Metaproteomics beyond databases: addressing the challenges and potentials of *de novo* sequencing

Tim Van Den Bossche[1,2], Denis Beslic[3], Sam van Puyenbroeck[1,2], Tomi Suomi[4], Tanja Holstein[1,2,5], Lennart Martens[1,2], Laura L. Elo[6#] and Thilo Muth[5#]*

(1) VIB - UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium

(2) Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9052 Ghent, Belgium

(3) ZKI-PH 3, ZKI-PH, Robert Koch Institute, 13353 Berlin, Germany

(4) Turku Bioscience Centre, University of Turku and Åbo Akademi University, 20520 Turku, Finland

(5) Data Competence Center MF 2, Robert Koch Institute, 13353 Berlin, Germany

(6) Institute of Biomedicine, University of Turku, 20520 Turku, Finland

[#] equal contributions

* corresponding author: Lennart.Martens@UGent.be

Abstract

Metaproteomics enables the large-scale characterization of microbial community proteins, offering crucial insights into their taxonomic composition, functional activities, and interactions within their environments. By directly analyzing proteins, metaproteomics offers insights on community phenotypes and the roles individual members play in diverse ecosystems. While database-dependent search engines are commonly used for peptide identification, they rely on pre-existing protein databases, which can be limiting for complex, poorly characterized microbiomes. *De novo* sequencing presents a promising alternative, which derives peptide sequences directly from mass spectra without requiring a database. Over time, this approach has evolved from manual annotation to advanced graph-based, tag-based, and deep learning-based methods, significantly improving the accuracy of peptide identification. This Viewpoint explores the evolution, advantages, limitations, and future opportunities of *de novo* sequencing in metaproteomics. We highlight recent technological advancements that have improved its potential for detecting unsequenced species and for providing deeper functional insights into microbial communities.

1. Introduction and motivation

Metaproteomics is the large-scale characterization of the entire protein complement of environmental microbiomes, providing insights into the taxonomic composition, functional activities, and interactions of microbial communities within their habitats [1]. It allows the direct study of microbial community functions in a wide range of environments, from human microbiomes to environmental ecosystems like soil and oceans, and offers insights into both the overall community phenotype and the contribution of individual members to the community biomass [2].

However, considerable bottlenecks remain in metaproteomics data analysis, especially for data obtained from complex or poorly characterized environments. This is because metaproteomics studies traditionally identify peptides using search engines that rely on pre-existing protein databases. While this approach works generally well in single-species proteomic studies, where annotated genomes can be used to create comprehensive protein sequence databases, it struggles with the large diversity and unknown nature of many microbial communities, which leads to incomplete, yet often overly large, databases. Recent advances in the accuracy and scalability of deep learning-based methods, however, have now made *de novo* sequencing a compelling alternative for metaproteomics, as *de novo* sequencing derives peptide sequences directly from the observed mass spectra, thus obviating the need for pre-existing sequence databases.

This Viewpoint explores how *de novo* sequencing can address the inherent limitations of database-dependent methods in metaproteomics. We discuss its evolution, highlight its advantages, and provide a future perspective on how recent technological progress is expected to advance the field.

2. Challenges of database-dependent search engines

Database-dependent search engines are the most commonly used approach in metaproteomics, relying on the comparison of experimentally acquired spectra to theoretical spectra derived from protein sequence databases. However, their effectiveness depends on their accuracy and completeness with regards to the sample's actual composition, which can be difficult to achieve for microbiomes [3].

In single-species proteomics, protein databases are typically derived from well-characterized genomes. In metaproteomics, however, the organisms present in the sample may be unknown, and their genomes might not yet be available in public databases, making database construction more complex. Blakeley-Ruiz and Kleiner [4] describe four main approaches for building metaproteomics databases, each with strengths and limitations: matched or unmatched metagenomes, and restricted or unrestricted reference databases (**Figure 1**).

The first type, a *matched metagenomic* protein database, is built from metagenomes specifically assembled from samples similar to the metaproteomic sample. This approach captures the sample's complexity while minimizing the number of irrelevant sequences, as it is directly based on the genomes of the organisms found in (similar) samples. However, assembling a metagenome from DNA reads and translating it into a metaproteome can introduce errors, potentially reducing accuracy. Nevertheless, this approach remains the preferred method when the required resources and expertise are available.

The second type, *restricted reference* databases, is built from taxonomically relevant proteomes, informed by prior knowledge of the sample’s composition. This approach is particularly useful for defined microbial communities, such as the SIHUMIx mock community [5] used in the CAMPI study [6], or when taxonomic data from methods like 16S rRNA gene sequencing is available. However, it is limited by the availability of comprehensive reference proteomes in public repositories, which are often sparse or incomplete for many bacterial species.

The third type, unmatched metagenomic protein databases, is built from metagenomic data from the same overall ecosystem rather than the specific sample. Examples include the Integrated Gene Catalog (IGC) of the human gut microbiome [7], the (expanded) human oral microbiome database (HOMD) [8,9], and the unified Global Ocean Microbiome Catalog (GOMC) [10]. These databases, commonly used for well-characterized ecosystems, are more accessible and enable cross-study comparisons. However, they are less specific, and (much) larger in size than matched metagenome or restricted reference databases, often resulting in lower identification rates and possibly inflated FDR due to increased database size and complexity [11].

The fourth type, *unrestricted reference* databases, includes all available sequences from major repositories like UniProtKB [12], UniRef [13], or NCBI RefSeq [14]. While these very large databases offer the most extensive coverage, they are non-specific and suffer from incomplete or sparse proteomes. This incompleteness, combined with the vast size of these databases, can lead to lower identification rates and potentially inflated FDR [11].

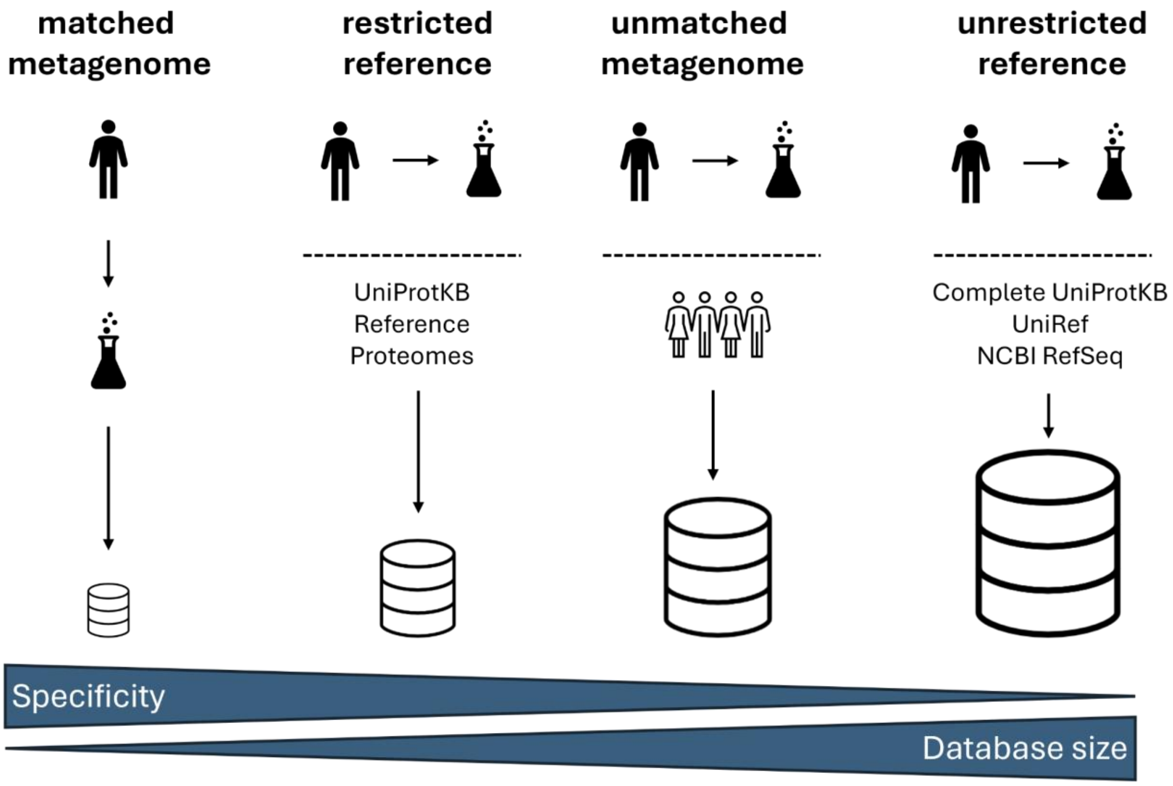

Figure 1. Four main approaches to construct protein databases for metaproteomics. *Matched metagenomes* offer high specificity but require more resources. *Restricted reference* databases use taxonomically relevant proteomes, limited by availability and completeness. *Unmatched metagenomes* combine data from similar ecosystems, enabling cross-study comparisons but increasing size and complexity, reducing accuracy. *Unrestricted reference* databases, like UniProtKB, UniRef, and NCBI RefSeq, provide broad proteome coverage but lower accuracy due to their large size. Gradient bars show the trade-off between specificity and database size.

3. Evolution of *de novo sequencing* in metaproteomics

Construction of databases in metaproteomics is challenging, particularly as large databases increase computational demands, extend search times, and risk missed identifications or false positives [3]. Additionally, identifications are constrained to peptides present in the database, meaning that most sequence variants or novel peptides will be missed. These limitations have driven interest in *de novo* sequencing, which altogether eliminates the need for a protein database during the analysis of the acquired fragmentation (MS/MS) spectra. In a typical *de novo* workflow, these MS/MS spectra are analyzed using a *de novo* sequencing algorithm to infer peptide sequences. Candidate peptides then undergo validation and post-processing before downstream analyses, which can follow either a peptide-centric, or a protein-centric approach. In the peptide-centric approach the obtained peptide sequences are matched exactly to reference databases by tools like Unipept [15], which use precomputed indices to perform taxonomic profiling using the lowest common ancestor algorithm [16], and which provide functional annotations based on the matched proteins. The peptides can also be mapped to KEGG pathways using the PathwayPilot [17]. In the protein-centric approach, however, the obtained peptide sequences are mapped to reference databases using a similarity-based homology search instead, thus allowing the peptides to deviate somewhat from those found in the reference protein sequences. Traditionally, this homology search is carried out using BLAST+ [18–20], and the functions and taxa associated with the thus-identified proteins are then grouped and quantified from the mapped peptides [21].

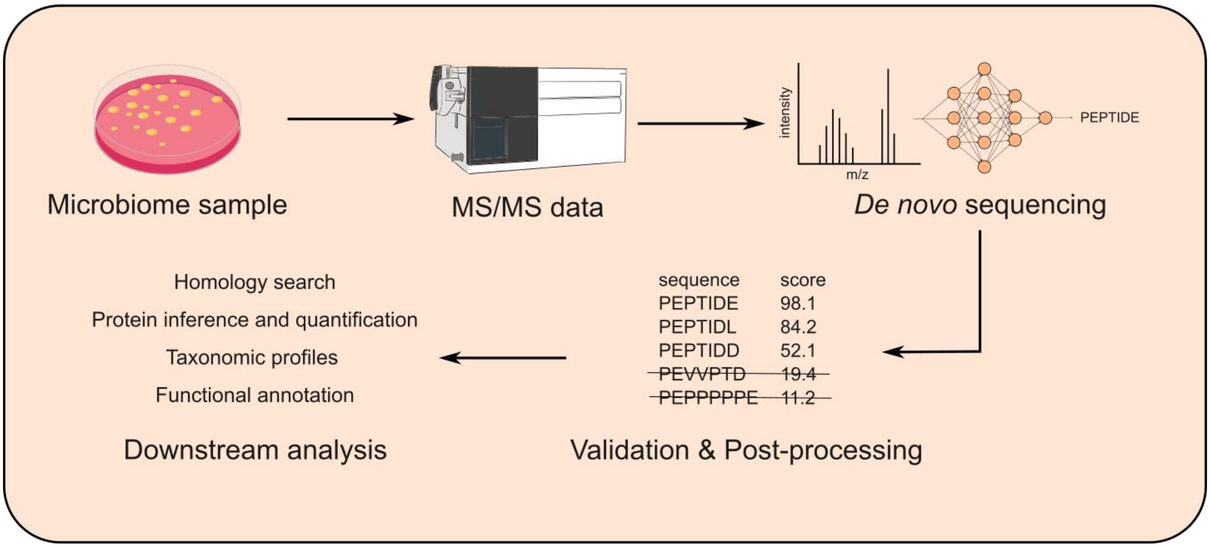

Figure 2. *De novo* sequencing workflow in metaproteomics. A microbiome sample is analyzed by mass spectrometry to generate MS/MS spectra, which are processed using one or more *de novo* algorithms to infer candidate peptide sequences. These sequences undergo validation and post-processing before being used for downstream analyses.

Even though several authors in the metaproteomics field have recommended *de novo* sequencing [11,22,23], its application has remained limited. However, with recent technological advancements, we believe its use may gain renewed attention in the field. In this section, we provide a comprehensive overview of the progress and applications of *de novo* sequencing in metaproteomics.

3.1 Early *de novo* sequencing: from manual annotation to the first algorithmic approaches

In the early days of *de novo* sequencing, spectra were manually annotated using the b and y ion series in the acquired fragmentation spectra, where researchers visually inspected and labeled fragment ions to infer peptide sequences. This manual method was applied in the foundational metaproteomics study by Wilmes and Bond [1] in 2004, where Q-ToF MS on excised 2D-gel spots identified porin, acetyl coenzyme A acetyltransferase, and a component of an ABC-type transport system, likely from dominant, uncultured *Rhodocyclus*-type organisms in activated sludge. Interestingly, modern user interfaces often retain support for this practice, as evidenced by the spectrum viewer in the CompOmics utilities library [24], as used in, amongst others, PeptideShaker [25], which supports metaproteomics data analysis [26].

The first commonly used, automated *de novo* sequencing methods employed approaches like integer linear optimization [27,28], divide-and-conquer [29], and hidden Markov models [30,31]. Graph-based approaches, such as Lutefisk97 [32], Sherenga [33], PEAKS [34], and PepNovo [35], represented MS/MS peaks as nodes, and mass differences between two observed peaks as edges. The highest-scoring path through the graph, from N- to C-terminus, was then used to infer the peptide sequence. In 2015, Muth *et al.* [11] showed that, while *de novo* sequencing only had a 25% peptide overlap with traditional database searches, it often identified novel, unique peptides not present in the database, and provided critical insights by mapping these sequences to reference proteomes. From this, the authors considered *de novo* sequencing as a complementary approach, able to uncover additional peptide identifications that had been missed by traditional database search methods. However, they also noted that a robust method for validation of *de novo* sequencing results would be beneficial.

Tag-based methods, such as MetaNovo [36], further demonstrated the applicability of *de novo* sequencing in metaproteomics. MetaNovo used *de novo* sequence tags to create a sample-specific sequence database in two steps: first, *de novo* sequencing generates *de novo* sequence tags using DirecTag [37], which are mapped to a protein sequence database via PeptideMapper [38]. This step incorporates probabilistic protein inference to rank and filter proteins based on estimated species and protein abundances. In the second step, the filtered sequence database undergoes a conventional database search using MaxQuant [39] for peptide identification, followed by taxonomic analysis using Unipept [15]. When tested on eight human mucosal-luminal interface samples, MetaNovo identified a similar number of peptides and bacterial taxa as MetaPro-IQ [40]. However, it detected significantly more non-bacterial peptides and outperformed workflows relying on matched metagenomes and whole-genome sequencing. MetaNovo improved taxon resolution, identified more taxa of interest, and flagged experimental contaminants, enhancing the accuracy and quality of the sequencing results.

3.2 *De novo* sequencing in the age of machine learning and deep learning

With advancements in machine learning, methods like pNovo [41] and Novor [42] introduced algorithms such as random forests and decision trees to improve *de novo* sequencing accuracy. In 2020, Johnson *et al.* [22] showed that *de novo* sequencing using Novor can be a useful metric for assessing the quality of metaproteomics data, particularly when proteome databases are insufficient. By appending *de novo* sequences to a FASTA file and comparing the subsequent results with a database search, researchers can evaluate the suitability of the protein database. This method is particularly valuable for species without sequenced genomes, where taxonomically related databases are inadequate. The study revealed that poor matches are often due to factors such as post-translational modifications (PTMs), incomplete sequence data, and noise in the spectra. While these issues highlight the limitations of *de novo* sequencing under such conditions, Johnson *et al.* emphasized the need for more sophisticated *de novo* sequencing tools to handle noisy data and PTMs, rather than solely relying on high-quality genome annotations.

In recent years, deep learning approaches have further improved the accuracy and reliability of *de novo* sequencing predictions, particularly in metaproteomics. These advancements are largely driven by neural networks that automatically learn and detect essential features from spectra and peptide sequences, focusing on relevant fragment peaks and subtle patterns. DeepNovo [43], for example, uses a convolutional neural network (CNN) to encode the spectrum and a recurrent neural network (RNN) to decode peptides, predicting one amino acid at a time. Newer architectures, such as transformer models, further enhance precision by accurately mapping even incomplete MS/MS fragmentation patterns to peptide sequences. Tools like DeepNovo [43], Graphnovo [44], Casanovo [45], and PointNovo [46] now achieve amino acid recall rates exceeding 70% on single-organism datasets, compared to only 40% with earlier methods like PepNovo [35].

This improved accuracy of *de novo* sequencing with DeepNovo inspired the development of deep learning workflows for metaproteomics. In 2021, Kleikamp *et al.* [18] developed NovoBridge, a pipeline that starts with DeepNovo or PEAKS results [34] combined with Unipept queries [15], enabling the identification of bacteria absent from existing databases. For example, in an ocean metaproteomics dataset, NovoBridge detected *Alphaproteobacteria* in higher abundance than conventional 16S rRNA sequencing, revealing microorganisms lacking prior sequencing data. Moreover, Kleikamp *et al.* developed a taxonomy validation method using randomized peptide sequences as a control. By comparing correct and randomized sequences, they showed that while random sequences could occasionally match at higher taxonomic levels (like phylum), false positives dropped significantly at lower levels, such as genus or species, leading to more accurate community representation. NovoBridge further uses BLAST+ [47] to homology-match high-quality sequences that could not be annotated by an exact match in Unipept, allowing them to be linked to closely related organisms or assigned to higher taxonomic ranks. Building on their findings with homology searches, they developed NovoLign [48], a pipeline that aligns de novo sequences from metaproteomics experiments using DIAMOND instead of BLAST+, with parameters optimized for short sequences and typical de novo sequencing errors. Applied to Orbitrap Astral MS/MS data, NovoLign significantly improved peptide-spectrum matches (PSMs) and identified additional taxa missed by conventional database searches. These studies highlight the potential of homology-based searches for metaproteomics while demonstrating how deep learning tools like DeepNovo provide de novo predictions that enable these methods, even though their models were not trained on metaproteomics data.

Subsequently, Lee *et al.* [49] demonstrated the utility of such metaproteomics-based training with the Kaiko deep learning model, trained on five million PSMs from 55 phylogenetically diverse bacteria using metaproteomics-specific datasets. Based on DeepNovo's architecture [43], Kaiko outperformed DeepNovo, which was trained on a less representative, nine-species dataset, in identifying organisms from soil and synthetic communities. Here, Kaiko was first used to annotate spectra, followed by taxonomic assignment through DIAMOND against the UniRef100 database, identifying the most abundant organisms in the sample. These identified taxa were used to generate a sample-specific protein database, albeit with the risk of increasing false positive matches [11,50]. This database was then used for traditional database searches, where it uncovered previously unrecognized species like *Candidatus Rokubacteria* and *Candidatus Tectomicrobia*, both absent in 16S rRNA sequencing but critical for carbon and nutrient cycling in terrestrial ecosystems. Kaiko also identified key functional roles through Unipept [15], linking novel species to enzymes like thioredoxin-dependent peroxiredoxin, benzoate degradation enzymes, and streptomycin biosynthesis. Additionally, it revealed that abundant taxa such as *Verrucomicrobia* and *Actinobacteria* are well-represented in metaproteomic data, but missed by traditional sequencing methods, which showcases Kaiko's ability to detect abundant microbial taxa overlooked by other methods. Similarly, π-HelixNovo [51], a transformer-based de novo sequencing model, has advanced the detection of novel peptides by incorporating complementary spectra to enhance ion information. Applied to gut metaproteomes, π-HelixNovo identified significantly more taxon-specific and novel peptides compared to Casanovo, improving taxonomic resolution and peptide recall. For example, π-HelixNovo uncovered a larger number of bacterial-specific peptides (586 versus 350) and species-specific peptides (63 versus 15), demonstrating its capacity to identify previously unseen members of the microbiome. Recently, π-PrimeNovo showed similar improvements on the same dataset, while significantly speeding up the model by making the decoding process non-autoregressive [52].

In addition to these advancements, Kleikamp *et al.* [53] focused on improving the detection of specific PTMs by incorporating SMSNet [54] into their metaproteomics pipeline. Moreover, SMSNet also enhances *de novo* sequencing by addressing a key challenge: the absence of robust methods for FDR estimation. While most *de novo* methods struggle with confidence evaluation due to a lack of decoy sequences [55–57], SMSNet refines its low-confidence predictions by incorporating a peptide database, providing an additional layer of validation. Paired with MSFragger for identifying peptides with PTMs [58], it effectively identified rare modifications like the sulfur-to-selenium (S−Se) mass shift, crucial for detecting selenocysteine, a key adaptation mechanism to oxidative stress in microbial communities. By incorporating DIAMOND [59] for peptide alignment and a reverse decoy strategy, the pipeline maintained a strict 5% FDR, improving identification confidence and greatly speeding up processing compared to BLAST+ [47].

Related to the issues in metaproteomics, forensic samples present unique challenges due to their unknown origin, complicating the identification of the appropriate protein sequence database—similar to studying an organism with an unsequenced genome in microbial communities. Jenson *et al.* developed MARLOWE [60], a computational tool that addresses this by characterizing source organisms in unknown forensic samples. It uses PEAKS, Novor, or Casanovo for de novo sequencing, performs peptide tag extraction and filtering by peptide strength (as defined by Jarman *et al.* previously [61]), assigns tags to taxonomic sources, adjusts counts for tags shared between organisms, and finally scores source organisms using non-negative least squares regression. Tested on biodiversity and Bacillus cereus datasets, MARLOWE successfully identified true contributors in single-source and binary mixtures and distinguished species within a bacterial group, demonstrating its potential for generating forensic leads and aiding follow-up analyses.

In **Table 1**, we provide an overview of tools that have been specifically developed for metaproteomics or have been applied in metaproteomics, either in their original publication or elsewhere. While this table excludes reviews, we would like to highlight the works of Muth and Renard (2018) [62], O’Bryon *et al.* (2020) [23] and Beslic *et al.* (2023) [63] for comprehensive overviews of traditional de novo sequencing tools, as well as Bittremieux *et al.* (2024) [64] for an in-depth review of deep learning de novo sequencing tools.

Table 1. Overview of de novo tools used in, or specifically developed for metaproteomics applications. The Table contains *de novo* tools that either have been specifically developed for metaproteomics applications, or have been applied in metaproteomics - in the original publication or elsewhere. The tools are listed in chronological order based on the original publication date.

Tool	Models used	Original publication	Applications

PEAKS *(commercial software)*	spectrum graph, deep learning (since v10+)	Bin M *et al.*, 2003 [34]	Lacerda CMR *et al.*, 2007 [65] Cantarel BL *et al.*, 2011 [66] Savidor A *et al.*, 2017 [67] Karaduta O *et al.*, 2020 [68]
PepNovo	spectrum graph, probabilistic network modelling	Frank A and Pevzner P, 2005 [35]	Benndorf D *et al.*, 2009 [69] Kuhn R *et al.*, 2011 [70] Cantarel *et al.*, 2011 [66] Hanreich A *et al.*, 2012 [71] Muth T *et al.*, 2015 [11]
Novor	spectrum graph, machine learning, decision tree	Ma B, 2015 [42]	Johnson RS *et al.*, 2020 [22] Thuy-Boun PS *et al.*, 2022 [72]
SMSNet	deep learning	Karunratanakul K *et al.*, 2019 [54]	Kleikamp HBC *et al.*, 2024 [53]
NovoBridge	starts from DeepNovo or PEAKS results	Kleikamp HBC *et al.*, 2021 [18]	
Kaiko	based on DeepNovo	Lee JY *et al.*, 2022 [49]	
MetaNovo	hybrid approach combining de novo sequencing and database search	Potgieter MG *et al.*, 2023 [36]	
π-HelixNovo	deep learning, transformer model	Yang T *et al.*, 2024 [51]	
Casanovo	deep learning, transformer model	Yilmaz M *et al.*, 2024 [45]	Yang T *et al.*, 2024 [51]
NovoLign	starts from DeepNovo or PEAKS results	Kleikamp HBC *et al.*, 2024 [48]	
MARLOWE	starts from PEAKS, Novor, or CasaNovo results (or customized input from other de novo	Jenson SC *et al.*, 2024 [60]	

	sequencing tool)		
π-PrimeNovo	deep learning, transformer model	Xiang Z *et al.*, 2025 [52]	

4. *De novo* sequencing in metaproteomics: *quo vadis*?

4.1 Advantages of *de novo* sequencing in metaproteomics

De novo sequencing offers a promising path forward in metaproteomics by directly identifying peptides from MS/MS spectra without relying on databases, effectively bypassing corresponding limitations [23,73].

De novo sequencing is particularly useful for detecting unsequenced microbial community members, isoforms, and mutations, all of which are often missed by traditional database-dependent approaches. As microbial diversity is often poorly represented in current databases, partly due to the difficulty of culturing many microbes, *de novo* sequencing enables a more complete analysis of microbial communities, even in novel or underexplored environments. By identifying sequence variants and novel peptides, it can therefore provide deeper insights into microbial diversity, and into their (unique) contributions to functional roles in the community.

Furthermore, *de novo* sequencing is much more efficient than database searching in the context of large search spaces, as it only needs to process spectrum information. This efficiency can become beneficial in time-critical scenarios, such as pathogen diagnostics or monitoring.

4.2 Limitations of *de novo* sequencing in metaproteomics

While the potential of *de novo* sequencing is high, it comes with notable limitations. One of the primary challenges is spectral quality. Historically, *de novo* sequencing has been more error-prone than database searches, especially with poor-quality MS/MS spectra. Reliable *de novo* sequencing heavily depends on complete ion fragmentation and minimal noise, as missing fragment ions quickly leads to considerable sequencing ambiguity [57]. This might lead to errors such as substitution errors, mass ambiguities in which a set of residues has the exact or similar combined mass as another residue set; or to inversion errors which are caused by mixing up the b-type fragment ion ladder with the y-type counterpart [62]. In cases of low spectral quality, database searches avoid this ambiguity issue by restricting the allowed sequence space to known peptides, whereas *de novo* methods must contend with a much larger search space. Approaches such as the masking of low-confidence sequence parts with mass deltas, or applying transfer learning to incorporate species-specific data have been proposed to mitigate these issues [18,54]. However, ambiguity-based sequence errors from the *de novo* process can still propagate, complicating the mapping of peptides to proteins, and further downstream to taxa and functions.

Another key challenge in *de novo* sequencing is the absence of a robust method for FDR estimation, as it lacks the decoy sequences used in database-driven methods to control FDR. This makes evaluating confidence in novel peptide identifications more difficult. NovoBoard [74] has proposed a solution using decoy spectra to set a lower-bound on FDR, as had been established in spectral library search approaches [75,76]. Moreover, several other post-processing tools have emerged to refine *de novo* sequencing results: pNovo3 [77] incorporates fragment ion intensity predictions, Spectralis [78] applies Levenshtein distance metrics, PostNovo [79] combines multiple sequencing methods to rescore predictions, SMSNet [54] improves low-confidence predictions with a peptide database, and Instanovo+ [80] includes a diffusion-based decoder to refine sequences. Although these tools all help quantify confidence, no validated method currently exists to set an FDR threshold with strong statistical support, or to compare confidence levels across different *de novo* tools, complicating their integration into broader workflows.

Additionally, while *de novo* sequencing can offer advantages in time-critical assessments by bypassing the need for comprehensive reference databases, its computational demands present a major challenge. Modern *de novo* algorithms often require additional hardware resources such as GPU-based processing, which may not be available in all research environments. The lack of streamlined pipelines also hinders the broader adoption of *de novo* sequencing in metaproteomics. Furthermore, infrequent updates to some post-processing tools can cause compatibility issues with emerging models. These limitations highlight the need for a user-friendly, modular, open-source solution that integrates multiple algorithms, ensures robust validation, and offers scalability, making *de novo* sequencing accessible to more researchers.

4.3 Future opportunities

The first major opportunity lies in advances in deep learning, significantly enhancing both database-dependent (meta)proteomics [81] and *de novo* sequencing. Optimizing deep learning methods can enhance performance of *de novo* sequencing in two critical ways: (i) adapting training data specifically for metaproteomics, and (ii) refining model architectures.

Regarding (i), adapting training data specifically for metaproteomics, *de novo* sequencing methods depend on sequence patterns and training data quality, making the integration of relevant metaproteomics datasets in model training crucial. As more public datasets become available, training deep learning models on more diverse data should further improve their ability to predict novel peptides from unexplored microbial environments. A promising opportunity lies in the transfer learning approach, in which existing models are further ‘fine-tuned’ using the provision of additional training data, though no detailed studies have so far assessed its benefits for metaproteomics. Pre-trained *de novo* models combined with transfer learning could improve performance, particularly for samples with varying spectral noise. However, overfitting of the resulting models remains a concern, and care should always be taken to ensure adequate generalizability of the transfer learned model.

For (ii), refining model architectures, further improvements in *de novo* sequencing can be achieved by optimizing model architectures. In a recent review, Bittremieux *et al.* [64] summarized key advances, such as removing the m/z binning requirement introduced by PointNovo, allowing it to better leverage the high-resolution provided by mass spectrometers such as Thermo’s Astral and Bruker’s TimsTOF Ultra 2. Furthermore, transformer architectures, as seen in Casanovo, simplify model training with built-in attention mechanisms, improving both stability and training speed. Other promising architectures include GraphNovo, which combines a graph-based approach with deep learning, and models like PepNet [82] and π-PrimeNovo [52], which increase inference speed by removing the autoregressive nature of previous models. Although these models show promise, independent benchmarking on metaproteomics datasets is still needed to assess their performance in this context.

The second major opportunity lies in leveraging hardware advancements, which play a crucial role in boosting the performance of *de novo* sequencing models. The integration of graphical processing units (GPUs) allows for faster and more efficient processing of large datasets, while tensor processing units (TPUs) are specifically optimized for accelerating deep learning model training and inference. Furthermore, field-programmable gate arrays (FPGAs), such as those used in RapidOMS [83] for spectral library searching, offer a 60-fold speedup compared to CPUs, and a 2.72-fold speedup compared to GPUs, along with an 11-fold improvement in energy efficiency compared to GPUs. Such efficiency improvements are essential as deep learning models become more complex and datasets grow in size.

The third major opportunity comes from recent advances in mass spectrometry, such as the abovementioned timsTOF Pro and Astral instruments, whose improved sensitivity improves the number of high-quality spectra acquired from complex samples, essential for *de novo* sequencing. The timsTOF Pro combines trapped ion mobility spectrometry (TIMS) with parallel accumulation-serial fragmentation (PASEF), producing cleaner spectra with more finely resolved peaks, which reduces misidentifications and enhances peptide predictions. Similarly, the Astral offers ultra-high mass accuracy and resolution, enabling the detection of low-abundance peptides, thereby expanding proteome coverage. These instruments, along with orthogonal separation techniques like ion mobility spectrometry, help reduce spectral overlap and noise, allowing improved peptide identification and PTM detection. Moreover, *in silico* efforts such as filtering spectra based on quality metrics, such as signal-to-noise ratios [84–89], is therefore crucial for maximizing the accuracy of *de novo* sequencing. Additionally, clustering spectra into consensus representations can yield more comprehensive ion series information [90–93]. Notably, clustering MS/MS data has proven effective for rapid quantitative profiling of metaproteomic samples, reducing the need for extensive database searches [94].

A fourth opportunity lies in the growing adaptation of data-independent acquisition mass spectrometry (DIA-MS) in the proteomics field. However, its adoption in metaproteomics has been slow, partly due to challenges highlighted in a recent benchmark study [95], which noted that the large database sizes required for DIA-metaproteomics need to be reduced [96,97], possibly increasing the risk of false positives. Despite this potential risk, Aakko *et al.* [98] demonstrated the feasibility of DIA-MS for consistent and accurate quantification of microbial samples, proving its applicability in complex metaproteomes using synthetic mixtures and human fecal samples. Moreover, Pietilä *et al.* [99] advanced this further with glaDIAtor, a tool that enables untargeted DIA metaproteomics by creating a pseudospectral library directly from DIA data, eliminating the need for DDA-based libraries and reducing MS runs, thus improving workflow efficiency. However, most *de novo* sequencing tools are trained on DDA-MS data and are generally unsuitable for DIA-MS. Recent tools like DeepNovo-DIA [100], Transformer-DIA [101], PepNet [82], and Cascadia [102] offer promising solutions to this challenge, demonstrating the potential of *de novo* sequencing for DIA-MS. DeepNovo-DIA [100] restructures the long short-term memory (LSTM) network to leverage the extra dimensionality of DIA data, identifying co-eluting precursor ions and their fragments, as well as fragment ions across multiple neighboring spectra. The networks learn 3D shapes of fragment ions, correlations between precursors and their fragments, and peptide sequence patterns. Transformer-DIA [101], an extension of Casanovo, uses self-attention layers and an encoder block to integrate precursor and fragment information, directly incorporating DIA-MS features. Cascadia [102] expands *de novo* sequencing of DIA-MS by using an augmented spectrum, including fragmentation peaks from nearby peptides to capture more signals associated with each peptide.

A fifth opportunity lies in developing user-friendly, scalable workflows that are accessible to researchers without extensive computational expertise. These workflows should integrate advanced *de novo* sequencing tools with key post-processing steps such as FDR estimation, protein inference, taxonomic assignment, and functional annotation. As *de novo* sequencing improves, downstream analysis methods must also advance, particularly technologies that go beyond BLAST+ [47] to handle large metaproteomic references. Hybrid approaches, like MetaNovo [36], which combine *de novo* sequencing with conventional database searches, offer great promise by incorporating probabilistic methods to refine protein and taxon inference. Streamlining these workflows will enable researchers to better understand microbial community dynamics and discover novel proteins, taxa, and functions.

Despite these technical advancements and the availability of benchmark studies in metaproteomics [6,103], there remains a pressing need for dedicated benchmarking efforts focused on *de novo* sequencing methods tailored specifically to metaproteomics. Current benchmarks often focus on single-organism or human-centric datasets, which overlook the complexity and diversity of microbial communities in environmental samples. Effective benchmarks must evaluate factors like accuracy, precision, recall, speed, candidate validation, PTMs, and novel peptides, all crucial for reliable application of *de novo* sequencing in real-world metaproteomics.

5. Conclusion

De novo sequencing has the potential to transform metaproteomics by enabling the identification of novel peptides and thus providing a deeper understanding of microbial communities in environments where traditional database-dependent approaches fall short. While recent advancements in deep learning have substantially improved the accuracy and efficiency of these tools, some challenges remain, particularly in handling poor-quality spectra, managing the large search space inherent in *de novo* approaches, and developing robust validation methods - similar to FDR estimation in database-dependent methods. Despite these challenges, the future of *de novo* sequencing in metaproteomics looks promising. Continued innovation in algorithms, hardware, and software will be essential to unlocking its full potential, making it a critical tool for advancing our understanding of microbial diversity and function in complex environments.

Acknowledgements

This work has benefited from collaborations facilitated by the Metaproteomics Initiative (https://metaproteomics.org/) whose goals are to promote, improve, and standardize metaproteomics [104]. T.V.D.B. acknowledges funding from the Research Foundation Flanders (FWO) [1286824N]. T.H. acknowledges funding from the Joachim-Herz-Foundation. L.M. acknowledges funding by the Research Foundation Flanders (FWO) (G010023N, G028821N), a Ghent University Concerted Research Action (BOF21/GOA/033), and the European Union's Horizon Europe Programme (101080544, 101103253, 101195186, and 10119173). L.L.E. acknowledges funding from the Research Council of Finland [329278, 341342].

Conflict of Interest

The authors declare no conflict of interest.

Abbreviations

CNN	convolutional neural network
CPU	central processing unit
DDA	data-dependent acquisition
DIA	data-independent acquisition
GOMC	global ocean microbiome catalog
GPU	graphical processing unit
HOMD	human oral microbiome database
IGC	integrated gene catalog
LSTM	long short-term memory
RNN	recurrent neural network
TPU	tensor processing unit

References

[1] Wilmes, P., Bond, P.L., The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* 2004, 6, 911–920.

[2] Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., et al., Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* 2017, 8, 1558.

[3] Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B.Y., et al., Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics* 2019, 16, 375–390.

[4] Blakeley-Ruiz, J.A., Kleiner, M., Considerations for constructing a protein sequence

database for metaproteomics. *Comput. Struct. Biotechnol. J.* 2022, 20, 937–952.

[5] Schäpe, S.S., Krause, J.L., Engelmann, B., Fritz-Wallace, K., et al., The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors. *Microorganisms* 2019, 7, 641.

[6] Van Den Bossche, T., Kunath, B.J., Schallert, K., Schäpe, S.S., et al., Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat. Commun.* 2021, 12, 7305.

[7] Li, J., Jia, H., Cai, X., Zhong, H., et al., An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 2014, 32, 834–841.

[8] Chen, T., Yu, W.-H., Izard, J., Baranova, O.V., et al., The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* 2010, 2010, baq013.

[9] Escapa, I.F., Chen, T., Huang, Y., Gajare, P., et al., New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems* 2018, 3, 10.1128/msystems.00187-18.

[10] Chen, J., Jia, Y., Sun, Y., Liu, K., et al., Global marine microbial diversity and its potential in bioprospecting. *Nature* 2024, 633, 371–379.

[11] Muth, T., Kolmeder, C.A., Salojärvi, J., Keskitalo, S., et al., Navigating through metaproteomics data: A logbook of database searching. *PROTEOMICS* 2015, 15, 3439–3453.

[12] The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 2023, 51, D523–D531.

[13] Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., et al., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015, 31, 926–932.

[14] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016, 44, D733–D745.

[15] Vande Moortele, T., Devlaminck, B., Vyver, S.V. de, Van Den Bossche, T., et al., Unipept 6.0: Expanding metaproteomics analysis with support for missed cleavages, semi-tryptic and non-tryptic peptides 2024, 2024.09.26.615136.

[16] Van Den Bossche, T., Verschaffelt, P., Vande Moortele, T., Dawyndt, P., et al., in:, Lisacek F (Ed.), *Protein Bioinforma.*, Springer US, New York, NY 2024, pp. 183–215.

[17] Vande Moortele, T., Verschaffelt, P., Huang, Q., Doncheva, N.T., et al., PathwayPilot: A User-Friendly Tool for Visualizing and Navigating Metabolic Pathways 2024, 2024.06.21.599989.

[18] Kleikamp, H.B.C., Pronk, M., Tugui, C., Silva, L.G. da, et al., Database-independent de novo metaproteomics of complex microbial communities. *Cell Syst.* 2021, 12, 375-383.e5.

[19] Ma, B., Johnson, R., De Novo Sequencing and Homology Searching‡‡. *Mol. Cell. Proteomics MCP* 2012, 11, O111.014902.

[20] Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., et al., Charting the Proteomes of Organisms with Unsequenced Genomes by MALDI-Quadrupole Time-of-Flight Mass Spectrometry and BLAST Homology Searching. *Anal. Chem.* 2001, 73, 1917–1926.

[21] Schallert, K., Verschaffelt, P., Mesuere, B., Benndorf, D., et al., Pout2Prot: An Efficient Tool to Create Protein (Sub)groups from Percolator Output Files. *J. Proteome Res.* 2022, 21, 1175–1180.

[22] Johnson, R.S., Searle, B.C., Nunn, B.L., Gilmore, J.M., et al., Assessing Protein Sequence Database Suitability Using De Novo Sequencing *. *Mol. Cell. Proteomics* 2020, 19, 198–208.

[23] O'Bryon, I., Jenson, S.C., Merkley, E.D., Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide

identification. *Protein Sci.* 2020, 29, 1864–1878.

[24] Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., et al., compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* 2011, 12, 70.

[25] Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 2015, 33, 22–24.

[26] Van Den Bossche, T., Verschaffelt, P., Schallert, K., Barsnes, H., et al., Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for Seamless End-to-End Metaproteomics Data Analysis. *J. Proteome Res.* 2020, 19, 3562–3566.

[27] DiMaggio, P.A., Floudas, C.A., De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal. Chem.* 2007, 79, 1433–1446.

[28] Andreotti, S., Klau, G.W., Reinert, K., Antilope—A Lagrangian Relaxation Approach to the de novo Peptide Sequencing Problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2012, 9, 385–394.

[29] Zhang, Z., De Novo Peptide Sequencing Based on a Divide-and-Conquer Algorithm and Peptide Tandem Spectrum Simulation. *Anal. Chem.* 2004, 76, 6374–6383.

[30] Fischer, B., Roth, V., Roos, F., Grossmann, J., et al., NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing. *Anal. Chem.* 2005, 77, 7265–7273.

[31] Horton, A.P., Robotham, S.A., Cannon, J.R., Holden, D.D., et al., Comprehensive de Novo Peptide Sequencing from MS/MS Pairs Generated through Complementary Collision Induced Dissociation and 351 nm Ultraviolet Photodissociation. *Anal. Chem.* 2017, 89, 3747–3753.

[32] Taylor, J.A., Johnson, R.S., Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 1997, 11, 1067–1075.

[33] Dančík, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A., De Novo Peptide Sequencing via Tandem Mass Spectrometry. *J. Comput. Biol.* 1999, 6, 327–342.

[34] Ma, B., Zhang, K., Hendrie, C., Liang, C., et al., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, 17, 2337–2342.

[35] Frank, A., Pevzner, P., PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* 2005, 77, 964–973.

[36] Potgieter, M.G., Nel, A.J.M., Fortuin, S., Garnett, S., et al., MetaNovo: An open-source pipeline for probabilistic peptide discovery in complex metaproteomic datasets. *PLOS Comput. Biol.* 2023, 19, e1011163.

[37] Tabb, D.L., Ma, Z.-Q., Martin, D.B., Ham, A.-J.L., Chambers, M.C., DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring. *J. Proteome Res.* 2008, 7, 3838–3846.

[38] Kopczynski, D., Barsnes, H., Njølstad, P.R., Sickmann, A., et al., PeptideMapper: efficient and versatile amino acid sequence and tag mapping. *Bioinformatics* 2017, 33, 2042–2044.

[39] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, 26, 1367–1372.

[40] Zhang, X., Ning, Z., Mayne, J., Moore, J.I., et al., MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* 2016, 4, 31.

[41] Chi, H., Sun, R.-X., Yang, B., Song, C.-Q., et al., pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra. *J. Proteome Res.* 2010, 9, 2713–2724.

[42] Ma, B., Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.* 2015, 26, 1885–1894.

[43] Tran, N.H., Zhang, X., Xin, L., Shan, B., Li, M., De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci.* 2017, 114, 8247–8252.

[44] Mao, Z., Zhang, R., Xin, L., Li, M., Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model. *Nat. Mach. Intell.* 2023, 5, 1250–1260.

[45] Yilmaz, M., Fondrie, W.E., Bittremieux, W., Melendez, C.F., et al., Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat. Commun.* 2024, 15, 6427.

[46] Qiao, R., Tran, N.H., Xin, L., Chen, X., et al., Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat. Mach. Intell.* 2021, 3, 420–425.

[47] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., et al., BLAST+: architecture and applications. *BMC Bioinformatics* 2009, 10, 421.

[48] Kleikamp, H.B.C., van der Zwaan, R., van Valderen, R., van Ede, J.M., et al., NovoLign: metaproteomics by sequence alignment. *ISME Commun.* 2024, 4, ycae121.

[49] Lee, J.-Y., Mitchell, H.D., Burnet, M.C., Wu, R., et al., Uncovering Hidden Members and Functions of the Soil Microbiome Using De Novo Metaproteomics. *J. Proteome Res.* 2022, 21, 2023–2035.

[50] Nalpas, N., Hoyles, L., Anselm, V., Ganief, T., et al., An integrated workflow for enhanced taxonomic and functional coverage of the mouse fecal metaproteome. *Gut Microbes* 2021, 13, 1994836.

[51] Yang, T., Ling, T., Sun, B., Liang, Z., et al., Introducing π-HelixNovo for practical large-scale de novo peptide sequencing. *Brief. Bioinform.* 2024, 25, bbae021.

[52] Zhang, X., Ling, T., Jin, Z., Xu, S., et al., π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nat. Commun.* 2025, 16, 267.

[53] Kleikamp, H.B.C., Palacios, P.A., Kofoed, M.V.W., Papacharalampos, G., et al., The Selenoproteome as a Dynamic Response Mechanism to Oxidative Stress in Hydrogenotrophic Methanogenic Communities. *Environ. Sci. Technol.* 2024, 58, 6637–6646.

[54] Karunratanakul, K., Tang, H.-Y., Speicher, D.W., Chuangsuwanich, E., Sriswasdi, S., Uncovering Thousands of New Peptides with Sequence-Mask-Search Hybrid De Novo Peptide Sequencing Framework. *Mol. Cell. Proteomics* 2019, 18, 2478–2491.

[55] Debrie, E., Malfait, M., Gabriels, R., Declerq, A., et al., Quality Control for the Target Decoy Approach for Peptide Identification. *J. Proteome Res.* 2023, 22, 350–358.

[56] Käll, L., Storey, J.D., MacCoss, M.J., Noble, W.S., Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* 2008, 7, 29–34.

[57] Colaert, N., Degroeve, S., Helsens, K., Martens, L., Analysis of the Resolution Limitations of Peptide Identification Algorithms. *J. Proteome Res.* 2011, 10, 5555–5561.

[58] Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., Nesvizhskii, A.I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* 2017, 14, 513–520.

[59] Buchfink, B., Xie, C., Huson, D.H., Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2015, 12, 59–60.

[60] Jenson, S.C., Chu, F., Barente, A.S., Crockett, D.L., et al., MARLOWE: Taxonomic Characterization of Unknown Samples for Forensics Using De Novo Peptide Identification 2024, 2024.09.30.615220.

[61] Jarman, K.H., Heller, N.C., Jenson, S.C., Hutchison, J.R., et al., Proteomics Goes to Court: A Statistical Foundation for Forensic Toxin/Organism Identification Using Bottom-Up Proteomics. *J. Proteome Res.* 2018, 17, 3075–3085.

[62] Muth, T., Renard, B.Y., Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief. Bioinform.* 2018, 19, 954–970.

[63] Beslic, D., Tscheuschner, G., Renard, B.Y., Weller, M.G., Muth, T., Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Brief. Bioinform.* 2023, 24, bbac542.

[64] Bittremieux, W., Ananth, V., Fondrie, W.E., Melendez, C., et al., Deep Learning Methods for De Novo Peptide Sequencing. *Mass Spectrom. Rev.* n.d., n/a.

[65] Lacerda, C.M.R., Choe, L.H., Reardon, K.F., Metaproteomic Analysis of a Bacterial

Community Response to Cadmium Exposure. *J. Proteome Res.* 2007, 6, 1145–1152.

[66] Cantarel, B.L., Erickson, A.R., VerBerkmoes, N.C., Erickson, B.K., et al., Strategies for Metagenomic-Guided Whole-Community Proteomics of Complex Microbial Environments. *PLOS ONE* 2011, 6, e27173.

[67] Savidor, A., Barzilay, R., Elinger, D., Yarden, Y., et al., Database-independent Protein Sequencing (DiPS) Enables Full-length *de Novo* Protein and Antibody Sequence Determination*. *Mol. Cell. Proteomics* 2017, 16, 1151–1161.

[68] Karaduta, O., Glazko, G., Dvanajscak, Z., Arthur, J., et al., Resistant starch slows the progression of CKD in the 5/6 nephrectomy mouse model. *Physiol. Rep.* 2020, 8, e14610.

[69] Benndorf, D., Vogt, C., Jehmlich, N., Schmidt, Y., et al., Improving protein extraction and separation methods for investigating the metaproteome of anaerobic benzene communities within sediments. *Biodegradation* 2009, 20, 737–750.

[70] Kuhn, R., Benndorf, D., Rapp, E., Reichl, U., et al., Metaproteome analysis of sewage sludge from membrane bioreactors. *PROTEOMICS* 2011, 11, 2738–2744.

[71] Hanreich, A., Heyer, R., Benndorf, D., Rapp, E., et al., Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. *Can. J. Microbiol.* 2012, 58, 917–922.

[72] Thuy-Boun, P.S., Wang, A.Y., Crissien-Martinez, A., Xu, J.H., et al., Quantitative Metaproteomics and Activity-based Protein Profiling of Patient Fecal Microbiome Identifies Host and Microbial Serine-type Endopeptidase Activity Associated With Ulcerative Colitis. *Mol. Cell. Proteomics* 2022, 21.

[73] Muth, T., Hartkopf, F., Vaudel, M., Renard, B.Y., A Potential Golden Age to Come—Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *PROTEOMICS* 2018, 18, 1700150.

[74] Tran, N.H., Qiao, R., Mao, Z., Pan, S., et al., NovoBoard: a comprehensive framework for evaluating the false discovery rate and accuracy of de novo peptide sequencing. *Mol. Cell. Proteomics* 2024, 0.

[75] Shiferaw, G.A., Gabriels, R., Bouwmeester, R., Van Den Bossche, T., et al., Sensitive and Specific Spectral Library Searching with CompOmics Spectral Library Searching Tool and Percolator. *J. Proteome Res.* 2022, 21, 1365–1370.

[76] Zhang, Z., Burke, M., Mirokhin, Y.A., Tchekhovskoi, D.V., et al., Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *J. Proteome Res.* 2018, 17, 846–857.

[77] Yang, H., Chi, H., Zeng, W.-F., Zhou, W.-J., He, S.-M., pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* 2019, 35, i183–i190.

[78] Klaproth-Andrade, D., Hingerl, J., Bruns, Y., Smith, N.H., et al., Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. *Nat. Commun.* 2024, 15, 151.

[79] Miller, S.E., Rizzo, A.I., Waldbauer, J.R., Postnovo: Postprocessing Enables Accurate and FDR-Controlled de Novo Peptide Sequencing. *J. Proteome Res.* 2018, 17, 3671–3680.

[80] Eloff, K., Kalogeropoulos, K., Morell, O., Mabona, A., et al., De novo peptide sequencing with InstaNovo: Accurate, database-free peptide identification for large scale proteomics experiments 2024, 2023.08.30.555055.

[81] Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L., Degroeve, S., The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *PROTEOMICS* 2020, 20, 1900351.

[82] Liu, K., Ye, Y., Li, S., Tang, H., Accurate de novo peptide sequencing using fully convolutional neural networks. *Nat. Commun.* 2023, 14, 7974.

[83] Pinge, S., Xu, W., Bittremieux, W., Moshiri, N., et al., RapidOMS: FPGA-based Open Modification Spectral Library Searching with HD Computing. *arXiv.org* 2024.

[84] Xu, H., Freitas, M.A., A Dynamic Noise Level Algorithm for Spectral Screening of Peptide MS/MS Spectra. *BMC Bioinformatics* 2010, 11, 436.

[85] Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., Eidhammer, I., Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *PROTEOMICS* 2006, 6, 2086–2094.

[86] Bielow, C., Hoffmann, N., Jimenez-Morales, D., Van Den Bossche, T., et al., Communicating Mass Spectrometry Quality Information in mzQC with Python, R, and Java. *J. Am. Soc. Mass Spectrom.* 2024, 35, 1875–1882.

[87] Ma, Z.-Q., Chambers, M.C., Ham, A.-J.L., Cheek, K.L., et al., ScanRanker: Quality Assessment of Tandem Mass Spectra via Sequence Tagging. *J. Proteome Res.* 2011, 10, 2896–2904.

[88] Foster, J.M., Degroeve, S., Gatto, L., Visser, M., et al., A posteriori quality control for the curation and reuse of public proteomics data. *PROTEOMICS* 2011, 11, 2182–2194.

[89] Bittremieux, W., Tabb, D.L., Impens, F., Staes, A., et al., Quality control in mass spectrometry-based proteomics. *Mass Spectrom. Rev.* 2018, 37, 697–711.

[90] Flikka, K., Meukens, J., Helsens, K., Vandekerckhove, J., et al., Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *PROTEOMICS* 2007, 7, 3245–3258.

[91] The, M., Käll, L., MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *J. Proteome Res.* 2016, 15, 713–720.

[92] Saeed, F., Hoffert, J.D., Knepper, M.A., CAMS-RS: Clustering Algorithm for Large-Scale Mass Spectrometry Data Using Restricted Search Space and Intelligent Random Sampling. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2014, 11, 128–141.

[93] Falkner, J.A., Falkner, J.W., Yocum, A.K., Andrews, P.C., A Spectral Clustering Approach to MS/MS Identification of Post-Translational Modifications. *J. Proteome Res.* 2008, 7, 4614–4622.

[94] Hao, C., Elias, J.E., Lee, P.K.H., Lam, H., metaSpectraST: an unsupervised and database-independent analysis workflow for metaproteomic MS/MS data using spectrum clustering. *Microbiome* 2023, 11, 176.

[95] Rajczewski, A.T., Blakeley-Ruiz, J.A., Meyer, A., Vintila, S., et al., Data-Independent Acquisition Mass Spectrometry as a Tool for Metaproteomics: Interlaboratory Comparison Using a Model Microbiome 2024, 2024.09.18.613707.

[96] Gómez-Varela, D., Xian, F., Grundtner, S., Sondermann, J.R., et al., Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Front. Microbiol.* 2023, 14.

[97] Dumas, T., Martinez Pinna, R., Lozano, C., Radau, S., et al., The astounding exhaustiveness and speed of the Astral mass analyzer for highly complex samples is a quantum leap in the functional analysis of microbiomes. *Microbiome* 2024, 12, 46.

[98] Aakko, J., Pietilä, S., Suomi, T., Mahmoudian, M., et al., Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota—Implementation and Computational Analysis. *J. Proteome Res.* 2020, 19, 432–436.

[99] Pietilä, S., Suomi, T., Elo, L.L., Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples. *ISME Commun.* 2022, 2, 51.

[100]Tran, N.H., Qiao, R., Xin, L., Chen, X., et al., Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* 2019, 16, 63–66.

[101]Ebrahimi, S., Guo, X., in:, *2023 IEEE 23rd Int. Conf. Bioinforma. Bioeng. BIBE*, 2023, pp. 28–35.

[102]Sanders, J., Wen, B., Rudnick, P., Johnson, R., et al., A transformer model for de novo sequencing of data-independent acquisition mass spectrometry data 2024, 2024.06.03.597251.

[103]Saito, M.A., Saunders, J.K., McIlvin, M.R., Bertrand, E.M., et al., Results from a Multi-Laboratory Ocean Metaproteomic Intercomparison: Effects of LC-MS Acquisition and Data Analysis Procedures. *EGUsphere* 2024, 1–41.

[104]Van Den Bossche, T., Arntzen, M.Ø., Becher, D., Benndorf, D., et al., The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome* 2021, 9, 243.