, De Bruijn I. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;**11**:1144–6. https://doi.org/10.1038/nmeth.3103
58. Wang Z, Huang P, You R. *et al.* MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol* 2023;**24**:1.
59. Wajid B, Anwar F, Wajid I. *et al.* Music of metagenomics—a review of its applications, analysis pipeline, and associated tools. *Funct Integr Genomics* 2022;**22**:1–24. https://doi.org/10.1007/s10142-021-00810-y
60. Zeng S, Patangia D, Almeida A. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun* 2022;**13**:5139. https://doi.org/10.1038/s41467-022-32805-z
61. Lavrinienko A, Tukalenko E, Mousseau TA. *et al.* Two hundred and fifty-four metagenome-assembled bacterial genomes from the bank vole gut microbiota. *Sci Data* 2020;**7**:312. https://doi.org/10.1038/s41597-020-00656-2
62. Chen C, Zhou Y, Hao F. *et al.* Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat Commun* 2021;**12**:1106. https://doi.org/10.1038/s41467-021-21295-0
63. Pasolli E, Asnicar F, Manara S. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**:649–62. https://doi.org/10.1016/j.cell.2019.01.001
64. Li D, Liu C-M, Luo R. *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6. https://doi.org/10.1093/bioinformatics/btv033
65. Parks DH, Imelfort M, Skennerton CT. *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55. https://doi.org/10.1101/gr.186072.114
66. Yang X, Song Z, King I. *et al.* A survey on deep semi-supervised learning. *IEEE Trans Knowl Data Eng* 2022;**35**:8934–54.
67. Ouali Y, Hudelot C, Tami M. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*. 2020.
68. Zha Y, Chong H, Yang P. *et al.* Microbial dark matter: from discovery to applications. *Genom Proteom Bioinform* 2022;**20**:867–81. https://doi.org/10.1016/j.gpb.2022.02.007
69. Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* 2021;**37**:i25–33. https://doi.org/10.1093/bioinformatics/btab293
70. Kohls M, Kircher M, Krepel J. *et al.* Correcting the estimation of viral taxa distributions in next-generation sequencing data after applying artificial neural networks. *Genes* 2021;**12**:1755. https://doi.org/10.3390/genes12111755
71. Rodney Brister J, Ako-Adjei D, Bao Y. *et al.* Ncbi viral genomes resource. *Nucleic Acids Res* 2015;**43**:D571–7. https://doi.org/10.1093/nar/gku1207
72. Rosales SM, Thurber RV. Brain meta-transcriptomics from harbor seals to infer the role of the microbiome and virome in a stranding event. *PloS One* 2015;**10**:e0143944. https://doi.org/10.1371/journal.pone.0143944
73. Shang J, Sun Y. Cheer: hierarchical taxonomic classification for viral metagenomic data via deep learning. *Methods* 2021;**189**:95–103. https://doi.org/10.1016/j.ymeth.2020.05.018
74. Wang Q, Garrity GM, Tiedje JM. *et al.* Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;**73**:5261–7. https://doi.org/10.1128/AEM.00062-07
75. Fiannaca A, La Paglia L, La Rosa M. *et al.* Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform* 2018;**19**:61–76.
76. Zhao G, Wu G, Lim ES. *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* 2017;**503**:21–30. https://doi.org/10.1016/j.virol.2017.01.005
77. Li H. wgsim-read simulator for next generation sequencing. Github Repository; 2011. http://github.com/lh3/wgsim
78. Cres CM, Tritt A, Bouchard KE. *et al.* DL-TODA: a deep learning tool for omics data analysis. *Biomolecules* 2023;**13**:585. https://doi.org/10.3390/biom13040585

79. Wood DE, Jennifer L, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol* 2019;**20**:1–13. https://doi.org/10.1186/s13059-019-1891-0
80. Kim D, Song L, Breitwieser FP. *et al.* Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;**26**:1721–9. https://doi.org/10.1101/gr.210641.116
81. Shaiber A, Willis AD, Delmont TO. *et al.* Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* 2020;**21**:1–35. https://doi.org/10.1186/s13059-020-02195-w
82. Bao Y, Wadden J, Erb-Downward JR. *et al.* SquiggleNet: real-time, direct classification of nanopore signals. *Genome Biol* 2021;**22**:1–16.
83. Payne A, Holmes N, Clarke T. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* 2021;**39**:442–50. https://doi.org/10.1038/s41587-020-00746-x
84. Kovaka S, Fan Y, Ni B. *et al.* Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled. *Nat Biotechnol* 2021;**39**:431–41. https://doi.org/10.1038/s41587-020-0731-9
85. Bartoszewicz JM, Seidel A, Rentzsch R. *et al.* DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics* 2020;**36**:81–9. https://doi.org/10.1093/bioinformatics/btz541
86. Altschul SF, Gish W, Miller W. *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
87. Deneke C, Rentzsch R, Renard BY. PaPrBaG: a machine learning approach for the detection of novel pathogens from ngs data. *Sci Rep* 2017;**7**:39194. https://doi.org/10.1038/srep39194
88. Barash E, Sal-Man N, Sabato S. *et al.* BacPaCS—bacterial pathogenicity classification via sparse-svm. *Bioinformatics* 2019;**35**:2001–8. https://doi.org/10.1093/bioinformatics/bty928
89. Cosentino S, Larsen MV, Aarestrup FM. *et al.* Pathogenfinder-distinguishing friend from foe using bacterial whole genome sequence data. *PloS One* 2013;**8**:e77302. https://doi.org/10.1371/journal.pone.0077302
90. Chen I-MA, Chu K, Palaniappan K. *et al.* IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 2019;**47**:D666–77. https://doi.org/10.1093/nar/gky901
91. Manara S, Pasolli E, Dolce D. *et al.* Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of staphylococcus aureus strains in a paediatric hospital. *Genome Med* 2018;**10**:1–19. https://doi.org/10.1186/s13073-018-0593-7
92. Liang Q, Bible PW, Liu Y. *et al.* Deepmicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom Bioinform* 2020;**2**:lqaa009. https://doi.org/10.1093/nargab/lqaa009
93. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:1–12.
94. Ounit R, Wanamaker S, Close TJ. *et al.* CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 2015;**16**:1–13.
95. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* 2016;**32**:3823–5. https://doi.org/10.1093/bioinformatics/btw542
96. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 2016;**7**:11257.
97. Huson DH, Beier S, Flade I. *et al.* Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol* 2016;**12**:e1004957. https://doi.org/10.1371/journal.pcbi.1004957
98. Almeida A, Mitchell AL, Boland M. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 2019;**568**:499–504. https://doi.org/10.1038/s41586-019-0965-1
99. Miao Y, Bian J, Dong G. *et al.* DETIRE: a hybrid deep learning model for identifying viral sequences from metagenomes. *Front Microbiol* 2023;**14**:1169791. https://doi.org/10.3389/fmicb.2023.1169791
100. Ren J, Song K, Deng C. *et al.* Identifying viruses from metagenomic data using deep learning. *Quant Biol* 2020 Cited By:149;**8**:64–77. https://doi.org/10.1007/s40484-019-0187-4
101. Fang Z, Tan J, Wu S. *et al.* PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* 2019;**8**:giz066. https://doi.org/10.1093/gigascience/giz066
102. Amgarten D, Braga LPP, Da Silva AM. *et al.* Marvel, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front Genet* 2018;**9**:304. https://doi.org/10.3389/fgene.2018.00304
103. Tang X, Shang J, Sun Y. RdRp-based sensitive taxonomic classification of RNA viruses for metagenomic data. *Brief Bioinform* 2022;**23**:bbac011. https://doi.org/10.1093/bib/bbac011
104. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using diamond. *Nat Methods* 2021;**18**:366–8. https://doi.org/10.1038/s41592-021-01101-x
105. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. https://doi.org/10.1038/nbt.3988
106. Mistry J, Chuguransky S, Williams L. *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913
107. Horton T, Kroh A, Ahyong S. *et al. World Register of Marine Species (WoRMS)* 2021. https://www.marinespecies.org (27 October 2021, date last accessed).
108. Wichmann A, Buschong E, Müller A. *et al.* MetaTransformer: deep metagenomic sequencing read classification using self-attention models. *NAR Genom Bioinform* 2023;**5**:lqad082. https://doi.org/10.1093/nargab/lqad082
109. Gwak H-J, Rho M. ViBE: a hierarchical BERT model to identify eukaryotic viruses using metagenome sequencing data. *Brief Bioinform* 2022;**23**:bbac204. https://doi.org/10.1093/bib/bbac204
110. Roux S, Enault F, Hurwitz BL. *et al.* VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015;**3**:e985. https://doi.org/10.7717/peerj.985
111. Beghini F, McIver LJ, Blanco-Míguez A. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 2021;**10**:e65088. https://doi.org/10.7554/eLife.65088
112. Voigt B, Fischer O, Krumnow C. *et al.* NGS read classification using AI. *PloS One* 2021;**16**:e0261548. https://doi.org/10.1371/journal.pone.0261548
113. Al-Ajlan A, El Allali A. CNN-MGP: convolutional neural networks for metagenomics gene prediction. *Interdiscip Sci: Comput Life Sci* 2019;**11**:628–35. https://doi.org/10.1007/s12539-018-0313-4
114. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 2010;**38**:e191–1. https://doi.org/10.1093/nar/gkq747
115. Ke-Lin D, Leung C-S, Mow WH. *et al.* Perceptron: learning, generalization, model selection, fault tolerance, and role in the deep learning era. *Mathematics* 2022;**10**:4730.

116. Zhao X, Wang L, Zhang Y. *et al.* A review of convolutional neural networks in computer vision. *Artif Intell Rev* 2024;**57**:99. https://doi.org/10.1007/s10462-024-10721-6
117. O'Dwyer DN, Ashley SL, Gurczynski SJ. *et al.* Lung microbiota contribute to pulmonary inflammation and disease progression in pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;**199**:1127–38. https://doi.org/10.1164/rccm.201809-1650OC
118. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* 2020;**404**:132306. https://doi.org/10.1016/j.physd.2019.132306
119. Cock PJA, Antao T, Chang JT. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3. https://doi.org/10.1093/bioinformatics/btp163
120. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15. https://doi.org/10.1093/nar/gky1049
121. O'Leary NA, Wright MW, Brister JR. *et al.* Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45. https://doi.org/10.1093/nar/gkv1189
122. Monteiro LMO, Saraiva JP, Toscan RB. *et al.* PredicTF: prediction of bacterial transcription factors in complex microbial communities using deep learning. *Environ Microbiome* 2022;**17**:1–11.
123. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9. https://doi.org/10.1093/bioinformatics/btu153
124. Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 2022 Cited By:8; **38**:344–50. https://doi.org/10.1093/bioinformatics/btab672
125. West PT, Probst AJ, Grigoriev IV. *et al.* Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* 2018;**28**:569–80. https://doi.org/10.1101/gr.228429.117
126. Muñoz-Gómez SA, Kreutz M, Hess S. A microbial eukaryote with a unique combination of purple bacteria and green algae as endosymbionts. *Sci Adv* 2021;**7**:eabg4102. https://doi.org/10.1126/sciadv.abg4102
127. Ren J, Ahlgren NA, Yang Young L. *et al.* VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017;**5**:1–20.
128. Dasari CM, Bhukya R. Explainable deep neural networks for novel viral genome prediction. *Appl Intell* 2022 Cited By:12; **52**:3002–17. https://doi.org/10.1007/s10489-021-02572-3
129. Tampuu A, Bzhalava Z, Dillner J. *et al.* ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PloS One* 2019;**14**:e0222271. https://doi.org/10.1371/journal.pone.0222271
130. Bai X, Ren J, Sun F. MLR-OOD: a Markov chain based likelihood ratio method for out-of-distribution detection of genomic sequences. *J Mol Biol* 2022;**434**:167586. https://doi.org/10.1016/j.jmb.2022.167586
131. Ren J, Liu PJ, Fertig E. *et al.* Likelihood ratios for out-of-distribution detection. *Adv Neural Inf Process Syst* 2019;**32**:14680–91.
132. Shang J, Tang X, Guo R. *et al.* Accurate identification of bacteriophages from metagenomic data using Transformer. *Brief Bioinform* 2022;**23**:bbac258. https://doi.org/10.1093/bib/bbac258
133. Auslander N, Gussow AB, Benler S. *et al. Nucleic Acids Res* 2020;**48**:E121. Cited By:34
134. Roux S, Páez-Espino D, Chen I-MA. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res* 2021;**49**:D764–75. https://doi.org/10.1093/nar/gkaa946
135. Arango-Argoty G, Garner E, Pruden A. *et al.* Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 2018;**6**:1–15.
136. Kiliç S, White ER, Sagitova DM. *et al.* CollecTF: a database of experimentally validated transcription factor-binding sites in bacteria. *Nucleic Acids Res* 2014;**42**:D156–60. https://doi.org/10.1093/nar/gkt1123
137. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9. https://doi.org/10.1093/nar/gkaa1100
138. Sayers EW, Agarwala R, Bolton EE. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2019;**47**:D23–8. https://doi.org/10.1093/nar/gky1069
139. Grigoriev IV, Nordberg H, Shabalov I. *et al.* The genome portal of the department of energy joint genome institute. *Nucleic Acids Res* 2012;**40**:D26–32. https://doi.org/10.1093/nar/gkr947
140. Pesant S, Not F, Picheral M. *et al.* Open science resources for the discovery and analysis of tara oceans data. *Sci Data* 2015;**2**:1–16. https://doi.org/10.1038/sdata.2015.23
141. Nowicki M, Bzhalava D, Bała P. Massively parallel implementation of sequence alignment with basic local alignment search tool using parallel computing in Java library. *J Comput Biol* 2018;**25**:871–81. https://doi.org/10.1089/cmb.2018.0079
142. Fritz A, Hofmann P, Majda S. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 2019;**7**:1–12. https://doi.org/10.1186/s40168-019-0633-6
143. Kleiner M, Thorson E, Sharp CE. *et al.* Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun* 2017;**8**:1558. https://doi.org/10.1038/s41467-017-01544-x
144. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:1–23.
145. He Q, Gao Y, Jie Z. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* 2017;**6**:1–11. https://doi.org/10.1093/gigascience/gix050
146. Matougui B, Boukelia A, Belhadef H. *et al.* NLP-MeTaxa: a natural language processing approach for metagenomic taxonomic binning based on deep learning. *Current Bioinformatics* 2021;**16**:992–1003. https://doi.org/10.2174/1574893616666210621101150
147. Gregor I, Dröge J, Schirmer M. *et al.* PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* 2016;**4**:e1603. https://doi.org/10.7717/peerj.1603
148. Dröge J, Gregor I, McHardy AC. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* 2015;**31**:817–24. https://doi.org/10.1093/bioinformatics/btu745
149. Sczyrba A, Hofmann P, Belmann P. *et al.* Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat Methods* 2017;**14**:1063–71. https://doi.org/10.1038/nmeth.4458
150. Maimon O, Rokach L. *Data Mining and Knowledge Discovery Handbook*, Vol. **2**. New York: Springer; 2005.
151. Tan R, Yu A, Liu Z. *et al.* Prediction of minimal inhibitory concentration of meropenem against Klebsiella pneumoniae using metagenomic data. *Front Microbiol* 2021;**12**:712886. https://doi.org/10.3389/fmicb.2021.712886

152. Nguyen M, Brettin T, Long SW. *et al.* Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci Rep* 2018;**8**:421. https://doi.org/10.1038/s41598-017-18972-w
153. Karagöz MA, Nalbantoglu OU. Taxonomic classification of metagenomic sequences from relative abundance index profiles using deep learning. *Biomed Signal Process Control* 2021;**67**:102539. https://doi.org/10.1016/j.bspc.2021.102539
154. Nalbantoglu OU, Way SF, Hinrichs SH. *et al.* RaiPHY: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinform* 2011;**12**:1–14.
155. Fang Z, Zhou H. VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. *Front Microbiol* 2021;**12**:615711.
156. Seguritan V, Alves NJr, Arnoult M. *et al.* Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput Biol* 2012;**8**:e1002657. https://doi.org/10.1371/journal.pcbi.1002657
157. Ding H, Feng P-M, Chen W. *et al.* Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol Biosyst* 2014;**10**:2229–35. https://doi.org/10.1039/C4MB00316K
158. Manavalan B, Shin TH, Lee G. PVP-SCM: sequence-based prediction of phage virion proteins using a support vector machine. *Front Microbiol* 2018;**9**:476. https://doi.org/10.3389/fmicb.2018.00476
159. Charoenkwan P, Kanthawong S, Schaduangrat N. *et al.* PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method. *Cells* 2020;**9**:353. https://doi.org/10.3390/cells9020353
160. Charoenkwan P, Nantasenamat C, Hasan MM. *et al.* Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *J Comput Aided Mol Des* 2020;**34**:1105–16. https://doi.org/10.1007/s10822-020-00323-z
161. Norman JM, Handley SA, Baldridge MT. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 2015;**160**:447–60. https://doi.org/10.1016/j.cell.2015.01.002
162. Yan Miao F, Liu TH, Liu Y. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* 2022;**38**:1216–22. https://doi.org/10.1093/bioinformatics/btab845
163. Mock F, Kretschmer F, Kriese A. *et al.* Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proc Natl Acad Sci* 2022;**119**:e2122636119. https://doi.org/10.1073/pnas.2122636119
164. Titus Brown C, Irber L. Sourmash: a library for minhash sketching of DNA. *J Open Source Softw* 2016;**1**:27. https://doi.org/10.21105/joss.00027
165. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
166. Bao HQ, Vinh LV, Van Hoai T. A deep embedded clustering algorithm for the binning of metagenomic sequences. *IEEE Access* 2022 Cited By:2; **10**:54348–57. https://doi.org/10.1109/ACCESS.2022.3176954
167. Van Vinh L, Van Lang T, Binh LT. *et al.* A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms Mol Biol* 2015;**10**:1–12. https://doi.org/10.1186/s13015-014-0030-4
168. Tyson GW, Chapman J, Hugenholtz P. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;**428**:37–43. https://doi.org/10.1038/nature02340
169. Yang B, Yu P, Leung HCM. *et al.* MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. In: *Proceedings of the first ACM international conference on bioinformatics and computational biology*. New York: Association for Computing Machinery; 2010. pp. 170–9.
170. Wang Y, Leung HCM, Yiu S-M. *et al.* MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* 2012;**28**:i356–62. https://doi.org/10.1093/bioinformatics/bts397
171. Líndez PP, Johansen J, Kutuzova S. *et al.* Adversarial and variational autoencoders improve metagenomic binning. *Commun Biol* 2023;**6**:1073. https://doi.org/10.1038/s42003-023-05452-3
172. Lloyd-Price J, Arze C, Ananthakrishnan AN. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62. https://doi.org/10.1038/s41586-019-1237-9
173. Nissen JN, Johansen J, Allesœ RL. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;**39**:555–60. https://doi.org/10.1038/s41587-020-00777-4
174. Pan S, Zhu C, Zhao X-M. *et al.* A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat Commun* 2022;**13**:2326. https://doi.org/10.1038/s41467-022-29843-y
175. Liu C-C, Dong S-S, Chen J-B. *et al.* MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 2022;**10**:1–16.
176. Madival SD, Mishra DC, Sharma A. *et al.* A deep clustering-based novel approach for binning of metagenomics data. *Curr Genomics* 2022;**23**:353–68. https://doi.org/10.2174/1389202923666220928150100
177. Sharon I, Morowitz MJ, Thomas BC. *et al.* Time series communitygenomics analysis reveals rapid shifts. *Genome Res* 2013;**23**:111–20.
178. Herath D, Tang S-L, Tandon K. *et al.* Comet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinform* 2017;**18**:161–72.
179. Sinha D, Sharma A, Mishra DC. *et al.* Metaconclust-unsupervised binning of metagenomics data using consensus clustering. *Curr Genomics* 2022;**23**:137. https://doi.org/10.2174/1389202923666220413114659
180. Kang DD, Froula J, Egan R. *et al.* Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015;**3**:e1165. https://doi.org/10.7717/peerj.1165
181. Wu Y-W, Tang Y-H, Tringe SG. *et al.* MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2014;**2**:1–18.
182. Lamurias A, Sereika M, Albertsen M. *et al.* Metagenomic binning with assembly graph embeddings. *Bioinformatics* 2022 Cited By:2; **38**:4481–7. https://doi.org/10.1093/bioinformatics/btac557
183. Wick RR. Badread: simulation of error-prone long reads. *J Open Source Softw* 2019;**4**:1316. https://doi.org/10.21105/joss.01316
184. Pan S, Zhu C, Zhao X-M. *et al.* SemiBin: incorporating information from reference genomes with semi-supervised deep learning leads to better metagenomic assembled genomes (MAGs). *Nat Commun* 2022;**13**:2326. https://doi.org/10.1038/s41467-022-29843-y

185. Mallawaarachchi V, Wickramarachchi A, Yu L. GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* 2020;**36**:3307–13. https://doi.org/10.1093/bioinformatics/btaa180
186. Wickramarachchi A, Lin Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol Biol* 2022;**17**:14. https://doi.org/10.1186/s13015-022-00221-z
187. Stöcker BK, Köster J, Rahmann S. SimLoRD: simulation of long read data. *Bioinformatics* 2016;**32**:2704–6. https://doi.org/10.1093/bioinformatics/btw286
188. Wickramarachchi A, Mallawaarachchi V, Rajan V. *et al.* MetaBCC-LR: meta genomics binning by coverage and composition for long reads. *Bioinformatics* 2020;**36**:i3–11. https://doi.org/10.1093/bioinformatics/btaa441
189. Arias PM, Hill KA, Kari L. i deLUCS: a deep learning interactive tool for alignment-free clustering of DNA sequences. *Bioinformatics* 2023;**39**:btad508. https://doi.org/10.1093/bioinformatics/btad508
190. Arias PM, Alipour F, Hill KA. *et al.* DeLUCS: deep learning for unsupervised clustering of DNA sequences. *PloS One* 2022;**17**:e0261531. https://doi.org/10.1371/journal.pone.0261531
191. Girgis HZ. MeShClust v3. 0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *BMC Genomics* 2022;**23**:423. https://doi.org/10.1186/s12864-022-08619-0
192. Arisdakessian CG, Nigro OD, Steward GF. *et al.* CoCoNet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics* 2021 Cited By:5; **37**:2803–10. https://doi.org/10.1093/bioinformatics/btab213
193. Wang Z, Wang Z, Lu YY. *et al.* SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* 2019;**35**:4229–38. https://doi.org/10.1093/bioinformatics/btz253
194. Yang Young L, Chen T, Fuhrman JA. *et al.* COCACOLA: binning metagenomic contigs using sequence composition, read coverage, co-alignment and paired-end read linkage. *Bioinformatics* 2017;**33**:791–8. https://doi.org/10.1093/bioinformatics/btw290
195. Huang W, Li L, Myers JR. *et al.* ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;**28**:593–4. https://doi.org/10.1093/bioinformatics/btr708
196. Jang HB, Bolduc B, Zablocki O. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 2019;**37**:632–9. https://doi.org/10.1038/s41587-019-0100-8
197. Kristensen DM, Waller AS, Yamada T. *et al.* Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J Bacteriol* 2013;**195**:941–50. https://doi.org/10.1128/JB.01801-12
198. Chibani CM, Farr A, Klama S. *et al.* Classifying the unclassified: a phage classification method. *Viruses* 2019;**11**:195. https://doi.org/10.3390/v11020195
199. Berahmand K, Daneshfar F, Salehi ES. *et al.* Autoencoders and their applications in machine learning: a survey. *Artif Intell Rev* 2024;**57**:28. https://doi.org/10.1007/s10462-023-10662-6
200. Singleton CM, Petriglieri F, Kristensen JM. *et al.* Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun* 2021;**12**:2009. https://doi.org/10.1038/s41467-021-22203-2
201. Brunbjerg AK, Bruun HH, Brøndum L. *et al.* A systematic survey of regional multi-taxon biodiversity: evaluating strategies and coverage. *BMC Ecol* 2019;**19**:1–15.
202. Beaulaurier J, Luo E, Eppley JM. *et al.* Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res* 2020;**30**:437–46. https://doi.org/10.1101/gr.251686.119
203. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci* 2008;**105**:1118–23. https://doi.org/10.1073/pnas.0706851105
204. Wirbel J, Pyl PT, Kartal E. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 2019;**25**:679–89. https://doi.org/10.1038/s41591-019-0406-6
205. Coelho LP, Kultima JR, Costea PI. *et al.* Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* 2018;**6**:1–11.
206. Sunagawa S, Coelho LP, Chaffron S. *et al.* Structure and function of the global ocean microbiome. *Science* 2015;**348**:1261359.
207. Olm MR, Butterfield CN, Alex Copeland T. *et al.* The source and evolutionary history of a microbial contaminant identified through soil metagenomic analysis. *MBio* 2017;**8**:10–1128.
208. Louis S, Tappu R-M, Damms-Machado A. *et al.* Characterization of the gut microbial community of obese patients following a weight-loss intervention using whole metagenome shotgun sequencing. *PloS One* 2016;**11**:e0149564. https://doi.org/10.1371/journal.pone.0149564
209. Song K, Ren J, Sun F. Reads binning improves alignment-free metagenome comparison. *Front Genet* 2019;**10**:1156.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf065
https://doi.org/10.1093/bib/bbaf065
Review