Received: 31 May 2023 | Revised: 3 August 2023 | Accepted: 10 August 2023

DOI: 10.1002/pmic.202200421

Proteomics
Proteomics and Systems Biology

REVIEW

Small proteins in bacteria – Big challenges in prediction and identification

Stephan Fuchs[1] | **Susanne Engelmann**[2,3]

[1]Genome Competence Center (MF1), Department MFI, Robert-Koch-Institut, Berlin, Germany

[2]Institute for Microbiology, Technische Universität Braunschweig, Braunschweig, Germany

[3]Microbial Proteomics, Helmholtzzentrum für Infektionsforschung GmbH, Braunschweig, Germany

Correspondence
Susanne Engelmann, Microbial Proteomics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany.
Email: Susanne.Engelmann@helmholtz-hzi.de

Funding information
DFG priority program, Grant/Award Numbers: SPP2002, EN 712/4-1; Institutional Fund (DFG), Grant/Award Number: INST 188/365-1 FUGG DFG; GRK PROCOMPAS (DFG)

Abstract

Proteins with up to 100 amino acids have been largely overlooked due to the challenges associated with predicting and identifying them using traditional methods. Recent advances in bioinformatics and machine learning, DNA sequencing, RNA and Ribo-seq technologies, and mass spectrometry (MS) have greatly facilitated the detection and characterisation of these elusive proteins in recent years. This has revealed their crucial role in various cellular processes including regulation, signalling and transport, as toxins and as folding helpers for protein complexes. Consequently, the systematic identification and characterisation of these proteins in bacteria have emerged as a prominent field of interest within the microbial research community. This review provides an overview of different strategies for predicting and identifying these proteins on a large scale, leveraging the power of these advanced technologies. Furthermore, the review offers insights into the future developments that may be expected in this field.

KEYWORDS
bioinformatics, bottom-up proteomics, databases, mass spectrometry, protein identification, top-down proteomics, proteogenomics

1 | INTRODUCTION

The last few decades have seen remarkable progress in the understanding of bacterial proteomes. The discovery of numerous small bacterial mRNAs and the ongoing identification of proteins of up to 100 amino acids (aa) have revealed a level of complexity in bacterial proteomes that far exceeds previous expectations. The term 'small protein' is not clearly defined and is used in different studies for proteins with sometimes very different length restrictions (25, 50, 70 or 100 aa) [1–9]. In the following, proteins with a length of up to 100 aa are referred to as small. There is increasing evidence that these proteins play essential roles in a wide range of cellular processes including cell signalling or regulation, toxins/anti-toxin systems, membrane functions, protein folding, and the formation and stabilisation of protein complexes [10–13]. Using sophisticated bioinformatics, sequencing and proteomics tools to identify the entire microproteome in bacteria is therefore a worthwhile, albeit challenging, approach [1, 14].

Automatic annotation and prediction of open reading frames (ORFs) encoding proteins of less than 100 aa is difficult for several reasons. These include insufficient sequence information for domain and homology searches, a limited number of experimentally validated ORFs, and the tendency of these proteins to be species-specific [15]. This makes it extremely difficult to distinguish between small open reading frames (sORFs) with low and high coding potential, and the

Abbreviations: aa, amino acid(s); bp, base pair(s); CDSs, coding sequences; CNNs, convolutional neural networks; DAA, data dependent acquisition mode; dN/dS, nonsynonymous to synonymous substitutions; HMMs, Hidden Markov models; iPtgxDB, integrated proteogenomics database; ML, machine learning; nt, nucleotides; ONT, Oxford Nanopore Technologies; PSMs, peptide spectrum matches; RBS, ribosomal binding site; RF, random forest; RNNs, recurrent neural networks; sORF, small open reading frame; sORFs, small open reading frames; TIS, translation initiation site; TTS, translation termination site.

number of false positives among predicted sORFs is very high [2, 16, 17]. Since start and stop codons are usually AT-rich, predicting protein-coding sORFs in GC-rich bacteria is even more challenging [2]. Given these facts, arbitrary cut-offs for minimal ORF lengths ranging from 50 to 100 codons were routinely applied to annotate bacterial genomes [18]. In addition, sequence characteristics, such as ribosomal binding sites (RBSs), codon usage and aa conservation are increasingly being used to distinguish between coding and non-coding ORFs [15]. This significantly reduces the number of false positives among the annotated sORFs in the databases, but also leads to the exclusion of bona fide small protein-coding genes. Therefore, in order to obtain more qualified protein databases for bacteria of interest, with a particular focus on small proteins, more accurate sORF prediction algorithms need to be established.

In addition, innovative developments in the direct detection of translated ORFs and proteins are having a major impact on small protein research. Ribosome profiling and mass spectrometry (MS)-based proteomics are currently the methods of choice for the experimental identification of protein-coding ORFs on a global scale. Ribosome profiling is much more sensitive, but provides only indirect evidence for the presence or absence of a small protein [19]. MS-based protein detection is currently the best method for not only directly detecting small proteins that accumulate at meaningful levels in the cell, but also for providing information about existing proteoforms and post-translational modifications [20, 21]. In addition, proteomics can also provide clues to the global subcellular localisation of proteins and protein–protein interactions, as well as insights into the molecular structure of protein complexes. The classical bottom-up proteomics approach, based on LC-MS/MS analysis of highly complex peptide mixtures derived from complex protein crude extracts after proteolytic digestion, provides deep insights into the bacterial proteome [22–24]. How complete the experimentally determined proteome of a bacterium ultimately depends not only on the quality of the MS/MS data but also on the quality of the protein databases used for the analysis of the MS/MS data. In contrast to genomic and transcriptomic technologies, where the DNA or RNA fragments are actually sequenced, MS-based proteomics mainly identifies peptides by matching MS/MS spectra against theoretical spectra of all candidate peptides present in a reference protein database (peptide spectrum matches [PSMs]) [25]. Classical bottom-up proteomics therefore only supports the identification of proteins that are expected to be produced by a given organism and is clearly biased towards the study of proteins with more than 100 aa. In addition, the low molecular weight of small proteins complicates experimental preparation and reduces the number of peptides that can be detected by MS. Therefore, the systematic MS-based identification of small proteins has been an obstacle for a long time. The increasing realisation that small proteins exist and play an essential role in many cellular processes [8, 10–13, 26] has led to several activities in recent years to improve MS-workflows and protein databases that support the identification of small proteins in bacteria [1].

In this review, we aim to highlight the challenges, advances and future prospects in the prediction of short protein-coding sORFs and the MS-based identification of the resulting small proteins in bacteria.

2 | IN SILICO PREDICTION OF SMALL OPEN READING FRAMES

The computational prediction of sORFs in bacteria has demonstrated immense potential in uncovering the existence and functions of small proteins. Additionally, it serves as a prerequisite for different experimental detection methods, such as classical shotgun proteomics. Bacteria exhibit distinct characteristics in genomic organisation and mRNA translation compared to eukaryotes, necessitating the adaptation of gene-finding algorithms tailored to these unique features. For instance, bacterial genomes often contain a higher density of coding sequences (CDSs) and lack introns, unlike eukaryotic genomes. Additionally, bacterial mRNA translation may commence with different start codons, and the operon structure in prokaryotic genomes can lead to polycistronic mRNAs. The heterogeneity in genomic characteristics intrinsic to different bacterial species can significantly affect the performance and effectiveness of predictive algorithms. For example, the task of detecting genes in GC-rich genomes becomes significantly more challenging due to the increased likelihood of encountering random ORFs [2]. Metrics, such as the codon adaptation index, lose their robustness when dealing with more recent horizontal gene transfers [27]. Unlike their eukaryotic counterparts, bacterial transcripts are typically polycistronic and generally lacking splicing and polyadenylation signals, and do not always have a clear translation initiation site (TIS) or RBS. Other gene-specific properties such as size, start codon and nested localisation have also been shown to significantly influence the accuracy of gene prediction [9, 14, 28–30]. Small ORFs, in particular, have divergent features, such as atypical nucleotide composition, lack of RBS and non-canonical start codons, which contribute to the computational challenges of prediction [17, 31, 32].

In the following sections, we will discuss different in silico methods for predicting ORFs from genome sequences, highlighting their advantages and remaining limitations, particularly with regard to the detection of sORFs (Table 1). It is vital to understand that prediction (and detection in this context) is not synonymous with identification. The predictive methods and models are inherently linked to presumptions and probabilities and may be influenced by factors such as false discovery rate (FDR). This complex nature of prediction underscores the urgent need for validation, as it is essential to verify the authenticity of the predicted ORFs, especially those pertaining to sORFs. The focus on this validation, and its critical importance, will be further explored in Section 5.

Our review provides only an overview of the different bioinformatics strategies and tools currently available, outlining their key strengths and limitations. However, we do not engage in comparative benchmarking of these tools. For those readers looking for a more detailed analysis and performance comparison of the tools mentioned, we would like to refer to the excellent benchmark studies by Dimonaco et al., Gelhausen et al. and Korandla et al. [14, 50, 51].

TABLE 1 Gene prediction programs mentioned in this review.

Tool	Approach	(Primary) input	(Main) method	Availability	Reference
MetaProdigal[a]	Ab initio	Genome sequences	Dynamic programming and comparative genomics[b]	Bioconda: https://anaconda.org/bioconda/prodigal Github: https://github.com/hyattpd/Prodigal	[33]
Prodigal	Ab initio	Genome sequences	Dynamic programming and Comparative genomics[b]	Bioconda: https://anaconda.org/bioconda/prodigal Github: https://github.com/hyattpd/Prodigal	[34]
RNACode	Ab initio	RNA Sequence Alignment	Comparative genomics[b]	Bioconda: https://anaconda.org/bioconda/rnacode Github: https://github.com/ViennaRNA/RNAcode	[35]
sORF finder[c]	Ab initio	Genome sequences	Comparative genomics[b]	Download: http://labo.bio.kyutech.ac.jp/~kohanada/sORFfinder2.tar.gz	[36]
μProteIns[c]	Evidence-based	RNA-seq and MS data	Reference-guided assembly and peptide mapping	Github: https://github.com/Eduardo-vsouza/uproteins	[37]
DeepRibo	Evidence-based	Ribo-seq	Deep learning	Github: https://github.com/Biobix/DeepRibo	[38]
Pepper[c]	Evidence-based	Peptide identifications and MS data	Peptide mapping	Gitlab: https://gitlab.com/s.fuchs/pepper	
REPARATION	Evidence-based	Ribo-seq	Random forest	Bioconda: https://anaconda.org/bioconda/reparation_blast[d] Github: https://github.com/Biobix/REPARATION	[39]
smORFer[c]	Evidence-based	Ribo-seq	Fourier transform	Github: https://github.com/Alexander Bartholomaeus/smORFer	[40]
CRITICA	Hybrid	Genome sequences	Comparative genomics[b]	Download: http://www.ttaxus.com/software.html	[41]
OCCAM[c]	Hybrid	Genome sequences	Comparative genomics[b]	Download: http://www.labinfo.lncc.br/occam	[42]
SearchDOGS Bacteria	Hybrid	Genome sequences	Comparative genomics[b]	Download: http://wolfe.ucd.ie	[43]
AUGUSTUS	ML-based	Genome sequences	Ribo-seq supported Hidden Markov models	Bioconda: https://anaconda.org/bioconda/augustus Github: https://github.com/Gaius-Augustus/Augustus Web service: http://bioinf.uni-greifswald.de/webaugustus	[44]
Balrog	ML-based	Genome sequences	Deep learning	Bioconda: https://anaconda.org/bioconda/balrog Github: https://github.com/salzberg-lab/Balrog	[45]
GeneMarkS	ML-based	Genome sequences	Hidden Markov Models	Web service: http://exon.gatech.edu/genemark/genemarks.cgi	[46]
Glimmer	ML-based	Genome sequences	Interpolated Markov models	Bioconda: https://anaconda.org/bioconda/glimmer Download: http://ccb.jhu.edu/software/glimmer/index.shtml	[47]
RanSEP[c]	ML-based	Genome sequences	Random Forest Classifier	Github: https://github.com/samuelmiver/RanSEPs Download: http://ranseps.crg.es	[4]
BLAST+[e]	Similarity-based	Genome sequences	Comparative genomics[b]	Bioconda: https://anaconda.org/bioconda/blast Download: https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html Web service: https://blast.ncbi.nlm.nih.gov/Blast.cgi	[48]
FASTA3[e]	Similarity-based	Genome sequences	Comparative genomics[b]	Bioconda: https://anaconda.org/bioconda/fasta3 Web service: https://www.ebi.ac.uk/Tools/sss/fasta	[49]

[a] Shows improved accuracy for small open reading frame (sORF) detection compared to prodigal.

[b] Comparative genomics-based metrics such as sequence similarities, codon substitution rates and/or nucleotide compositions.

[c] Specifically designed for sORF detection.

[d] In this version the commercial 'usearch' program has been substituted with 'blast'.

[e] Although these programs are not gene prediction programs in the strict sense (as discussed in Section 2.1), their inclusion here is for completeness.

2.1 | Similarity-based prediction methods

Phylogenetic conservation is intrinsically linked to biological information, such as CDSs, and conserved sequences across species often indicate important functions. This evolutionary constraint enables the identification and understanding of functionally important genes and their encoded proteins. Similarity-based approaches exploit evolutionary conservation and homology between known and unknown proteins by searching for sequence similarities between query sequences and annotated protein databases. This allows the identification of putative small proteins based on their resemblance to known proteins.

Both FASTA and BLAST mark important milestones in computer-assisted homology searches with a large number of target sequences [52, 53]. These tools have significantly improved the speed and efficiency of sequence comparison, enabling researchers to search increasingly large sequence databases. Despite their utility, FASTA and BLAST are primarily sequence alignment tools and not stand-alone gene prediction tools.

In a unique study, researchers relied solely on a similarity-based approach using mpiBLAST [54], a high-performance variant of the original BLAST algorithm, to compare intergenic sequences from 1474 fully assembled replicon sequences of 780 distinct prokaryotic genomes [31]. The database used for this comparison included not only annotated CDSs from all genomes examined, but also all potential intergenic ORFs of at least 99 bp. This approach led to the detection of 38,895 unannotated genes across all genomes and 1100 ‘missing’ genes that did not align with any known sequences, suggesting they may belong to so far undiscovered gene families. The vast majority of these unannotated and missed genes were short, encoding no more than 100 aa, consistent with the arbitrary length cut-offs of conventional gene annotation pipelines. In addition, a significant proportion of these proteins appear to be of foreign origin, lacking the usual sequence features shared by other protein-coding genes within the same organism. This absence of familiar traits could have contributed to their initial evasion from detection [31]. However, this study only considered potential ORFs with canonical start codons, excluding a significant number of non-canonical sequences. It is therefore reasonable to assume that the actual number of small, undiscovered genes is likely to be much higher.

While similarity-based approaches can reveal structural properties, conserved motifs and functional roles, they have limitations when analysing shorter sequences. Shorter sequences increase the likelihood of random matches, leading to false positives and difficulties identifying true homologous relationships. As a result, the statistical power of comparisons obtained when searching with short query sequences is reduced [55]. Tools such as BLAST rely on the calculation of statistical significance (e.g., *E*-value) to assess the reliability of the matches found. Shorter sequences typically yield lower alignment scores, resulting in higher *E*-values and lower statistical significance, making it more challenging to distinguish biologically meaningful matches from random ones. Additionally, short genes are less conserved and often species-specific, complicating the application of this approach in non-model organisms. Short protein sequences may also lack conserved domains or functional motifs, making it difficult for similarity searches to detect distant homologues or identify functionally related proteins with limited sequence identity. In addition, sensitivity is generally lower when searching for shorter sequences, particularly DNA-to-DNA comparisons, which have 5–10 times lower sensitivity than searches using translated sequences [55]. Therefore, these methods are mainly used in combination with other approaches, such as ab initio methods, improve gene predictions [36] (see Section 2.4).

2.2 | Ab initio prediction methods

Ab initio methods rely on intrinsic sequence properties, such as codon usage, ribosome binding site motifs and secondary structure predictions, to identify genes. They use algorithms based on predetermined rules and statistical models to recognise patterns within the genomic sequence that indicate the presence of genes. Unlike similarity-based approaches, ab initio methods do not require prior knowledge of existing proteins, making them particularly valuable for the discovery of novel proteins in non-model bacterial species. Numerous tools, such as GLIMMER and AUGUSTUS (both of which utilise Markov models, which is why they can also be classified as ML-based methods; see next section), have gained prominence by adopting this strategy (Table 1) [47, 44]. However, the selection of specialised programs for prokaryotic gene prediction is limited, and becomes even more constrained when considering the unique requirements for sORF detection.

Prodigal is a widely used ab initio gene-finding algorithm specifically designed for prokaryotic genomes. It employs a dynamic programming approach to predict coding regions based on input sequence features [34]. The algorithm evaluates gene coding potential by combining coding scores derived from in-frame hexamer statistics and start scores based on start codon and RBS motif frequencies. However, Prodigal is not tailored for sORF prediction. To control FDRs, it favours longer genes through specific rules, such as penalising final scores for genes shorter than 250 bp and excluding genes no longer than 90 bp.

RNAcode is a computational tool that specialises in predicting conserved coding regions within RNA sequences by utilising multiple sequence alignment of homologous RNA sequences [35]. It assesses both synonymous and non-synonymous substitution rates (see also Section 2.3), operating on the underlying principle that coding regions are more inclined to display specific evolutionary patterns compared to non-coding regions. As emphasised by its authors, RNAcode can achieve satisfactory results using alignments of just four sequences that are less than 90% identical [35]. Thus, RNAcode enables the de novo prediction of unknown CDSs not only in model but also in non-model organisms, as well as microbial communities (see also Section 5). Moreover, RNAcode has been successfully applied in identifying previously undiscovered sORFs, while also shedding light on their functional roles [56, 57].

Other mainly ab initio tools that are more specialised in the detection of small proteins include sORF finder [36] and MetaProdigal [33]. sORF finder predicts sORFs by analysing the nucleotide pentamer and hexamer composition bias between coding and non-CDSs. To make accurate predictions, the method requires a significant amount of known coding and non-CDSs from the organism under investigation. Bayes' estimation of coding probability enhances the accuracy of gene prediction, while optional homology searches based on BLAST can provide additional support in assessing the coding potential of identified gene candidates. Estimating synonymous and non-synonymous substitutions in homologous sequences, followed by a chi-square test, further supports coding potential evaluation.

MetaProdigal, an extension of Prodigal, is tailored for metagenomic data and can predict small proteins with improved accuracy using specific parameters and training data [56, 58]. A major advantage is the ability to apply training data from different species to evaluate gene candidates. To save computational resources, the selection of training data is based on the GC content of the input sequence, considering the problem of multiple testing. The computation of confidence values for each gene candidate, representing the logarithm of the probability that a gene is genuine compared to the background, further improves the result evaluation.

While ab initio methods provide valuable insights, they also have several limitations and challenges. Because they are based on predefined rules and models, these methods may have limited adaptability and generalisability to new or different genomic contexts, as their performance depends on the accuracy of the underlying assumptions. The complexity of those can make the quality of outcomes unpredictable. For example, Dimonaco et al. recently reported that Augustus (version 3.3.3) underperformed when analysing *Staphylococcus aureus* data using the *S. aureus* model, achieving a detection rate of only a 21% [14]. Surprisingly, the performance improved significantly when using a *Homo sapiens* model, reaching a detection rate of 79%. On the other hand, Augustus detected 96.64% of *Pseudomonas fluorescens* genes when using the *Escherichia coli* model, indicating that genes from both organisms share common features and characteristics captured by the model [14]. These findings emphasise the potential variability in the performance of ab initio methods, depending on the selected model. Compared to similarity-based methods, ab initio approaches often provide only limited information about function or biological roles of detected genes. Analysing the distinctive architecture of gene clusters in prokaryotes can complement this, particularly for sORFs, where homology-based information is mostly scarce. By integrating information of unique gene arrangements with associated molecular processes, a comprehensive understanding of specialised metabolic and cellular pathways can be achieved which can also help to shed light on the physiological roles of co-localised small proteins [59].

2.3 | Machine learning (ML)-based methods

ML-based approaches have become increasingly popular in gene prediction because they provide a data-driven approach to identifying novel protein-CDSs (Table 1). These methods use computational models to learn and recognise complex patterns and features in genomic data, improving the prediction of gene candidates. Unlike similarity-based methods, ML methods use algorithms to learn and generate models from training data, which typically consists of annotated genes and non-coding regions. These models can capture complex relationships in the data, potentially allowing better generalisation to different genomic contexts.

GeneMarkS [46] is an ML-based gene prediction tool that combines hidden Markov models (HMMs, see Info Box 1) for protein-coding and non-coding regions with models of regulatory sites near gene starts. It learns species-specific parameters from prokaryotic input sequences without prior knowledge of any protein or rRNA genes. In tests with the genome sequence of *Bacillus subtilis*, GeneMarkS demonstrated similar accuracy in detecting gene starts for both genes shorter than 300 nt and long genes. As expected, the fraction of accurately detected genes shorter than 300 bp improved to up to 90% as more known genes shared significant sequence similarity [46]. In the same study, Glimmer showed reduced performance, only accurately detecting up to 72% of short genes. This difference may due to Glimmer's default setting, which predicts the gene start at the start codon of the longest ORF containing the predicted gene.

RanSEP is specifically designed for identifying sORFs in bacterial genomes [4]. Utilising random forest (RF) classifiers, it distinguishes between coding and non-coding sORFs by computing various features, such as aa composition, hydrophobicity and secondary structure. The method also incorporates additional features like start codon prevalence, GC content and RBS information to enhance prediction accuracy and adapt to different organisms by learning species-specific parameters from provided training data. Feature selection is carried out to minimise overfitting, while an out-of-bag approach is implemented for feature importance estimation, offering insights into the significance of each feature in the classification process, enabling users to thoroughly analyse its performance and predictions. RanSEP has been successfully applied to screen systematically for sORFs in various bacteria [4].

Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have also been utilised successfully for sORF detection (see Info Box 1). These methods are capable of capturing intricate sequence features and context dependencies by processing genomic data through multiple layers of interconnected neurons. One example is DeepRibo [38], which is built upon an artificial neural network that employs both CNN and RNN architectures. This neural network integrates recurrent memory cells and convolutional layers, merging the information obtained from high-throughput ribosome profiling data and ribosome binding translation initiation sequence regions into a single model. DeepRibo is specifically designed as a unified model, trained on different ribosome profiling experiments, allowing the identification of ORFs in prokaryotes without prior knowledge of the translational landscape. In a recent evaluation of ORF prediction tools based on Ribo-seq, DeepRibo demonstrated exceptional performance across a wide range of bacteria, distinguishing itself as a robust and reliable choice [50] (see also Section 3.2).

Recently, Balrog has been introduced as gene-finding algorithm across different prokaryotic species [45]. It utilises a universal model of prokaryotic genes based on a temporal convolutional network. While Balrog aims to detect smaller genes with its default minimum ORF

Info Box 1

Hidden Markov models (HMMs)

HMMs describe any DNA or protein sequence as a series of observable symbols (nucleotides or amino acids [aas]) generated by hidden states. These hidden states represent different features in the DNA sequence, such as coding regions, non-coding regions or regulatory elements, each with its own probabilities of 'emitting' specific nucleotides or aas. These emission probabilities represent the likelihood of observing each nucleotide when the DNA sequence is in a particular state. For example, a coding region might have a higher probability of emitting certain nucleotide patterns than a non-coding region. Transition probabilities between the hidden states represent the likelihood of transitioning from one state to another as we move along the DNA sequence. These probabilities help capture the structure and organisation of the DNA sequence, such as the tendency for coding regions to be followed by non-coding regions, and vice versa. Once the HMM is defined, it can be used to determine the likely series of hidden states behind a DNA sequence. This allows to identify key features, like coding and non-coding regions, as well as the location of regulatory elements.

Random forests

A random forest is a machine learning (ML) technique particularly useful for classification tasks. It consists of multiple decision trees that helps classify an input, like a DNA sequence, based on certain features, such as nucleotide composition and predicted secondary structure. Each tree in the forest is built using a random portion of the available data, which helps make the overall model diverse and less prone to errors. When a new input, like a DNA sequence, needs to be classified, it is passed through each tree in the forest. Each tree makes a decision about the classification based on the input's features. To get the final prediction, the decisions of all trees are combined, usually by taking the most common decision.

Deep learning

Deep learning models are a subset of ML and are designed to automatically learn intricate patterns and features from complex data. Deep learning models such as convolutional neural networks (CNNs) process DNA sequences, learning to detect specific combinations of nucleotides, known as motifs, which are often indicative of genes or other significant features by applying a series of filters across the sequence. They can analyse large-scale DNA sequences and automatically identify relevant features, enhancing the efficiency of gene prediction. Additionally, these models consider the sequential nature of DNA sequences. Some, like recurrent neural networks (RNNs) and derivates, even retain information from previous positions in the sequence, enabling them to capture dependencies and patterns across the entire sequence. This ability to consider sequence context can be crucial for gene prediction, where the order and combination of nucleotides can determine gene functionality and location.

Rule-based expert systems

Rule-based expert systems rely on a predefined set of rules or heuristics to make decisions or solve problems. These systems are particularly useful in domains where expert knowledge can be formalised into clear, logical rules. In the context of biology, these rules could include criteria for identifying coding regions, non-coding regions, regulatory elements or other functional elements within the genome. The rule-based expert system then processes the input DNA sequence by applying the predefined rules, determining which rules apply to each part of the sequence, and ultimately making predictions or classifications based on the combined results. Compared to the ML-based method, rule-based expert systems impress with transparency and interpretability. Since the rules are explicitly defined by experts, it is easier to understand and explain the reasoning behind the system's predictions.

length of 60 nucleotides, the primary focus is not on accurate start site prediction. This limitation can lead to challenges in precisely identifying the boundaries of small genes, which may affect the subsequent analysis and functional characterisation of these small genes. In addition, Balrog may be less effective at detecting functionally uncharacterised gene families because hypothetical proteins were excluded from its training data.

ML-based methods are excellent at identifying complex patterns and relationships in genomic data. Recent studies have shown that these approaches can match or even outperform ab initio tools [14]. However, their effectiveness is strongly linked to the quality and representativeness of the training data and the selected features used. If the model has limited similarity to the genome or inadequately covers the intra-species genomic variation, prediction accuracy drops significantly [14]. Similarly, the selection of characteristics must be applied carefully and cannot be generalised. This highlights the critical role of high-quality, labelled training data from closely related genomes in building accurate prediction models. For example, the ratio of non-synonymous to synonymous substitutions (dN/dS) is a commonly used feature whose importance in discriminating between coding and non-CDSs has been widely demonstrated. This makes sense, because CDSs have lower ratios than non-CDSs because they have undergone purifying selection during evolution. A commonly used implementation of the dN/dS test uses PAML as the scoring matrix, which is based on substitution rates observed in standard-length ORFs, which in turn are not necessarily similar to those observed in shorter sequences, as shown for short exon sequences [60, 61].

2.4 | Hybrid methods

Hybrid methods exploit the strengths of ab initio, similarity-based and ML approaches to improve prediction accuracy and minimise false positives in the detection of small proteins without relying exclusively on prior gene annotation or homology-based comparisons with known proteins. These methods typically incorporate multiple algorithms and data sources, such as sequence features, conservation patterns and structural properties, to improve gene prediction. More recently, additional hybrid tools specifically designed for the detection of sORFs have been introduced (Table 1).

Tools such as COMBREX, CRITICA and SearchDOGs combine homology searches with statistical measures or synteny analysis [41]. All tools were successfully applied to predict bacterial sORFs [43, 62–65]. However, they are limited to comparing closely related organisms and cannot predict novel sORFs that do not share homology with an annotated gene. On the other hand, OCCAM [42] follows mainly the similarity-based workflow as proposed by Warren and colleagues but considers sORFs between 36 and 300 bp [31]. To increase sensitivity without compromising specificity, false positive BLAST hits are filtered using an ML-based target-decoy approach.

Evidence-based hybrid methods share the ability to integrate experimental information, such as ribosome profiling (Ribo-seq) and proteomics data, providing empirical evidence of active translation and improving the overall accuracy of small protein identification. By exploiting the complementary strengths of different approaches and data types, these methods can offer more comprehensive and reliable means of predicting small proteins within genomic data, and are discussed in the next chapter.

3 | EVIDENCE-BASED STRATEGIES OF PREDICTION OF SMALL OPEN READING FRAMES

In addition to bioinformatic algorithms, information from experimental approaches such as RNA-seq, Ribo-seq or MS data is increasingly being used for genome annotation and, in particular, to support in silico sORF prediction [2, 7, 9, 30, 56, 66–74]. Ribosome profiling and MS-based protein data have the advantage of detecting ORFs that are actually translated, whereas transcriptomic data exclusively provide evidence of transcriptional activity at the respective genomic region. However, experimentally based predictions are generally dependent on cultivation conditions. They only capture snapshots of (potential) sORFs. To get a more complete picture of protein-coding sORFs, it is therefore extremely useful to combine refined in silico bioinformatics predictions with evidence from experimental approaches.

3.1 | Transriptomics

High-throughput transcriptomic approaches based on RNA-seq technologies or tiling DNA microarrays have provided genome-wide maps for many bacteria, including transcription start sites and operons, and have led to the identification of a large number of previously unknown transcripts, many of which are very small and apparently lack long protein-coding ORFs [75–80]. They were assumed to be non-coding. While their existence was well documented, their biological function was controversial. For many of them it is now clear that the RNA-molecule itself has regulatory activities. However, a closer look revealed that some of them may have a dual function, acting both as mRNA and as regulatory RNA [81], raising the question of whether these are actually translated. Targeted or systemic detection of the translation activity and the resulting proteins showed that translation of these RNAs is much more widespread than expected [7, 67, 68, 71, 82]. Based on these data, software tools have been developed to support the prediction of ORFs, including sORFs and to generate protein databases. μProteIns employs RNA-seq data to execute a reference-based assembly, utilising both genomic and transcriptomic data to construct a protein database via 6- and 3-frame translation, respectively. Following the peptide search, an RF classifier is used to filter out low-quality spectra. The remaining unique peptides are then used to validate the identified sORFs, increasing the integrity of the findings [37].

A crucial prerequisite for this is a very comprehensive transcriptome analysis using RNA-seq [79], which makes it possible to create a database with almost all actively transcribed protein-coding genes of a bacterium. Large amounts of RNA-seq data are available in publicly accessible databases for many bacteria under a wide range of growth conditions and represent an important source for evidence-based global sORF predictions in these organisms.

3.2 | Ribosome profiling

Ribosome profiling provides direct evidence for the translation of potential ORFs, regardless of their length or start and stop codons and is therefore independent of genome annotation. It is based on deep sequencing of ribosome-protected mRNA fragments, also called as 'ribosome footprints', and has great potential to improve the annotation of prokaryotic genomes [19, 69, 71, 83, 84]. Advanced ribosome profiling approaches can resolve ORF translation down to the single-codon level. Precise identification of translation start and stop sites is facilitated by either stopping bacterial ribosomes at initiation with retapamulin, Onc12 or tetracyclin, also known as TIS-profiling, or stalling them at termination using apidaecin (translation termination site [TTS]-profiling). This increases the density of ribosomes at either translation start or translation stop codons [19, 71, 85, 86]. Progress has also been made in determining translation activity at ORFs during elongation [87, 88]. Application of the different approaches allows accurate mapping of ORFs, including sORFs and alternative ORFs by identifying also multiple alternative start and stop sites within or outside the coding regions [28]. In addition to ATG, TTG and GTG were found to function as translation start sites [73, 83]. Identification of multiple alternative start sites within the same reading frame provide evidence for the generation of proteoforms, that is, proteins whose translation is initiated from different start sites within the same reading frame and which are characterised by different N-termini but an identical C-terminus. In most cases, it remains to be seen whether they exist and have a distinct function. In addition, internal start codons belonging to alternative ORFs in a different reading frame, whose translation product is completely different from that encoded in the main ORF, can be detected [28, 89]. Alternative ORFs have so far been excluded from the conventional annotation of bacterial genomes to maintain a low FDR, but may play an important role in the production of small proteins. These studies also show that RBSs are not essential for ribosomes to find the TIS, but that they do influence the efficiency of translational initiation [90]. The integration of RNA-seq data revealed that leaderless transcripts also serve as templates for protein synthesis [84, 91]. While ribosome profiling can be applied to a wide range of bacteria, this is not fully the case for special applications such as TIS- and TTS-profiling based on the activity of specific antibiotics. Various antibiotics are available for the latter two procedures [71, 85]. However, they are not equally effective against all bacteria and, it often takes a lot of effort to find the most appropriate drug and its effective concentration [67, 71, 85]. To increase the efficacy of antibiotics, it may also be necessary to genetically modify the bacterium of interest, as described for retapamulin in Gram-negative bacteria, where the drug had limited activity. Deletions of genes encoding major efflux components were the only way to map TISs in *E. coli* (*tolC*) and *Campylobacter jejuni* (*cmeB*) [67, 85].

Several algorithms have been developed to improve genome annotation on the basis of ribosome, TIS- and TTS-profiling data and more specifically to detect sORFs in bacteria [50, 39, 40]. REPARATION introduces an innovative algorithm designed to systematically identify putative protein-coding ORFs in bacterial genomes. It utilises an RF classifier, trained on patterns identified from Ribo-seq data within protein-coding ORFs. To mitigate the biases often associated with in silico prediction methods, it incorporates several validation steps. These include logistic regression models to represent the relationship between ribosome density and ribosome protected fragment (RPF) coverage, thereby estimating minimum read density and ORF RPF coverage. In addition, it implements a rule-based post-processing algorithm to filter out false positives, particularly those overlapping with confirmed coding ORFs.

With smORFer [40] another tool has been introduced that integrates genetic sequence's structural features, in-frame translation data and Fourier transform to generate a measurable score. By integrating TIS-profiling data, the sensitivity of the prediction of start codons of bacterial ORFs is improved. Its modular design allows users to customise sORF searches based on an organism's specific data. Like ab initio methods smORFer identifies putative sORFs without relying solely on existing gene annotations or homology comparisons, making it a useful tool for gaining new insights into the genomic landscape.

A thorough comparison study [50] was conducted on various Ribo-seq-based ORF prediction tools including REPARATION, DeepRibo and smORFer. Using Ribo-seq data from four distinct bacterial species, DeepRibo emerged as a robust tool across the testing data. Despite this, a notable limitation common to all tools was the lack of high sensitivity in detecting sORFs. Even with sufficient Ribo-seq signals, a significant number of experimentally validated sORFs remained undetected by any tool. This finding highlights an ongoing challenge in genomic analysis and emphasises the need for improvements in sORF detection capabilities.

3.3 | Proteogenomics

The aim of the MS-based proteomics is to provide a complete list of the proteins present in a bacterial cell, for example. Proteins are typically identified by matching MS/MS spectra of peptides against theoretical spectra of all candidate peptides derived from proteins encoded by annotated ORFs in the respective genome sequence. However, this also means that only those proteins for which an ORF has been predicted can be identified. To identify novel peptides that are missing from protein databases based on conventional genome annotations, a so-called 'proteogenomics' approach can be applied [92]. Instead of comparing peptide spectra to databases of previously annotated proteins, the MS/MS data need to be compared to customised protein databases [93].

Databases such as translation databases, which consider the full coding capacity of a given bacterial genome, are widely used to analyse MS/MS-data in proteogenomics [2, 72, 94]. Protein sequences are generated using six-frame translation of the genomic sequence. SALT is a freely available algorithm that supports the global translation of bacterial genomes using all six reading frames [2]. A limitation of this strategy is the extremely large size of the resulting database, with a significant (or even excessive) proportion of non-existent protein sequences, which complicates protein identification. Different strategies can be applied to create translational databases. Using stop-to stop translation, a separate protein entry is created for each sequence between two stop codons. This results in a non-redundant database that represents the full coding potential of a given genome sequence [2]. The resulting database size depends on the minimum sequence length cutoff used. An obvious drawback of this database is that the majority of protein sequences obtained are artificial. Additionally, the resulting protein databases almost invariably fail to capture real N-terminal peptides. Alternatively, start-to-stop translation can be applied. The validity of this database depends on the selection of potential translation start sites. Extensive characterisation of the activity of several start codons in *E. coli* and ribosome profiling in several bacteria revealed that ATG in bacteria is mainly used by ribosomes for translation initiation, followed by TTG and GTG [9, 73, 83, 95]. However, other codons such as CTG, ATT, ATC and ATA, which are characterised by differing from ATG in only one position, showed translation initiation in *E. coli* [95]. Which of these codons, and how often, should be used to artificially translate bacterial chromosomes? One strategy would be to prefer ATG and translation of the longest ORF. If there is no ATG, the next alternative start codon that results in the longest translation product could be considered. This also generates a non-redundant database that covers almost the entire coding potential, but similar to stop-to-stop translation, many protein sequences obtained are artificial and do not contain the true N-terminus. This problem can be solved by a start-to-stop translation that considers all possible translation products starting from both the ATG and all non-canonical start codons. The resulting database would have the clear advantage of including all possible N-terminal peptides and would also be suitable for proteoform detection. However, this is a very large and highly redundant database. For a more qualified MS/MS data analysis, it is recommended to create a non-redundant peptide database [9, 72].

To capture the full protein-coding potential of a given bacterial genome sequence, Omasits et al. [94] created an integrated proteogenomics database (iPtgxDB) by combining reference genome annotations with the results of the ab initio gene prediction algorithms Prodigal [34] and ChemGenome [96], and all potential in silico ORFs obtained by a modified six-frame translation considering alternative start codons above a defined length threshold. This database is more specific and therefore less complex than the translational databases described above, as the proportion of non-existent protein sequences is significantly reduced. This has a positive effect on the reliability of MS-based protein identifications. iPtgxDBs have already been successfully used to identify small proteins in a number of bacteria, including *Bartonella henselae*, *B. subtilis*, *Listeria monocytogenes* and *Sinorhizobium meliloti* (Table 2) [68, 94, 97, 98].

When using translation databases that mostly represent artificial peptide sequences, the next critical step is to derive the ORFs from the lists of identified peptides. For this purpose, Pepper, a rule-based expert system (see Info Box 1), has been developed that integrates MS proteomic data with genomic information. The fully automated pipeline processes peptide identifications obtained from searches against a database created from full-genome translations. For ab initio ORF prediction, the pipeline applies a set of expertly curated rules, considering factors such as genomic locations of identified peptides, types of start codons, presence of RBSs and RBS spacer region lengths. [2]. The successful application of Pepper has led to the detection of previously unknown small proteins in different bacterial species, including *S. aureus* and *C. jejuni*, as well as the archaeon *Haloferax volcanii* [2, 7, 67].

It is becoming increasingly clear, at least from extensive TIS-profiling studies, that more than one variant of a protein can arise from a single ORF either through multiple translation or through post-translational modifications [28, 85, 89, 103]. This can have significant effect on the length of the predicted gene products and therefore their functional properties. To increase the accuracy of ORF prediction, especially for sORFs, MS-based proteomics can also assist in the identification of 5-end(s) of ORFs among them sORFs by experimentally identifying the N-terminal end of the proteins [102]. Selective enrichment of N-terminal peptides in combination with high-throughput MS is currently the standard approach for the identification of protein N-terminal sequences, also known as N-terminomics [104, 105]. Blocking free amino groups on the intact proteins by acetylation and subsequent proteolysis of the acetylated proteins results in a mixture of N-terminally acetylated (true N-terminal) and non-acetylated (internal and carboxy-terminal) peptides. The non-acetylated peptides are then removed from acetylated and formylated peptides by amino-specific affinity chromatography. The resulting unbound fraction is highly enriched in N-terminal peptides. These can be analysed by LC-MS/MS [105].

4 | EXPERIMENTAL IDENTIFICATION OF SMALL PROTEINS BY MS-BASED PROTEOMICS IN BACTERIA

In recent years, the identification of small proteins using MS-based techniques has been intensified. Although MS has great potential for the discovery, validation and functional characterisation of small proteins, standard MS approaches have limited applicability for the identification of known and novel small proteins. The identification of small proteins by MS-based proteomics has been hampered not only by the fact that they were ignored by conventional genome annotation algorithms and were therefore missing from protein databases, but also by their low molecular weight, which makes them difficult to prepare and also reduces the number of MS-compatible peptides. As shown for *Salmonella typhimurium*, small proteins clearly suffer from a lower peptide identification rate. It was almost 9% for proteins with more than 100 aa, but only 2.5% for proteins up to 50 aa and 4.5% for proteins between 50 and 100 aa [9].

As a result, the detection of a small protein in a given sample using conventional protocols was more or less random, even if it was present in the database. These problems have been approached in different ways by studies focusing on the systematic identification of small proteins in bacteria. A key objective of method development was to increase the number and intensity of MS-compatible peptides from small proteins in order to improve the sequence coverage and the quality of MS/MS spectra. Critical steps are sample preparation, protease digestion, liquid chromatography (LC), MS data acquisition, peptide spectrum matching, MS-data analysis and protein identification.

4.1 | Enrichment and digestion of small proteins

Several analytical pipelines have been developed to separate small proteins from larger proteins in complex bacterial protein samples prior to digestion and highly sensitive LC-MS analysis. This has been successfully achieved using solid-phase enrichment columns that combine reversed-phase binding and size-based separation [7, 68, 97] (Table 2). During the enrichment step, only small proteins can enter the pores of the column and interact with the column material, while larger proteins pass directly through the column. In *B. subtilis*, this has been shown to increase the absolute number of small proteins identified by a factor of two [97]. Alternatively, intact bacterial proteins can be separated on a gel-free fractionation system [106]. The principle of this technique is based on that of classical SDS-PAGE, in which the gel is polymerised in a columnar cartridge. In this way, proteins migrate

TABLE 2 A selection of studies aimed at the systematic identification of small proteins in bacteria using MS-based shotgun proteomics approaches.

Organism	Protein fraction	Protein prefractionation	Endoprotease	Protein database	Reference
S. typhimurium	Soluble proteins	None	Trypsin LysC	- Reference protein database for *S. typhimurium* strain SL1344 – six frame translation with all ORFs of at least 30 bp initiated from ATG, GTG and TTG	[9]
S. typhimurium	Soluble proteins	None	Trypsin	- Six frame translation with all ORFs of at least 30 bp initiated from ATG, GTG and TTG	[72]
S. typhimurium	Soluble proteins	- GradSeq - 1D SDS-PAGE	Trypsin	- Reference protein database for *S. typhimurium* SL1344 combined with computational sORF predictions (sPepFinder and experimental data (RNA-seq and Ribo-seq data)	[66]
H. volcanii	Soluble proteins	- Solid phase enrichment	LysC	- *H. volcanii* reference protein database - Six frame translation from stop to stop	[7]
B. subtilis	Soluble proteins	- Solid phase enrichment	Trypsin LysC	- *B. subtilis* 168 reference protein database - iPtgxDB	[97]
P. aeruginosa *E. coli* *S. aureus* *M. pneumonia*	Soluble proteins	- Novex 10%—20% Tricine gels	LysC/Trypsin	- Customised protein databases based on RanSEP	[4]
S. aureus	Soluble proteins	- 1D SDS-PAGE	Trypsin LysC AspN	- Six frame translation from stop to stop	[2]
Simplified human intestinal microbiota (SIHUMIx)	Soluble proteins	- 1D SDS-PAGE - Tricine-SDS gel - FASP/MWCO filtration - C8 cartridges - Gelfree 8100 fractionator - Reversed acetone precipitation	Trypsin	- Reference protein database	[6]
M. mazei	Soluble proteins	- Gelfree 8100 fractionator	Trypsin	- Merged protein database, which includes the *M. mazei*, *M. barkeri* and *M. acetivorans* (non-redundant) protein sequences, combined with sORF predictions based on transcriptomic data	[99]
M. mazei	Soluble proteins	- 1D SDS-PAGE	Trypsin	- Reference protein database	[100]
M. mazei	Soluble proteins	- 1D SDS-PAGE	Trypsin Chymotrypsin GluC LysArginase LysC	- Reference protein database of *M. mazei* combined with sORF predictions based on transcriptomic data	[101]
M. mazei	Soluble proteins	None	Trypsin	- Reference protein database	[82]
M. mazei	Soluble proteins	- Acetonitrile-based protein precipitation in combination with 1D SDS-PAGE	Trypsin	- Reference protein database of *M. mazei* combined with sORF predictions based on transcriptomic data	[3]
B. henselae	Subcellular fractions (Cyt, TM, IM, OM)	none	Trypsin Chymotrypsin	- iPtgxDB	[94]
C. jejuni	Soluble proteins	- Gelfree 8100 fractionator	Trypsin Chymotrypsin	- Reference protein database of *C. jejuni* combined with sORF predictions based on Ribo-seq data - Six frame translation from stop to stop	[67]

(Continues)

TABLE 2 (Continued)

Organism	Protein fraction	Protein prefractionation	Endoprotease	Protein database	Reference
L. monocytogenes	Soluble proteins	None	LysC/Trypsin	- iPtgxDB	[98]
L. monocytogenes	Soluble proteins	Isolation of N-terminal peptides by COFRADIC	Trypsin GluC	- Six frame translation from any of the known start codons (ATG, GTG, TTG, CTG, ATT, ATA, ATC) with a minimal length of six amino acids	[102]
S. meliloti	Soluble proteins	- Solid phase enrichment	Trypsin LysC	- iPtgxDB	[68]
Human microbiome *B. thetaiotaomicron*	Soluble proteins	-Acetic acid-based protein precipitation and MWCO filtration	Trypsin	- Reference protein databases combined with customised protein databases with sORFs prediction using Prodigal [34] or MetaProdigal [33]	[56]

MS, mass spectrometry; sORF, small open reading frame.

through the gel and can be collected in soluble fractions. This strategy has been successfully used to increase the number of small proteins identified in several bacteria, both in pure cultures and in a simplified gut microbiome [6, 67, 99] (Table 2). In addition, specific precipitation with organic solvents, such as in acetone or different concentrations of acetonitrile, which depletes the majority of proteins above 15 kDa, has been used to identify small proteins in soluble bacterial protein extracts [3, 6] (Table 2). As is the case with large proteins, the physicochemical properties of small proteins are highly heterogeneous. It is therefore highly unlikely that any single approach will be able to enrich all small proteins expressed in a given organism. For a very comprehensive analysis of small proteins, several studies have combined multiple complementary methods to increase both the number and the confidence of protein identifications [3, 6, 99] (Table 2). In some cases, the total protein concentration can be significantly reduced by using these protocols. For very low concentrated small protein samples, protein digestion using the Single-Pot Solid-Phase-Enhanced Sample Preparation technique has been performed to obtain the maximum number of peptides detectable by MS [67, 107].

Clearly MS-detectability of peptides represents an important limiting factor for identification of small proteins. Peptide predictor tools based on ML have been established to identify theoretical peptides that are refractory towards MS-based detection [108, 109]. These tools can be used for selecting endoproteases more suitable for the identification of small proteins in a given protein database. As the total number of MS compatible peptides is usually very small for small proteins, multi-protease approaches are frequently applied to increase the number of MS-detectable peptides for identification of these proteins [2, 4, 9, 67, 68, 94, 97, 102, 101]. The suitability of the endoproteases used for the identification of small proteins may vary from bacterium to bacterium [2, 67, 97, 101] (Table 2). A very systematic approach using multiple proteases (trypsin, chymotrypsin, LysC, Lysargi-Nase and GluC) in GeLC–MS/MS analysis for the archaeon *Methanosarcina mazei* showed a significant improvement in the identification of small proteins. In total, 91 small proteins were identified with at least two unique peptides and for 39 small proteins a complete sequence coverage was achieved. Using trypsin alone, only 77 small proteins could be identified [101]. For *S. aureus*, the differences between the different endoproteases (LysC, AspN, Trypsin) in small protein identification were not as significant. The results showed that AspN with 48 SP100 was less efficient than LysC (69 SP100) and trypsin (104 SP100). However, a unique set of small proteins (55 SP100 for trypsin, 5 for LysC and 8 for AspN) was also identified for *S. aureus* using those enzymes [2].

4.2 | LC-MS/MS and data analyses

The low number of MS detecteable peptides is also a particular challenge for MS and data analyses in the detection of small proteins. Often only one unique peptide is available for protein identification. High quality MS/MS spectra, stringent filtering criteria and rigorous validation of identified peptides are therefore essential to facilitate high-confidence protein detection and to avoid the reporting of spurious novel protein identifications.

The majority of studies focusing on the identification of small proteins have used conventional bottom-up proteomics approaches based on electrospray ionisation (ESI) MS [1]. Matrix assisted laser desorption/ionisation (MALDI) ionisation has been used as a promising alternative for the identification of small membrane proteins and low complexity samples [1, 110]. Selection of peptide ions for MS/MS analysis is commonly based on peak intensity and resolution (=data dependent acquisition [DAA] mode). However, this results in a loss of peptide information, which is particularly critical for small proteins. Therefore, the use of the data-independent mode may prove to be a clear advantage for the identification of small proteins in the future, as all ions are fragmented for MS/MS [111]. This is also essential for a subsequent global quantification of these proteins using protein/peptide labelling approaches. To improve the detection and sequence coverage of peptides, additional fragmentation methods can be used including higher-energy collision dissociation (HCD), electron-transfer dissociation (ETD) and electron-transfer combined with higher-energy collision dissociation (EThcD) [5, 112–114]. This is useful when longer peptides are expected, for example when using endoproteases with fewer cleavage sites such as AspN or GluC, or when using top-down proteomics for small protein identification [113]. While the commonly used collision-induced dissociation (CID) method selectively fragments the most labile bonds in the peptides and therefore provides limited protein coverage, the use of electron capture/transfer dissociation (ECD/ETD) has been shown to improve sequence coverage [115–117].

When using comprehensive databases, such as translation databases (see above), the search space becomes much larger, resulting in a higher number of random PSMs with high scores and thus significantly reducing the sensitivity of peptide identification [25]. Because the vast majority of entries in a six-frame translation database belong to very small proteins, the likelihood of false positives is therefore greater in this subset [118]. For these reasons, and often only one unique peptide is available to identify small proteins, the FDR of novel peptides or peptides belonging to small proteins differs from the FDR of annotated peptides by several orders of magnitude. The degree of this effect strongly depends on the genome annotation completeness [119]. Therefore, for the identification of small proteins, many studies have used FDRs for PSMs or peptides below 0.1 [2, 7, 94, 101, 100]. In addition, to limit the analyses to high quality PSMs, various spectral and quality-based filtering criteria were applied. Specifically, only MS/MS spectra with sequence tags of at least five consecutive b or y fragment ions or two times four consecutive b or y ions were included [2, 7, 67, 120]. Proteogenomics tool Pepper enables automated assessment of MS/MS spectrum quality [2]. Assignment of correct peptide identifications can also be aided by re-scoring PSMs using ML tools such as Percolator that applies a target decoy approach [121]. Several features describing the quality of PSMs are used for scoring, for example, PSM features provided by the MS-GF+ search engine, including a specific MS-GF+ score and matched fragment peak mass deviations [122]. Other scoring features that were used by ML are the deviation of the predicted peptide retention time (RT) and the intensity of peptide fragment ions [123–125].

Due to their small size, small proteins are excellent candidates for top-down proteomics using native protein extracts. However, this requires special expertise and MS instruments that can routinely achieve high sensitivity, mass accuracy and isotopic resolution, as well as efficient fragmentation of peptides, resulting in high quality MS^2 and MS^3 spectra that provide the most information about the primary structure of peptides/proteins. This approach is promising in particular for the identification and characterisation of small proteins. However, it is not yet widely used because of the challenges it faces [110, 112, 126, 127]. Top-down data can also be used in combination with de novo sequencing to identify small proteins. This is a completely database-independent approach in which protein sequences are inferred from aa-specific mass increases between adjacent fragment peaks and has been successfully applied to the identification of small proteins encoded by the human genome. Meier-Credo et al. [110] present a top-down MALDI-MS/MS approach that is well suited for the identification and sequence analysis of membrane proteins. Using photosystem II as an example, they were able to analyse proteins in the range of 2.5 and 9 kDa with high accuracy and sensitivity. Recently, nanopore sequencing has emerged as an alternative for single-molecule-based primary sequencing and conformational analysis of peptides and proteins [128–134]. In the near future, this technique may also support the identification of small proteins and peptides, either by de novo sequencing or by protein fingerprinting, not only at a global level but also at the level of individual cells.

5 | CONSTRAINTS AND CRITICAL VALIDATION METHODS IN PREDICTING NOVEL SMALL PROTEINS

Bacteria are highly dependent on their environment and closely adapt the expression of their genes to the prevailing conditions. Factors such as temperature, pH, availability of nutrients and interaction with other microbial species as well as host cells all contribute to the gene expression within these microorganisms. For instance, the expression of CsrA, a small RNA-binding protein that serves as global regulator of carbon storage, is modulated in bacteria such as *E. coli* in response to biofilm formation and environmental stressors [135, 136]. Another example is RNAIII in *S. aureus*, which encodes delta-hemolysin, which is mainly expressed during the post-exponential growth phase and whose expression level varies greatly between different isolates [137, 138]. Beyond mere gene expression, the stability of the resulting RNAs and proteins is critical for successful detection, especially when using transcriptomics and/or proteomics methods. By using data from the transcriptome and proteome, activity of metabolic pathways under the conditions studied can be roughly estimated. Complementary genomic analyses can then help to complete our understanding of the role of small proteins. In a noteworthy genomic study involving 1773 metagenomes from four different human body microbiomes (mouth, gut, vagina and skin), more than 4500 conserved families of small proteins were predicted bona fide using MetaProdigal and RNAcode [56]. Surprisingly, the vast majority of these proteins were previously unknown and does not exhibit sequence homology to known protein domains. This work highlights the importance of methods that can be used to study entire microbial consortia to discover novel gene products and provide initial clues to their physiological role.

However, it is important to recognise that genome-wide evidence-based methods for predicting small proteins, including transcriptomics, ribosome profiling and MS-based proteogenomics, generate false positives. Thus, rigorous validation of small proteins predictions with a second independent experimental method is mandatory. For small proteins predicted from transcriptomic or ribosome profiling data, it is recommended that they be detected directly by MS-based methods by integrating the sequence information of the proteins into the database used for MS/MS data analysis [9, 66–68, 99]. Small proteins that have been identified solely by MS-based methods using a proteogenomics approach based on translational databases will also need to be confirmed by a second method. These include immunological methods using specific antibodies, the expression of tagged proteins encoded on a plasmid, or the use of synthetic peptides that match the experimentally detected peptides of the small proteins to check RTs and peptide fragmentation [7, 67, 94].

6 | A GLIMPSE INTO THE FUTURE

ML approaches, particularly deep learning algorithms, have the potential to infer peptide sequences directly from MS/MS spectra without the need for a protein database (for review see [139]). This is a much more challenging task, as it essentially involves de novo peptide sequencing, but it can also provide more comprehensive and unbiased identifications particularly for small proteins. DeepNovo and SMSNet, to name just two, are examples of a deep learning-based algorithms for de novo peptide sequencing. Both use deep CNNs and RNNs to model the patterns in the MS/MS spectra and infer the aa sequence of the peptide that produced the spectrum [140, 141] (see Info Box 1). These algorithms demonstrate the potential of ML for direct peptide identification from MS spectra. However, they also highlight the challenges involved in this task, particularly in terms of the complexity and variability of the spectral data and the need for large, high-quality training datasets. Therefore, although significant progress has been made, there is still much research to be done in this area. Oxford Nanopore Technologies (ONT) has made significant contributions to long-read sequencing technology in genomics [142]. The possibility of applying this technology to direct protein sequencing is currently being explored. The principle of the ONT's nanopore sequencing technology is to detect changes in electrical current as a single molecule (DNA, RNA or possibly protein) passes through a nanopore. These changes can identify single bases in DNA or RNA, or possibly aa in proteins. The adaptation of ONT technology to protein sequencing is still in its early stages and faces significant technical challenges. These include the need to distinguish between the 21 different aa in proteins (compared to the four bases in DNA or RNA), the variable charge and size of aa, and the complex three-dimensional structure of proteins [128–134]. However, if the application of nanopore sequencing to proteins is successful, it could amount to a revolution in proteomics, driving the discovery of new protein families.

The discovery of many small proteins whose CDSs are hidden in microbial genomes is opening up research opportunities across disciplines. Among the crucial questions that remain to be answered in this developing field are how many of these newly discovered proteins are functional and, if so, what their functional roles are—in particular, whether they are involved in pathogenicity or in interactions with the human host or other microorganisms—and how their expression is regulated.

ACKNOWLEDGEMENTS

The authors are very grateful to their colleagues and collaborators who have been involved in the prediction and identification of small proteins for fruitful discussions. This work was supported by an individual project grant within the DFG priority program SPP2002 'Small proteins in prokaryotes, an unexplored world' to S.E. and S.F. (EN 712/4-1), by a grant of the GRK PROCOMPAS (DFG) to S.E. and an institutional grant (INST 188/365-1 FUGG DFG) to S.E.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable – no new data generated, or the article describes entirely theoretical research.

ORCID

Susanne Engelmann https://orcid.org/0000-0002-1201-3488

REFERENCES

1. Ahrens, C. H., Wade, J. T., Champion, M. M., & Langer, J. D. (2022). A practical guide to small protein discovery and characterization using mass spectrometry. *Journal of Bacteriology*, *204*, e0035321. https://doi.org/10.1128/JB.00353-21
2. Fuchs, S., Kucklick, M., Lehmann, E., Beckmann, A., Wilkens, M., Kolte, B., Mustafayeva, A., Ludwig, T., Diwo, M., Wissing, J., Jänsch, L., Ahrens, C. H., Ignatova, Z., & Engelmann, S. (2021). Towards the characterization of the hidden world of small proteins in *Staphylococcus aureus*, a proteogenomics approach. *PLoS Genetics*, *17*, e1009585. https://doi.org/10.1371/journal.pgen.1009585
3. Cassidy, L., Kaulich, P. T., & Tholey, A. (2019). Depletion of high-molecular-mass proteins for the identification of small proteins and short open reading frame encoded peptides in cellular proteomes. *Journal of Proteome Research*, *18*, 1725–1734. https://doi.org/10.1021/acs.jproteome.8b00948
4. Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., & Lluch-Senar, M. (2019). Unraveling the hidden universe of small proteins in bacterial genomes. *Molecular Systems Biology*, *15*, e8290. https://doi.org/10.15252/msb.20188290
5. Müller, S. A., Kohajda, T., Findeiß, S., Stadler, P. F., Washietl, S., Kellis, M., Von Bergen, M., & Kalkhof, S. (2010). Optimization of parameters for coverage of low molecular weight proteins. *Analytical and Bioanalytical Chemistry*, *398*, 2867–2881. https://doi.org/10.1007/s00216-010-4093-x
6. Petruschke, H., Anders, J., Stadler, P. F., Jehmlich, N., & Von Bergen, M. (2020). Enrichment and identification of small proteins in a simplified human gut microbiome. *Journal of Proteomics*, *213*, 103604. https://doi.org/10.1016/j.jprot.2019.103604
7. Hadjeras, L., Bartel, J., Maier, L.-K., Maaß, S., Vogel, V., Svensson, S. L., Eggenhofer, F., Gelhausen, R., Müller, T., Alkhnbashi, O. S., Backofen, R., Becher, D., Sharma, C. M., & Marchfelder, A. (2023). Revealing the small proteome of *Haloferax volcanii* by combining ribosome profiling and small-protein optimized mass spectrometry. *microLife*, *4*, 17.
8. Hemm, M. R., Weaver, J., & Storz, G. (2020). *Escherichia coli* small proteome. *EcoSal Plus*, *9*, https://doi.org/10.1128/ecosalplus.ESP-0031-2019
9. Fijalkowski, I., Willems, P., Jonckheere, V., Simoens, L., & van Damme, P. (2022). Hidden in plain sight: Challenges in proteomics detection of small ORF-encoded polypeptides. *microLife*, *3*, 17.
10. Storz, G., Wolf, Y. I., & Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. *Annual Review of Biochemistry*, *83*, 753–777. https://doi.org/10.1146/annurev-biochem-070611-102400
11. Hobbs, E. C., Fontaine, F., Yin, X., & Storz, G. (2011). An expanding universe of small proteins. *Current Opinion in Microbiology*, *14*, 167–173. https://doi.org/10.1016/j.mib.2011.01.007
12. Gray, T., Storz, G., & Papenfort, K. (2022). Small proteins; Big questions. *Journal of Bacteriology*, *204*, e0034121. https://doi.org/10.1128/JB.00341-21

13. Garai, P., & Blanc-Potard, A. (2020). Uncovering small membrane proteins in pathogenic bacteria: Regulatory functions and therapeutic potential. *Molecular Microbiology*, *114*, 710–720. https://doi.org/10.1111/mmi.14564
14. Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A., & Creevey, C. J. (2022). No one tool to rule them all: Prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics*, *38*, 1198–1207. https://doi.org/10.1093/bioinformatics/btab827
15. Boekhorst, J., Wilson, G., & Siezen, R. J. (2011). Searching in microbial genomes for encoded small proteins. *Microbial Biotechnology*, *4*, 308–313. https://doi.org/10.1111/j.1751-7915.2011.00261.x
16. Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology*, *4*, e1000176. https://doi.org/10.1371/journal.pcbi.1000176
17. Hemm, M. R., Paul, B. J., Schneider, T. D., Storz, G., & Rudd, K. E. (2008). Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology*, *70*, 1487–1501. https://doi.org/10.1111/j.1365-2958.2008.06495.x
18. Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*, 6614–6624. https://doi.org/10.1093/nar/gkw569
19. Vazquez-Laslop, N., Sharma, C. M., Mankin, A., & Buskirk, A. R. (2022). Identifying small open reading frames in prokaryotes with ribosome profiling. *Journal of Bacteriology*, *204*, e0029421. https://doi.org/10.1128/JB.00294-21
20. Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*, 347–355. https://doi.org/10.1038/nature19949
21. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., & Yates, J. R. 3rd (2013). Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, *113*, 2343–2394. https://doi.org/10.1021/cr3003533
22. Maass, S., Moog, G., & Becher, D. (2019). Subcellular protein fractionation in *Legionella pneumophila* and preparation of the derived sub-proteomes for analysis by mass spectrometry. *Methods in Molecular Biology*, *1921*, 445–464. https://doi.org/10.1007/978-1-4939-9048-1_28
23. Omasits, U., Quebatte, M., Stekhoven, D. J., Fortes, C., Roschitzki, B., Robinson, M. D., Dehio, C., & Ahrens, C. H. (2013). Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. *Genome Research*, *23*, 1916–1927. https://doi.org/10.1101/gr.151035.112
24. Becher, D., Hempel, K., Sievers, S., Zühlke, D., Pané-Farré, J., Otto, A., Fuchs, S., Albrecht, D., Bernhardt, J., Engelmann, S., Völker, U., Van Dijl, J. M., & Hecker, M. (2009). A proteomic view of an important human pathogen – towards the quantification of the entire *staphylococcus aureus* proteome. *PLoS ONE*, *4*, e8176. https://doi.org/10.1371/journal.pone.0008176
25. Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, *73*, 2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009
26. Steinberg, R., & Koch, H.-G. (2021). The largely unexplored biology of small proteins in pro- and eukaryotes. *The FEBS Journal*, *288*, 7002–7024. https://doi.org/10.1111/febs.15845
27. Basrai, M. A., Hieter, P., & Boeke, J. D. (1997). Small open reading frames: Beautiful needles in the haystack. *Genome Research*, *7*, 768–771. https://doi.org/10.1101/gr.7.8.768
28. Orr, M. W., Mao, Y., Storz, G., & Qian, S.-B. (2020). Alternative ORFs and small ORFs: Shedding light on the dark proteome. *Nucleic Acids Research*, *48*, 1029–1042. https://doi.org/10.1093/nar/gkz734
29. Kreitmeier, M., Ardern, Z., Abele, M., Ludwig, C., Scherer, S., & Neuhaus, K. (2022). Spotlight on alternative frame coding: Two long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection. *iScience*, *25*, 103844. https://doi.org/10.1016/j.isci.2022.103844
30. Hücker, S. M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., Nelson, C. W., Schloter, M., Rost, B., Scherer, S., & Neuhaus, K. (2017). Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. *PLoS ONE*, *12*, e0184119. https://doi.org/10.1371/journal.pone.0184119
31. Warren, A. S., Archuleta, J., Feng, W.-C., & Setubal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, *11*, 131. https://doi.org/10.1186/1471-2105-11-131
32. Cheng, H., Soon Chan, W., Li, Z., Wang, D., Liu, S., & Zhou, Y. (2011). Small open reading frames: Current prediction techniques and future prospect. *Current Protein & Peptide Science*, *12*, 503–507. https://doi.org/10.2174/138920311796957667
33. Hyatt, D., Locascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, *28*, 2223–2230. https://doi.org/10.1093/bioinformatics/bts429
34. Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119
35. Washietl, S., Findeiß, S., Müller, S. A., Kalkhof, S., Von Bergen, M., Hofacker, I. L., Stadler, P. F., & Goldman, N. (2011). RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, *17*, 578–594. https://doi.org/10.1261/rna.2536111
36. Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S.-H. (2010). sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics*, *26*, 399–400. https://doi.org/10.1093/bioinformatics/btp688
37. de Souza, E. V., Dalberto, P. F., Machado, V. P., Canedo, A., Saghatelian, A., Machado, P., Basso, L. A., & Bizarro, C. V. (2022). μProteInS—a proteogenomics pipeline for finding novel bacterial microproteins encoded by small ORFs. *Bioinformatics*, *38*, 2612–2614. https://doi.org/10.1093/bioinformatics/btac115
38. Clauwaert, J., Menschaert, G., & Waegeman, W. (2019). DeepRibo: A neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Research*, *47*, e36. https://doi.org/10.1093/nar/gkz061
39. Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., & Van Damme, P. (2017). REPARATION: Ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Research*, *45*, e168. https://doi.org/10.1093/nar/gkx758
40. Bartholomäus, A., Kolte, B., Mustafayeva, A., Goebel, I., Fuchs, S., Benndorf, D., Engelmann, S., & Ignatova, Z. (2021). smORFer: A modular algorithm to detect small ORFs in prokaryotes. *Nucleic Acids Research*, *49*, e89. https://doi.org/10.1093/nar/gkab477
41. Badger, J. H., & Olsen, G. J. (1999). CRITICA: Coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, *16*, 512–524. https://doi.org/10.1093/oxfordjournals.molbev.a026133
42. Cerqueira, F. R., & Vasconcelos, A. T. R. (2020). OCCAM: Prediction of small ORFs in bacterial genomes by means of a target-decoy database approach and machine learning techniques. *Database (Oxford)*, *2020*, baaa067. https://doi.org/10.1093/database/baaa067
43. Oheigeartaigh, S. S., Armisen, D., Byrne, K. P., & Wolfe, K. H. (2014). SearchDOGS bacteria, software that provides automated identification of potentially missed genes in annotated bacterial genomes. *Journal of Bacteriology*, *196*, 2030–2042. https://doi.org/10.1128/JB.01368-13
44. Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence

alignments. *Bioinformatics*, *27*, 757–763. https://doi.org/10.1093/bioinformatics/btr010

45. Sommer, M. J., & Salzberg, S. L. (2021). Balrog: A universal protein model for prokaryotic gene prediction. *PLoS Computational Biology*, *17*, e1008727. https://doi.org/10.1371/journal.pcbi.1008727
46. Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*, 2607–2618. https://doi.org/10.1093/nar/29.12.2607
47. Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, *23*, 673–679. https://doi.org/10.1093/bioinformatics/btm009
48. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421
49. Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, *132*, 185–219. https://doi.org/10.1385/1-59259-192-2:185
50. Gelhausen, R., Müller, T., Svensson, S. L., Alkhnbashi, O. S., Sharma, C. M., Eggenhofer, F., & Backofen, R. (2022). RiboReport – benchmarking tools for ribosome profiling-based identification of open reading frames in bacteria. *Briefings in Bioinformatics*, *23*, bbab549. https://doi.org/10.1093/bib/bbab549
51. Korandla, D. R., Wozniak, J. M., Campeau, A., Gonzalez, D. J., & Wright, E. S. (2020). AssessORF: Combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. *Bioinformatics*, *36*, 1022–1029. https://doi.org/10.1093/bioinformatics/btz714
52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
53. Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*, 2444–2448. https://doi.org/10.1073/pnas.85.8.2444
54. Darling, A. E., Carey, L., & Feng, W.-c. (2003). The design, implementation and evaluation of mpiBLAST. In ClusterWorld Conference & Expo and the 4th International Conference on Linux Clusters: The HPC Revolution, June 23 to 26 in 2003 in San Jose, California.
55. Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. In *Current protocols in bioinformatics*, John Wiley & Sons, Inc. (Chapter 3, 3.1.1–3.1.8). https://doi.org/10.1002/0471250953.bi0301s42
56. Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., Pavlopoulos, G. A., Kyrpides, N. C., & Bhatt, A. S. (2019). Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell*, *178*, 1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016
57. Baumgartner, D., Kopf, M., Klähn, S., Steglich, C., & Hess, W. R. (2016). Small proteins in cyanobacteria provide a paradigm for the functional analysis of the bacterial micro-proteome. *BMC Microbiology*, *16*, 285. https://doi.org/10.1186/s12866-016-0896-z
58. Fremin, B. J., Bhatt, A. S., Kyrpides, N. C., & Global Phage Small Open Reading Frame Consortium. (2022). Thousands of small, novel genes predicted in global phage genomes. *Cell Reports*, *39*, 110984. https://doi.org/10.1016/j.celrep.2022.110984
59. Megrian, D., Taib, N., Jaffe, A. L., Banfield, J. F., & Gribaldo, S. (2022). Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria. *Nature Microbiology*, *7*, 2114–2127. https://doi.org/10.1038/s41564-022-01257-y
60. Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088
61. Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*, i275–i282. https://doi.org/10.1093/bioinformatics/btr209
62. Wood, D. E., Lin, H., Levy-Moonshine, A., Swaminathan, R., Chang, Y.-C., Anton, B. P., Osmani, L., Steffen, M., Kasif, S., & Salzberg, S. L. (2012). Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. *Biology Direct*, *7*, 37. https://doi.org/10.1186/1745-6150-7-37
63. Choi, H.-P., Juarez, S., Ciordia, S., Fernandez, M., Bargiela, R., Albar, J. P., Mazumdar, V., Anton, B. P., Kasif, S., Ferrer, M., & Steffen, M. (2013). Biochemical characterization of hypothetical proteins from *Helicobacter pylori*. *PLoS ONE*, *8*, e66605. https://doi.org/10.1371/journal.pone.0066605
64. Akita, H., Kimura, Z.-I., Yusoff, M. Z. M., Nakashima, N., & Hoshino, T. (2016). Draft genome sequence of *Burkholderia* sp. strain CCA53, isolated from leaf soil. *Genome Announcements*, *4*, e00630. https://doi.org/10.1128/genomeA.00630-16
65. Akita, H., Matsushika, A., & Kimura, Z.-I. (2019). *Enterobacter oligotrophica* sp. nov., a novel oligotroph isolated from leaf soil. *MicrobiologyOpen*, *8*, e00843. https://doi.org/10.1002/mbo3.843
66. Venturini, E., Svensson, S. L., Maaß, S., Gelhausen, R., Eggenhofer, F., Li, L., Cain, A. K., Parkhill, J., Becher, D., Backofen, R., Barquist, L., Sharma, C. M., Westermann, A. J., & Vogel, J. (2020). A global data-driven census of *Salmonella* small proteins and their potential functions in bacterial virulence. *microLife*, 1. https://doi.org/10.1093/femsml/uqaa002
67. Froschauer, K., Svensson, S. L., Gelhausen, R., Fiore, E., Kible, P., Klaude, A., Kucklick, M., Fuchs, S., Eggenhofer, F., Engelmann, S., Backofen, R., & Sharma, C. M. (2022). Complementary Ribo-seq approaches map the translatome and provide a small protein census in the foodborne pathogen Campylobacter jejuni. bioRxiv. https://doi.org/10.1101/2022.11.09.515450
68. Hadjeras, L., Heiniger, B., Maaß, S., Scheuer, R., Gelhausen, R., Azarderakhsh, S., Barth-Weber, S., Backofen, R., Becher, D., Ahrens, C. H., Sharma, C. M., & Evguenieva-Hackenberg, E. (2023). Unraveling the small proteome of the plant symbiont *Sinorhizobium meliloti* by ribosome profiling and proteogenomics. *microLife*, *4*, 22.
69. Baek, J., Lee, J., Yoon, K., & Lee, H. (2017). Identification of unannotated small genes in *Salmonella*. *G3: Genes, Genomes, Genetics*, *7*, 983–989. https://doi.org/10.1534/g3.116.036939
70. Stringer, A., Smith, C., Mangano, K., & Wade, J. T. (2021). Identification of novel translated small open reading frames in *Escherichia coli* using complementary ribosome profiling approaches. *Journal of Bacteriology*, *204*, JB0035221. https://doi.org/10.1128/JB.00352-21
71. Weaver, J., Mohammad, F., Buskirk, A. R., & Storz, G. (2019). Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio*, *10*, e02819 https://doi.org/10.1128/mBio.02819-18
72. Willems, P., Fijalkowski, I., & Van Damme, P. (2020). Lost and found: Re-searching and re-scoring proteomics data aids genome annotation and improves proteome coverage. *mSystems*, *5*, e00833–e00820. https://doi.org/10.1128/mSystems.00833-20
73. Smith, C., Canestrari, J. G., Wang, A. J., Champion, M. M., Derbyshire, K. M., Gray, T. A., & Wade, J. T. (2022). Pervasive translation in *Mycobacterium tuberculosis*. *eLife*, *11*, e73980. https://doi.org/10.7554/eLife.73980
74. Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., & Menschaert, G. (2013). Combining *in silico* prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, *14*, 648. https://doi.org/10.1186/1471-2164-14-648
75. Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., & Vogel, J. (2010). The primary transcriptome of the

major human pathogen *Helicobacter pylori*. *Nature*, *464*, 250–255. https://doi.org/10.1038/nature08756

76. Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E. L., Fogg, M. J., Fromion, V., Goelzer, A., … Noirot, P. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*, *335*, 1103–1106. https://doi.org/10.1126/science.1206848
77. Jäger, D., Sharma, C. M., Thomsen, J., Ehlers, C., Vogel, J., & Schmitz, R. A. (2009). Deep sequencing analysis of the *Methanosarcina mazei* Gö1 transcriptome in response to nitrogen availability. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 21878–21882. https://doi.org/10.1073/pnas.0909051106
78. Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., De Los Mozos, I. R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penadés, J. R., Valle, J., Solano, C., & Gingeras, T. R. (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 20172–20177. https://doi.org/10.1073/pnas.1113521108
79. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57–63. https://doi.org/10.1038/nrg2484
80. Laass, S., Monzon, V. A., Kliemt, J., Hammelmann, M., Pfeiffer, F., Förstner, K. U., & Soppa, J. (2019). Characterization of the transcriptome of *Haloferax volcanii*, grown under four different conditions, with mixed RNA-Seq. *PLoS ONE*, *14*, e0215986. https://doi.org/10.1371/journal.pone.0215986
81. Gimpel, M., & Brantl, S. (2017). Dual-function small regulatory RNAs in bacteria. *Molecular Microbiology*, *103*, 387–397. https://doi.org/10.1111/mmi.13558
82. Prasse, D., Thomsen, J., De Santis, R., Muntel, J., Becher, D., & Schmitz, R. A. (2015). First description of small proteins encoded by spRNAs in Methanosarcina mazei strain Gö1. *Biochimie*, *117*, 138–148. https://doi.org/10.1016/j.biochi.2015.04.007
83. Giess, A., Jonckheere, V., Ndah, E., Chyżyńska, K., van Damme, P., & Valen, E. (2017). Ribosome signatures aid bacterial translation initiation site identification. *BMC Biology*, *15*, 76. https://doi.org/10.1186/s12915-017-0416-0
84. Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015). Leaderless transcripts and small proteins are common features of the mycobacterial translational landscape. *PLoS Genetics*, *11*, e1005641. https://doi.org/10.1371/journal.pgen.1005641
85. Meydan, S., Marks, J., Klepacki, D., Sharma, V., Baranov, P. V., Firth, A. E., Margus, T., Kefi, A., Vázquez-Laslop, N., & Mankin, A. S. (2019). Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Molecular Cell*, *74*, 481–493.e6. https://doi.org/10.1016/j.molcel.2019.02.017
86. Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., Yoshikawa, H., Wanner, B. L., Ishihama, Y., & Mori, H. (2016). Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. *DNA Research*, *23*, 193–201. https://doi.org/10.1093/dnares/dsw008
87. Fremin, B. J., Nicolaou, C., & Bhatt, A. S. (2021). Simultaneous ribosome profiling of hundreds of microbes from the human microbiome. *Nature Protocols*, *16*, 4676–4691. https://doi.org/10.1038/s41596-021-00592-4
88. Hwang, J.-Y., & Buskirk, A. R. (2017). A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic Acids Research*, *45*, 327–336. https://doi.org/10.1093/nar/gkw944
89. Meydan, S., Vázquez-Laslop, N., & Mankin, A. S. (2018). Genes within genes in bacterial genomes. *Microbiology Spectrum*, *6*, https://doi.org/10.1128/microbiolspec.rwr-0020-2018
90. Saito, K., Green, R., & Buskirk, A. R. (2020). Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *eLife*, *9*, e55002. https://doi.org/10.7554/eLife.55002
91. Gelsinger, D. R., Dallon, E., Reddy, R., Mohammad, F., Buskirk, A. R., & Diruggiero, J. (2020). Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Research*, *48*, 5201–5216. https://doi.org/10.1093/nar/gkaa304
92. Jaffe, J. D., Berg, H. C., & Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, *4*, 59–77. https://doi.org/10.1002/pmic.200300511
93. Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, *11*, 1114–1125. https://doi.org/10.1038/nmeth.3144
94. Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Québatte, M., Patrignani, A., Dehio, C., Frey, J. E., Robinson, M. D., Wollscheid, B., & Ahrens, C. H. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Research*, *27*, 2083–2095. https://doi.org/10.1101/gr.218255.116
95. Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D., & Salit, M. (2017). Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Research*, *45*, 3615–3626. https://doi.org/10.1093/nar/gkx070
96. Singhal, P., Jayaram, B., Dixit, S. B., & Beveridge, D. L. (2008). Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophysical Journal*, *94*, 4173–4183. https://doi.org/10.1529/biophysj.107.116392
97. Bartel, J., Varadarajan, A. R., Sura, T., Ahrens, C. H., Maaß, S., & Becher, D. (2020). Optimized proteomics workflow for the detection of small proteins. *Journal of Proteome Research*, *19*, 4004–4018. https://doi.org/10.1021/acs.jproteome.0c00286
98. Varadarajan, A. R., Goetze, S., Pavlou, M. P., Grosboillot, V., Shen, Y., Loessner, M. J., Ahrens, C. H., & Wollscheid, B. (2020). A proteogenomic resource enabling integrated analysis of listeria genotype–proteotype–phenotype relationships. *Journal of Proteome Research*, *19*, 1647–1662. https://doi.org/10.1021/acs.jproteome.9b00842
99. Cassidy, L., Prasse, D., Linke, D., Schmitz, R. A., & Tholey, A. (2016). Combination of bottom-up 2D-LC-MS and semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the archaeon methanosarcina mazei. *Journal of Proteome Research*, *15*, 3773–3783. https://doi.org/10.1021/acs.jproteome.6b00569
100. Kaulich, P. T., Cassidy, L., Weidenbach, K., Schmitz, R. A., & Tholey, A. (2020). Complementarity of different SDS-PAGE gel staining methods for the identification of short open reading frame-encoded peptides. *Proteomics*, *20*, 2000084. https://doi.org/10.1002/pmic.202000084
101. Kaulich, P. T., Cassidy, L., Bartel, J., Schmitz, R. A., & Tholey, A. (2021). Multi-protease approach for the improved identification and molecular characterization of small proteins and short open reading frame-encoded peptides. *Journal of Proteome Research*, *20*, 2895–2903. https://doi.org/10.1021/acs.jproteome.1c00115
102. Impens, F., Rolhion, N., Radoshevich, L., Bécavin, C., Duval, M., Mellin, J., García Del Portillo, F., Pucciarelli, M. G, Williams, A. H., & Cossart, P. (2017). N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes*. *Nature Microbiology*, *2*, 17005. https://doi.org/10.1038/nmicrobiol.2017.5
103. Smith, L. M., & Kelleher, N. L. (2013). Proteoform: A single term describing protein complexity. *Nature Methods*, *10*, 186–187. https://doi.org/10.1038/nmeth.2369
104. Berry, I. J., Steele, J. R., Padula, M. P., & Djordjevic, S. P. (2016). The application of terminomics for the identification of protein start sites

and proteoforms in bacteria. *Proteomics*, *16*, 257–272. https://doi.org/10.1002/pmic.201500319

105. Mcdonald, L., & Beynon, R. J. (2006). Positional proteomics: Preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nature Protocols*, *1*, 1790–1798. https://doi.org/10.1038/nprot.2006.317

106. Tran, J. C., & Doucette, A. A. (2008). Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. *Analytical Chemistry*, *80*, 1568–1573. https://doi.org/10.1021/ac702197w

107. Hughes, C. S., Foehr, S., Garfield, D. A., Furlong, E. E., Steinmetz, L. M., & Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular Systems Biology*, *10*, 757. https://doi.org/10.15252/msb.20145625

108. Gao, Z., Chang, C., Yang, J., Zhu, Y., & Fu, Y. (2019). AP3: An advanced proteotypic peptide predictor for targeted proteomics by incorporating peptide digestibility. *Analytical Chemistry*, *91*, 8705–8711. https://doi.org/10.1021/acs.analchem.9b02520

109. Qeli, E., Omasits, U., Goetze, S., Stekhoven, D. J., Frey, J. E., Basler, K., Wollscheid, B., Brunner, E., & Ahrens, C. H. (2014). Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *Journal of Proteomics*, *108*, 269–283. https://doi.org/10.1016/j.jprot.2014.05.011

110. Meier-Credo, J., Preiss, L., Wüllenweber, I., Resemann, A., Nordmann, C., Zabret, J., Suckau, D., Michel, H., Nowaczyk, M. M., Meier, T., & Langer, J. D. (2022). Top-down identification and sequence analysis of small membrane proteins using MALDI-MS/MS. *Journal of the American Society for Mass Spectrometry*, *33*, 1293–1302. https://doi.org/10.1021/jasms.2c00102

111. Zhang, F., Ge, W., Ruan, G., Cai, X., & Guo, T. (2020). Data-independent acquisition mass spectrometry-based proteomics and software tools: A glimpse in 2020. *Proteomics*, *20*, 1900276. https://doi.org/10.1002/pmic.201900276

112. Cassidy, L., Helbig, A. O., Kaulich, P. T., Weidenbach, K., Schmitz, R. A., & Tholey, A. (2021). Multidimensional separation schemes enhance the identification and molecular characterization of low molecular weight proteomes and short open reading frame-encoded peptides in top-down proteomics. *Journal of Proteomics*, *230*, 103988. https://doi.org/10.1016/j.jprot.2020.103988

113. Cristobal, A., Marino, F., Post, H., van den Toorn, H. W. P., Mohammed, S., & Heck, A. J. R. (2017). Toward an optimized workflow for middle-down proteomics. *Analytical Chemistry*, *89*, 3318–3325. https://doi.org/10.1021/acs.analchem.6b03756

114. Olsen, J. V., & Mann, M. (2004). Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 13417–13422. https://doi.org/10.1073/pnas.0405549101

115. Wysocki, V. H., Tsaprailis, G., Smith, L. L., & Breci, L. A. (2000). Mobile and localized protons: A framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, *35*, 1399–1406. https://doi.org/10.1002/1096-9888(200012)35:12⟨1399::AID-JMS86⟩3.0.CO;2-R

116. Zubarev, R. A. (2004). Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology*, *15*, 12–16. https://doi.org/10.1016/j.copbio.2003.12.002

117. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., & Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 9528–9533. https://doi.org/10.1073/pnas.0402700101

118. Burger, T. (2018). Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *Journal of Proteome Research*, *17*, 12–22. https://doi.org/10.1021/acs.jproteome.7b00170

119. Zhang, K., Fu, Y., Zeng, W.-F., He, K., Chi, H., Liu, C., Li, Y.-C., Gao, Y., Xu, P., & He, S.-M. (2015). A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics*, *31*, 3249–3253. https://doi.org/10.1093/bioinformatics/btv340

120. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology*, *9*, 59–64. https://doi.org/10.1038/nchembio.1120

121. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & Maccoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*, 923–925. https://doi.org/10.1038/nmeth1113

122. Granholm, V., Kim, S., Navarro, J. C. F., Sjölund, E., Smith, R. D., & Käll, L. (2014). Fast and accurate database searches with MS-GF+Percolator. *Journal of Proteome Research*, *13*, 890–897. https://doi.org/10.1021/pr400937n

123. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., & Wilhelm, M. (2019). Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, *16*, 509–518. https://doi.org/10.1038/s41592-019-0426-7

124. Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., & Cox, J. (2019). High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, *16*, 519–525. https://doi.org/10.1038/s41592-019-0427-6

125. Silva, A. S. C., Bouwmeester, R., Martens, L., & Degroeve, S. (2019). Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics*, *35*, 5243–5248. https://doi.org/10.1093/bioinformatics/btz383

126. Ferguson, J. T., Wenger, C. D., Metcalf, W. W., & Kelleher, N. L. (2009). Top-down proteomics reveals novel protein forms expressed in *Methanosarcina acetivorans*. *Journal of the American Society for Mass Spectrometry*, *20*, 1743–1750. https://doi.org/10.1016/j.jasms.2009.05.014

127. Kohlstaedt, M., Buschmann, S., Xie, H., Resemann, A., Warkentin, E., Langer, J. D., & Michel, H. (2016). Identification and characterization of the novel subunit CcoM in the cbb3(3)cytochrome c oxidase from *Pseudomonas stutzeri* ZoBell. *mBio*, *7*, e01921–01915. https://doi.org/10.1128/mBio.01921-15

128. Ying, Y.-L., Hu, Z.-L., Zhang, S., Qing, Y., Fragasso, A., Maglia, G., Meller, A., Bayley, H., Dekker, C., & Long, Y.-T. (2022). Nanopore-based technologies beyond DNA sequencing. *Nature Nanotechnology*, *17*, 1136–1146. https://doi.org/10.1038/s41565-022-01193-2

129. Restrepo-Pérez, L., Joo, C., & Dekker, C. (2018). Paving the way to single-molecule protein sequencing. *Nature Nanotechnology*, *13*, 786–796. https://doi.org/10.1038/s41565-018-0236-6

130. Nivala, J., Marks, D. B., & Akeson, M. (2013). Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nature Biotechnology*, *31*, 247–250. https://doi.org/10.1038/nbt.2503

131. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A., & Dekker, C. (2021). Multiple rereads of single proteins at single–amino acid resolution using nanopores. *Science*, *374*, 1509–1513. https://doi.org/10.1126/science.abl4381

132. Hu, Z.-L., Huo, M.-Z., Ying, Y.-L., & Long, Y.-T. (2021). Biological nanopore approach for single-molecule protein sequencing. *Angewandte Chemie (International ed in English)*, *60*, 14738–14749. https://doi.org/10.1002/anie.202013462

133. Callahan, N., Tullman, J., Kelman, Z., & Marino, J. (2020). Strategies for development of a next-generation protein sequencing platform. *Trends in Biochemical Sciences*, *45*, 76–89. https://doi.org/10.1016/j.tibs.2019.09.005

134. Chinappi, M., & Cecconi, F. (2018). Protein sequencing via nanopore based devices: A nanofluidics perspective. *Journal of Physics: Condensed Matter*, *30*, 204002. https://doi.org/10.1088/1361-648X/aababe

135. Jackson, D. W., Suzuki, K., Oakford, L., Simecka, J. W., Hart, M. E., & Romeo, T. (2002). Biofilm formation and dispersal under the influence of the global regulator CsrA of *Escherichia coli*. *Journal of Bacteriology*, *184*, 290–301. https://doi.org/10.1128/JB.184.1.290-301.2002

136. Shimizu, K. (2013). Regulation systems of bacteria such as *Escherichia coli* in response to nutrient limitation and environmental stresses. *Metabolites*, *4*, 1–35. https://doi.org/10.3390/metabo4010001

137. Janzon, L., Löfdahl, S., & Arvidson, S. (1989). Identification and nucleotide sequence of the delta-lysin gene, *hld*, adjacent to the accessory gene regulator (*agr*) of *Staphylococcus aureus*. *Molecular & General Genetics*, *219*, 480–485. https://doi.org/10.1007/BF00259623

138. Ziebandt, A.-K., Kusch, H., Degner, M., Jaglitz, S., Sibbald, M. J. J. B., Arends, J. P., Chlebowicz, M. A., Albrecht, D., Pantuček, R., Doškar, J., Ziebuhr, W., Bröker, B. M., Hecker, M., Van Dijl, J. M., & Engelmann, S. (2010). Proteomics uncovers extreme heterogeneity in the *Staphylococcus aureus* exoproteome due to genomic plasticity and variant gene regulation. *Proteomics*, *10*, 1634–1644. https://doi.org/10.1002/pmic.200900313

139. Wen, B., Zeng, W.-F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., & Zhang, B. (2020). Deep learning in proteomics. *Proteomics*, *20*, 1900335. https://doi.org/10.1002/pmic.201900335

140. Tran, N. H., Zhang, X., Xin, L., Shan, B., & Li, M. (2017). De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 8247–8252. https://doi.org/10.1073/pnas.1705691114

141. Karunratanakul, K., Tang, H.-Y., Speicher, D. W., Chuangsuwanich, E., & Sriswasdi, S. (2019). Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular & Cellular Proteomics*, *18*, 2478–2491. https://doi.org/10.1074/mcp.TIR119.001656

142. Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, *12*, 213–218. https://doi.org/10.1038/nprot.2016.182

How to cite this article: Fuchs, S., & Engelmann, S. (2023). Small proteins in bacteria – big challenges in prediction and identification. *Proteomics*, *23*, e2200421. https://doi.org/10.1002/pmic.202200421