Natural Products and Bioprospecting

REVIEW

Open Access

Precision enzyme discovery through targeted mining of metagenomic data

Shohreh Ariaeenejad[1], Javad Gharechahi[2], Mehdi Foroozandeh Shahraki[3], Fereshteh Fallah Atanaki[3], Jian-Lin Han[4,5], Xue-Zhi Ding[6], Falk Hildebrand[7,8], Mohammad Bahram[9,10], Kaveh Kavousi[3*] and Ghasem Hosseini Salekdeh[11*]

Abstract

Metagenomics has opened new avenues for exploring the genetic potential of uncultured microorganisms, which may serve as promising sources of enzymes and natural products for industrial applications. Identifying enzymes with improved catalytic properties from the vast amount of available metagenomic data poses a significant challenge that demands the development of novel computational and functional screening tools. The catalytic properties of all enzymes are primarily dictated by their structures, which are predominantly determined by their amino acid sequences. However, this aspect has not been fully considered in the enzyme bioprospecting processes. With the accumulating number of available enzyme sequences and the increasing demand for discovering novel biocatalysts, structural and functional modeling can be employed to identify potential enzymes with novel catalytic properties. Recent efforts to discover new polysaccharide-degrading enzymes from rumen metagenome data using homology-based searches and machine learning-based models have shown significant promise. Here, we will explore various computational approaches that can be employed to screen and shortlist metagenome-derived enzymes as potential biocatalyst candidates, in conjunction with the wet lab analytical methods traditionally used for enzyme characterization.

Keywords Metagenomics, Enzyme bioprospecting, Functional-based screening, Sequence-based screening, Protein structure prediction, Natural products

*Correspondence:
Kaveh Kavousi
kkavousi@ut.ac.ir
Ghasem Hosseini Salekdeh
hsalekdeh@yahoo.com
[1] Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREEO), Karaj, Iran
[2] Human Genetics Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran
[3] Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran
[4] Livestock Genetics Program, International Livestock Research, Institute (ILRI), Nairobi 00100, Kenya
[5] CAAS-ILRI Joint Laboratory On Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing 100193, China
[6] Key Laboratory of Yak Breeding Engineering, Lanzhou Institute of Husbandry and Pharmaceutical Sciences, Chinese Academy of Agricultural Sciences (CAAS), Lanzhou 730050, China
[7] Gut Microbes and Health, Quadram Institute Bioscience, Norwich, Norfolk, UK
[8] Digital Biology, Earlham Institute, Norwich, Norfolk, UK
[9] Department of Ecology, Swedish University of Agricultural Sciences, Ulls Väg 16, 756 51 Uppsala, Sweden
[10] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St, Tartu, Estonia
[11] Faculty of Natural Sciences, Macquarie University, Sydney, NSW, Australia

1 Introduction

Enzymes are becoming increasingly valuable for the development of industrial catalysts due to their ability to significantly enhance the rate of biochemical reactions. High efficiency and selectivity are crucial characteristics for choosing enzymes for commercial applications in bio-catalysis, biofuels, and bioremediation. It is not surprising that microbial enzymes make up most commercial enzymes (88%) used in industry, as they offer several advantages over plant- or animal-derived enzymes [1, 2]. These advantages include higher stability, greater production yield, easier optimization, and increased cost-effectiveness in the industrial applications [3, 4]. Despite their advantages, the number of commercially available microbial enzymes is limited. The use of traditional culture-dependent microbiological methods to screen natural diversity for unknown enzymes is a common approach to obtaining appropriate microbial biocatalysts with desirable properties. This approach involves enriching microorganisms from environmental samples in the presence of appropriate substrates, isolating pure cultures, and screening microbial isolates to ultimately identify enzymes of interest [5]. While this method has proven successful in identifying many commercially available enzymes, more than 99% of microorganisms present in environmental samples cannot be cultured using standard laboratory techniques [6]. The consequence of this limitation is a potential loss of microbial diversity and the opportunity to discover novel enzymes with desired catalytic properties.

Advances in next-generation sequencing technologies have made it possible to access the genome sequences of all microorganisms present in an environment, without the need for their isolation and cultivation. The process of subjecting the DNA extracted from a community of microorganisms recovered from an environmental sample to whole-genome shotgun sequencing is referred to as metagenomics [7, 8]. The method enables direct sequencing of environmental DNA (eDNA) to explore community diversity, functional activities, and interactions of microorganisms inhabiting a specific environment [9]. Metagenomic sequences can be assembled de novo into contigs that represent the genomic segments of microorganisms from which they originate. This allows us to access the coding sequences of enzymes from uncultured microorganisms and predict their functional potentials under specific environmental conditions.

This approach can be used to explore the genomic sequences of unknown microorganisms residing in an environment for the discovery of novel enzymes with improved catalytic properties [10, 11]. It is demonstrated by the steadily increased number of predicted protein-coding sequences from metagenome sequencing of microbial communities obtained from diverse environments. Despite the current annotation pipelines, a significant portion of these sequences remains functionally uncharacterized, leaving many of them as unknown entities. The challenge is further magnified during the identification of a particular enzyme with an enhanced catalytic property. To tackle this problem, there is a growing interest in developing novel computational tools that can model the catalytic properties of enzymes by utilizing shared structural and functional features preserved in their amino acid sequences.

The current experimental approaches for identifying and characterizing new enzymes are limited in terms of speed and throughput, resulting in a gap between the numbers of discovered sequences and enzymes that are experimentally characterized with respect to their catalytic properties [12]. Assaying the activity of these enzyme sequences, especially the large number of novel enzymes predicted from metagenomic sequences, is impractical. Additionally, to realize the industrial application of an enzyme, it needs to be designed to meet specific process requirements [13]. All these limitations highlight the importance of physicochemical and structural features to be considered when searching for enzymes with properties suitable for a specific industrial or biotechnological application. The current approaches for *in-silico* enzyme discovery rely on the properties of enzymes inferred from phylogenetic analyses, sequence similarity searches, genomic positional information, three-dimensional (3D) structural modeling, and predictions based on machine learning [14]. Phylogenetic analyses can help infer the common ancestral origin of enzymes with shared catalytic properties. The sequence divergence that occurs during natural evolution can introduce variability in catalytic properties. The inclusion of sequences from catalytically efficient enzymes in phylogenetic analyses can help to identify distantly related sequences that may possess novel functional activities. Deep learning models can be employed to predict the structures of target enzymes by utilizing multiple sequence alignments and protein contact maps of many metagenomic sequences [15, 16]. Sequence similarity networks are valuable tools for identifying new candidate protein subfamily clusters by leveraging pairwise sequence similarities [17]. Genomic context provides important information regarding substrates, cofactors, bioactivity, and other co-regulated genes associated with the target enzymes [14]. For example, enzymes targeting a specific glycan substrate can be identified based on their genomic localization in polysaccharide utilization loci [18, 19].

Most deciphered 3D structures to date pertain to the enzymes isolated from cultured organisms, leaving limited experimental evidence regarding the structural characteristics of enzymes discovered through metagenome sequencing. Predicting protein function from structural data presents a significant challenge due to numerous instances of highly conserved protein folds that catalyze different reactions [14]. However, performing protein structural similarity searches is a crucial step in narrowing down the sequences that encode enzymes of interest from metagenome datasets. This approach aids in gaining functional insights into unknown sequences, eliminating the need for costly wet lab experiments. When combined with sequence homology-based searches, this approach possesses significant power in shortlisting specific enzymes within vast metagenomic datasets. While the 3D structure of enzymes plays a crucial role in their functionality, its utilization in the *in-silico* bioprospecting of novel enzymes remains limited.

Although several review papers have investigated the application of function-based, homology-based, and machine-learning-based methods to identify, predict, and annotate enzyme-encoding sequences in metagenome data [14, 20–24], there is a lack of knowledge regarding the integration of structural data into annotation pipelines. In this review, we provide an overview of various approaches and pipelines currently employed to explore metagenomic sequences for the discovery of new enzymes. In particular, we will emphasize the significance of incorporating structural data when searching for potential biocatalysts and natural bioproducts in large metagenome datasets.

2 Bioprospecting of novel enzymes from environmental samples

Microbes residing in diverse environments, including soil, hydrothermal vents, saline or alkaline lakes, acid mine drainage, permafrost, hot springs, wastewater treatment sludges, and animal guts, offer the potential for discovering novel enzymatic processes [9]. The microbial communities inhabiting these environments are typically complex in terms of their composition and abundance. It is also worth noting that most members within these communities may not possess desired functions. As a result, comprehensive screening approaches are necessary to identify the desired enzymatic process in a complex environmental sample. Traditional screening approaches involved cultivating microorganisms under defined culture conditions and subsequently screening for microbial clones that exhibit the function of interest. As noted earlier, this approach is unable to capture all microbial diversity in the environment, a phenomenon known as the “great plate count anomaly” [25].

To address the limitations associated with culture-based screening approaches, culture-independent methods were introduced. These methods are classified into two major approaches: functional-based screening (FBS) and sequence-based screening (SBS). Figure 1 provides a comprehensive summary of culture-independent screening approaches commonly used to search for novel biocatalysts in eDNA.

2.1 Function-based screening (FBS)

FBS involves direct cloning of eDNA libraries into suitable vectors, followed by functional screening in surrogate hosts such as *E. coli* [26]. During the screening process, the clones are examined to identify enzymes capable of utilizing a specific substrate or producing a specific product. After identifying the clones with desired functional properties, the DNA encoding for the function of interest is sequenced to identify the gene responsible for observed enzymatic activity. It is important to note that while this method has the potential to screen thousands of eDNA libraries, there are several limitations. The method becomes labor-intensive due to the necessity of analyzing a large number of clones to encompass the entire range of microorganisms present in an environmental sample [27]. The lack of expression of DNA originating from distantly related microorganisms in the surrogate host may result in a reduced representation of significant diversity in the screened samples. In addition, if the desired function relies on the coordinated activity of multiple enzymes, it is essential for all the encoding genes clustered in the same genomic region to be recovered in a single clone [28].

One significant advantage of FBS is its independence from prior knowledge of gene sequences or even the existence of such enzymes. Nevertheless, screening a typical metagenome library necessitates the evaluation of a substantial number of clones. As the complexity of the library increases, this process becomes labor-intensive, time-consuming, and expensive. To expedite the screening process, robotic instruments have been developed, allowing for efficient processing of complex eDNA libraries at a rate of up to 10 million per day. This method enables assaying for a single substrate, thereby reducing the chance of identifying highly promiscuous and multifunctional enzymes [20]. The success of identifying a target enzyme depends on several factors, including the assay method, gene size, gene abundance in the metagenomic sample, host-vector system, and the efficiency of gene expression in the surrogate host [26, 29].

FBS has been widely used for screening novel enzymes, including cellulase [30], esterase [31, 32], carboxylesterase [33], and lipases [34] from diverse environmental sources. In a recent study using activity-based screening through complementary sequence and structure analyses, a novel esterase was isolated by investigating lipolytic enzymes from a compost metagenome library [35]. The same

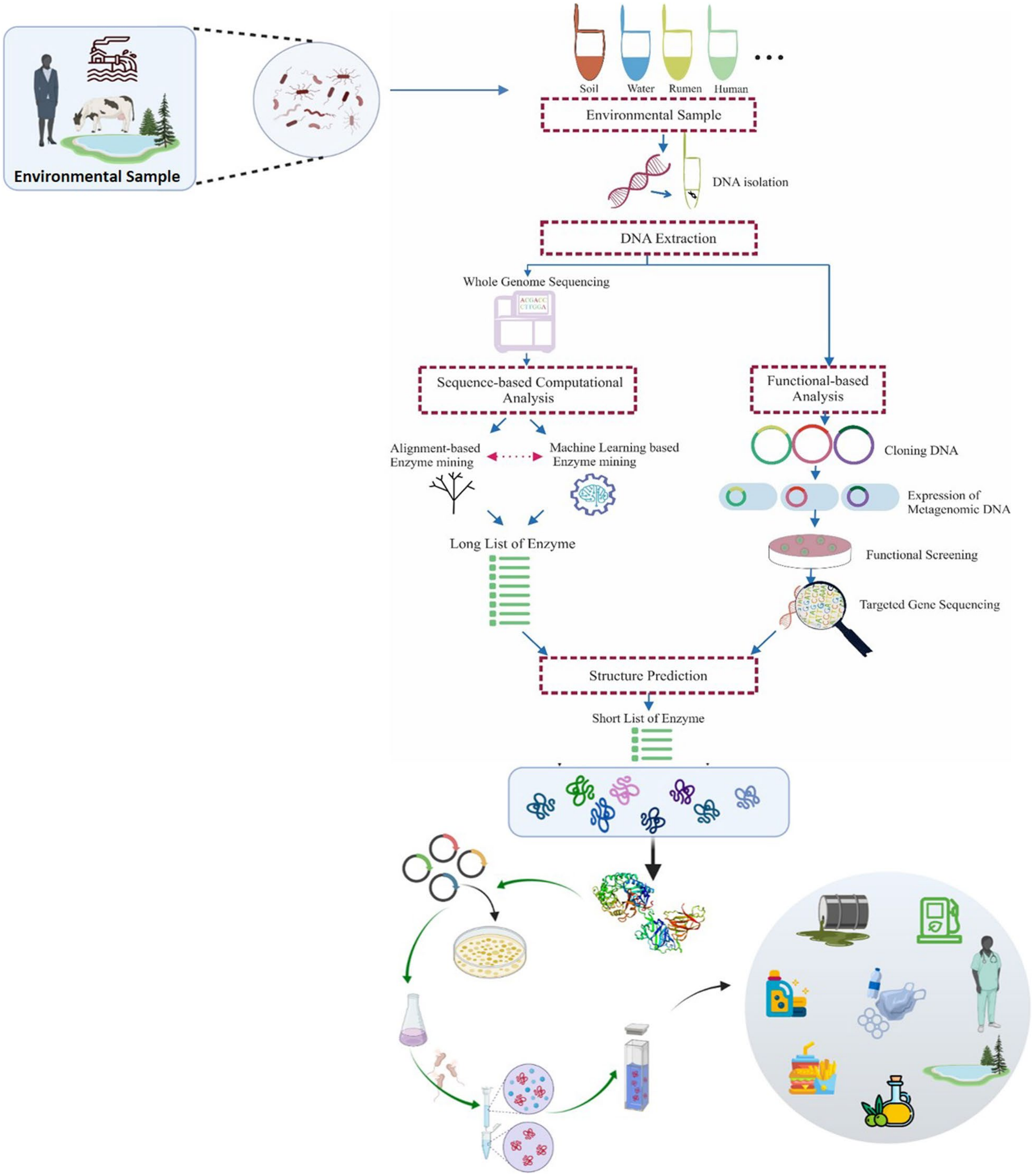

Fig. 1 Culture-independent screening methods for mining novel enzymes from environmental samples. Both function-based and sequence-based methods can benefit from the information gained through structural analysis to refine the initial list of candidate enzymes

approach was used to identify four thermo-alkaliphilic glycosyl hydrolases from wheat straw-degrading microbial consortia [36]. These enzymes hold the potential for utilization in lignocellulosic biomass-degrading cocktails.

2.2 Sequence-based screening (SBS)

The conventional methods for enzyme discovery are generally laborious, costly, resource-intensive, and time-consuming, with no guarantee of success. Due to the availability of a vast number of manually curated protein sequences as well as experimentally characterized enzymes in public databases, the development of novel computational approaches has become imperative to leverage this information in the enzyme discovery process [2, 24]. Specifically, these valuable resources can be utilized to construct machine-learning models that can aid in biocatalyst prospecting. SBS methods expedite the discovery of novel enzymes while minimizing resource usage and achieving a higher success rate.

Prior to the emergence of metagenome sequencing, SBS relied predominantly on the design of primers or probes derived from conserved regions of known enzymes to amplify or screen eDNA libraries in the quest for novel enzyme sequences. This method allows for the identification of novel candidate variants of known enzyme sequences but does not possess the capability to discover entirely new enzymes [27]. Metagenome sequencing has revolutionized the field by enabling the sequencing of complete DNA extracted from a specific environmental sample [37]. Considering this capability, we can now delve into the genetic constituent of every microorganism present in any environment and gain access to all coding sequences, enabling the exploration of any enzymes. The primary challenge associated with this approach lies in accurately annotating the coding sequences predicted in metagenomic sequences. Presently, the annotation process relies on sequence homology searches against known genes or pathways available in public databases. However, the process lacks optimal efficiency, with over 40% of protein-coding sequences remaining unannotated and labeled as unknown or hypothetical. The situation becomes more complex when searching for an enzyme that catalyzes a specific hydrolytic or biosynthetic reaction within a vast number of protein-coding sequences predicted in a metagenomic dataset. The search for a new enzyme through bioprospecting of metagenomic sequences can be carried out by using two general approaches: de novo and reference-based, depending on the availability of known enzyme families [14]. The de novo discovery of new biocatalysts using SBS is challenging, particularly when there is no prior knowledge about the function of interest. Recent studies suggest that predicting protein structures and comparing structural models using residue-residue contact maps can be used to model unknown structures and assist in identifying new biocatalysts in metagenomic datasets [38, 39]. Reference-based methods can be employed when there is existing knowledge about the members of a specific class of enzymes, but the search is for enzymes with distinct functionality. Identifying new enzymes may be less challenging when there are experimentally characterized members, compared to situations where there is a lack of prior knowledge about enzyme function and structure. Robinson et al. [14] proposed a roadmap for metagenomic enzyme discovery, termed "enzyme expansion", which aims to discover enzymes with novel catalysts, substrate specificities, and reaction conditions. Considering that both the reference-based and "enzyme expansion" methods aim to identify enzymes with novel catalytic properties, we have integrated them as a reference-based approach. It is clear that de novo and reference-based methods of enzyme discovery can effectively leverage homology-based (HB), structural-based (SB), and machine learning-based (MLB) analyses.

2.2.1 Homology-based (HB) analysis

Sequence homology search can be used to identify sequences that are closely or distantly related to known enzymes. The method holds significant potential for discovering novel functional homologs of known enzymes [2]. The approach is not only effective in expanding homologs of known enzymes but also capable of searching for enzymes with unique functions. The search is typically conducted on the sequences deposited in public databases, including Pfam, RefSeq, UniPort, and NCBI non-redundant protein (NCBI-nr). Due to inadequate annotations in most publicly available databases, the search results may include hits that are incorrectly annotated [15], thus the results must be manually curated. Several tools have been developed to facilitate the search for closely related sequences, including BLAST [40], DIAMOND [41], and USEARCH [42]. Search algorithms based on either profile Hidden Markov Models (HMMs) such as HMMER [43] or position-specific scoring matrices such as PSI-BLAST [44] can be utilized to identify distantly related sequences. In addition, automated annotation platforms such as MetaHMM [45] and ANASTASIA [46] have been developed to facilitate enzyme discovery through homology-based search. The success of the method hinges on selecting the appropriate target database for the homology search and ensuring the accuracy of annotations for the sequences within that database.

The sequence homology search can be utilized to narrow down a large set of ORFs predicted in a complex metagenome dataset, specifically focusing on enzymes with unique functional characteristics, including thermostability, pH stability, specific activity, and more [47]. In a study by Elbehery et al. [48], HB analysis was carried out to identify two antibiotic resistance genes from the metagenome of Atlantis II Deep Red Sea brine pool. Protein-coding sequences were annotated against sequences deposited in the Comprehensive Antibiotic Resistance Database (CARD, https://card.mcmaster.ca/) using BLASTx, leading to the successful identification of two ORFs encoding a class A beta-lactamase and an aminoglycoside-3' phosphotransferase. The properties of these enzymes were further elucidated through 3D structure prediction. Garg et al. [49] applied HB analysis to identify a novel cellulase (Cel5R) from a soil metagenome. The enzyme was subsequently characterized for its salt- and heat-stable properties. The 3D structure of the enzyme was determined through crystallography. In the landmark study leading to the development of the ANASTASIA platform, a novel esterase named EstDZ4 was mined in a hot spring metagenome [46]. The HB analysis proved successful in identifying EstDZ4, which showed thermostable properties, making it a promising candidate for biotechnological application. The result of this study demonstrated the efficacy of *in-silico* analyses in identifying enzymes that exhibit remote similarity to known sequences.

2.2.2 *Machine learning-based (MLB) analysis*

In HB analysis, it is assumed that homologous sequences share similar functions. However, it is important to acknowledge that there can be exceptions to this rule, where two closely related sequences may possess different functions. Consequently, relying solely on sequence homology may lead to wrongly interpreted or overlooked functional variations. To address these limitations, additional analyses and experimental validations are often necessary to accurately determine functional attributes of closely related sequences. To incorporate additional features in function prediction, methods that leverage MLB analysis can be employed. MLB analysis utilizes advanced algorithms and models to learn patterns and relationships from various data sources, including sequence information, structural properties, physicochemical characteristics, and functional annotations [50, 51]. By considering a broader range of features, MLB analysis can enhance the accuracy and specificity of function prediction, enabling the identification of enzymes with unique and diverse functional characteristics. Moreover, MLB algorithms can detect non-linear relationships and patterns in the data, increasing the likelihood of discovering novel enzymes compared to HB analysis. MLB analysis has demonstrated its effectiveness in uncovering hidden functional relationships and facilitating the discovery of novel biocatalysts with specific catalytic activities and desirable properties.

Several MLB approaches have been developed for the functional classification of the enzymes. Table 1 lists some of the methods that utilize MLB models for the annotation of protein sequences and the prediction of EC numbers. It is important to note that while the methods presented in Table 1 primarily focus on identifying mono-functional enzymes, there are specialized tools such as mlDEEPre [52] that enable the prediction of both multi-functional and mono-functional enzymes.

Table 1 Machine learning algorithms that are used for EC number prediction (all methods are accessible through web server)

Method	Feature type	Machine learning algorithm(s)	EC level prediction	Website	Refs.
EzyPred	pseudo PSSM (Pse-PSSM) and FunD encoding	OET-KNN	Three levels	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/	[53]
SVM-prot	AAC, polarity, hydrophobicity, surface tension, charge, normalized Van der Waals volume, polarizability, secondary structure, solvent accessibility, molecular weight, solubility, number of hydrogen bond donors in side chain, and number of hydrogen bond acceptors in side chain	SVM—KNN—probabilistic neural networks	Three levels	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi	[54]
DEEPre	sequence length-dependent sequence length independent	CNN—RNN	All levels	http://www.cbrc.kaust.edu.sa/DEEPre	[55]
ECPred	information from amino acid sequence alignment and physicochemical properties	KNN, SVM	All levels	https://ecpred.kansil.org/	[56]
CLEAN	–	Contrastive learning	All levels	https://clean.platform.moleculemaker.org/configuration	[57]
HDMLF	–	Deep learning techniques	All levels	https://ecrecer.biodesign.ac.cn	[58]
EnzBert	–	Transformer model techniques	All levels	https://gitlab.inria.fr/nbuton/tfpc	[102]

2.2.2.1 EzyPred EzyPred takes a protein sequence as input and then determines whether it is an enzyme. It then proceeds to classify the enzyme into its respective EC number, main EC class, and subclass. The classification of protein sequences in EzyPred is achieved through the implementation of a machine learning approach known as "optimized evidence-theoretic k-nearest neighbor (OET-KNN)" in conjunction with two types of features to capture information about the protein sequence [53].

2.2.2.2 SVM-prot SVM-prot was initially developed as a computational tool for predicting the EC number of enzymes. It utilizes a representation of the protein sequence using 13 different numerical properties. It employs composition, transition, and distribution to encode each property. The original version used support vector machines (SVM) as the classifier, while it was later updated to utilize two additional classifiers, namely K-nearest neighbors (KNN) and probabilistic neural networks, to expand its prediction capabilities [54]. The incorporation of newer classifiers has significantly improved the overall performance of the method in predicting the EC number of enzymes and their functionality.

2.2.2.3 DEEPre DEEPre is an EC number prediction tool that employs two types of features for mapping a protein sequence into a numerical space [55]. Sequence length-dependent features, such as position-specific scoring matrices (PSSM), and sequence length-independent features, such as functional domain-based encoding are used as input to a deep learning model comprised of a convolutional neural network (CNN) and recurrent neural network (RNN). DEEPre can predict enzyme function on all four levels of the EC classification system.

2.2.2.4 ECPred ECPred is another popular method for predicting the EC number of enzymes [56]. This method adopts an independent learning model for each EC number. The classification is carried out in two levels. In the first level, features based on PSSM and physicochemical properties are utilized, and an SVM classifier is employed. In the second level, features derived from sequence alignments are used for classification by a Nearest Neighbor (NN) classifier.

2.2.2.5 CLEAN CLEAN is a ML algorithm to assign EC number to less-studied proteins or those with uncharacterized functions [57]. CLEAN utilizes a contrastive learning framework, enabling it to confidently assign EC numbers to understudied enzymes, correct mislabeled enzymes, and identify promiscuous enzymes with multiple EC numbers. The effectiveness of CLEAN has been demonstrated through systematic in silico and in vitro experiments.

2.2.2.6 HDMLF HDMLF is a novel hierarchical dual-core multitask learning framework utilizing advanced deep learning techniques for protein sequence embedding and EC number prediction [58]. An attention layer and a greedy strategy optimize the EC prediction process, resulting in stable and superior performance compared to other representative methods. The tool is accessible through the user-friendly web platform ECRECer (https://ecrecer.biodesign.ac.cn) with a cloud-based serverless architecture and an offline package to enhance usability.

2.2.2.7 EnzBert EnzBert is a transformer model for sequence-based protein functional annotation [59]. It predicts the functional enzyme annotations by taking into account only sequence features. When compared to state-of-the-art tools, this model demonstrates superior performance in predicting EC numbers. Specifically, the EnzBert model significantly enhanced accuracy in monofunctional enzyme class prediction and achieved a notable improvement in EC number predictions at level two within the benchmark dataset.

2.2.3 The integrative approaches based on homology and machine learning

Both HB and MLB approaches can be used to discover novel microbial enzymes from environmental samples. Integrating HB and MLB methods increases the accuracy of enzyme discovery and allows for the targeted mining of novel enzymes, thereby reducing the need for costly and time-consuming wet lab experiments. In a previous study, thermostable xylanases were identified by the HB method and further analyzed using an ML-aided approach based on random forest classification [60]. Specifically, they developed a ML model called TAXyl, based on a SVM, which was trained using various sequence-based and length-independent protein features. The model was designed to discriminate between sequences encoding non-thermophilic, thermophilic, and hyperthermophilic xylanases. The model was successfully applied to predict three novel thermostable xylanases from sheep and cow rumen metagenomes.

Furthermore, by integrating HB and MLB approaches, the same group also developed an integrated tool called MCIC, which combines HB and MLB analyses to identify cellulases from metagenomic sequences [61]. MCIC focuses on screening novel cellulases based on their optimal pH and temperature dependencies. The machine learning model employed in MCIC was trained using various sequence-based features. The tool facilitates the comparison of metagenome datasets based on their cellulolytic capabilities. To validate the method, two candidate cellulase enzymes identified by MCIC were cloned and subjected to further characterization.

MeTarEnz (metagenomic targeted enzyme miner) (https://cbb.ut.ac.ir/MeTarEnz/) is a similar software providing various services for targeted isolation of different enzymes from user-defined databases. It accepts sequences in different formats including unassembled short reads, assembled contigs, and translated coding sequences. This software can also predict the optimum pH and temperature of lipolytic enzymes using regression models. It was implemented for an in-depth analysis of tannery wastewater metagenomic data followed by mining a thermophilic alkaline lipase [62].

2.3 Utilization of structural information

The primary goal of bioprospecting enzymes for many industrial applications is to identify those that exhibit optimal functionality under specific conditions. Overcoming obstacles and addressing challenges associated with screening methods will contribute to the development of novel tools and technologies for enzyme discovery through metagenomic analysis. By doing so, we can enhance the efficiency and effectiveness of the bioprospecting process, leading to the identification of enzymes with desired characteristics for various industrial applications. Both SBS and FBS methods generate extensive lists of candidate enzymes. However, characterizing these candidates and identifying specific enzymes with desired properties remains a challenging task. Structural analyses can play a crucial role in narrowing down the search space by reducing the candidate sequences to a limited subset. This targeted subset can then undergo further functional analysis through wet lab procedures. By integrating structural analyses, researchers can efficiently prioritize and focus their experimental efforts on a more manageable set of candidate enzymes, facilitating the identification of enzymes with the desired properties.

It is widely accepted that the 3D structure of an enzyme directly influences its function. However, there are also instances where proteins with similar sequences exhibit dissimilar structures [63]. Surprisingly, even highly similar sequences can lead to proteins with distinct structures. This observed structural dissimilarity often correlates with differences in their functions [63]. There are also examples of proteins with limited sequence similarities but the same folding structures, suggesting that conserved positions in proteins tend to preserve their folding and biological functionality [64]. These findings highlight the complex relationship between protein sequence, structure, and function, demonstrating that sequence similarity alone cannot reliably predict structural similarities or functional properties of enzymes.

The analysis of protein structure–function relationships can be conducted at three levels: amino acid sequence and composition, 3D structure, and spatial conformations of the active site [65]. Computational molecular simulation offers a robust approach for determining and analyzing enzyme structure, dynamics, and functional mechanisms within the framework of physical interactions. Analyzing the 3D structures of enzymes can provide valuable insights into their diverse properties, such as function, spatial conformation, thermal and pH stability.

Prominent methods for predicting 3D protein structures include comparative modeling and ab initio structure prediction [66]. Comparative modeling can be achieved through homology modeling or threading methods for fold recognition. In homology modeling, predictions are based on previously solved structures serving as templates, assuming that homologous proteins share similar 3D structures. Choosing an appropriate template model is crucial for achieving high-quality and accurate predictions. Threading methods involve scanning the primary structure of an unknown protein against a database of proteins with known structures [67, 68]. By employing scoring functions based on statistical or knowledge-based potentials, the compatibility of the query protein with known structure is evaluated. Commonly used tools for comparative modeling include I-TASSER [69], Phyre [70], MODELLER [71], SWISS-MODEL [72], and AlphaFold [73]. Particularly, AlphaFold represents a significant advancement in structure prediction methodologies, leveraging state-of-the-art neural network architectures and training procedures. By integrating evolutionary, physical, and geometric constraints specific to protein structures, AlphaFold achieves remarkable improvements in accuracy.

Ab initio protein structure modeling involves the prediction of protein structures from scratch, relying solely on physical forces and energy principles [74]. This approach is particularly valuable when experimental structural information or suitable template structures are unavailable. Various tools are available to perform ab initio structural prediction, each utilizing different algorithms and methodologies. Notable examples include GROMACS [75], NAMD [76], and TeraChem [77]. These tools employ advanced simulation techniques such as molecular dynamics to explore the conformational space and identify the most energetically favorable protein structure. By leveraging the principles of physics and energy minimization, ab initio modeling enables the generation of protein structures in the absence of prior structural knowledge.

Protein 3D structure modeling plays a crucial role in distinguishing proteins with similar sequences, allowing the exploration of hidden characteristics that cannot be revealed through conventional sequence homology searches alone. This capability becomes particularly valuable when searching for novel enzymes within protein sequences predicted from metagenome data. By providing detailed insights into the spatial arrangement of atoms within a protein, 3D structure modeling aids in the identification of unique structural features, functional regions, and key residues that contribute to enzyme activity. This deeper understanding of protein structure allows for more precise and comprehensive analysis, ultimately facilitating the discovery and characterization of novel enzymes with desired properties from metagenome-derived sequences.

The utilization of structural information has been extensively employed in enzyme bioprospecting from environmental samples, as demonstrated by various studies summarized in Table 2. The processes that lead to the identification of candidate enzymes are summarized into seven distinct stages (S1-S6), with each stage involving specific computational analyses. The different stages of enzyme bioprospecting and their corresponding computational analyses are presented in Table 2.

2.3.1 Predicting enzyme thermal stability through structural analysis

The 3D structure of native proteins is determined by a multitude of weak interactions, including hydrogen bonding, salt bridges, hydrophobic, and polar interactions. These non-covalent forces, along with covalent disulfide bonds between cysteine residues, play essential roles in stabilizing protein structure [78]. These interactions contribute to various structural properties such as protein stability, dynamics, recognition, catalysis, and degradation. Salt bridges are strong electrostatic interactions formed between negatively charged groups [79] that stabilize protein structure and protect the protein from aggregation [80]. The stability of salt bridges is influenced by factors such as pH, distance and geometric orientation of the residues involved. Predicting the presence and location of salt bridges in a protein provides valuable insight into protein stability. There are several freely available tools to predict salt bridges, including Tm predictor [http://tm.life.nthu.edu.tw/], PoPMusic [81], and SCooP [82]. These tools are mainly used to predict changes in the thermodynamic stability, melting temperature, and temperature-dependent stability of a protein.

Hydrogen bonds are another crucial type of interaction that contributes to protein structure. They play a key role in the formation of secondary structures, such as α-helices and β-sheets, by establishing bonds between carbonyl oxygen and amide nitrogen [83]. Several tools are available for predicting the number of hydrogen bonds in a protein, including HBPLUS [84], PyMol [85], and HAAD [86].

Disulfide bonds also play a vital role in the formation of protein structures. They contribute to the stability of protein structures under harsh environments, enhance their mechanical and thermodynamic stability, and minimize the likelihood of misfolding [87]. Computational tools have been developed to accurately predict disulfide-bonding networks and patterns in a protein, thereby aiding in the correct modeling of protein structure. Fariselli et al. [78] introduced a tool for predicting the disulfide bonding state of cysteines in proteins with a prediction accuracy of over 90%.

3 Natural product discovery through metagenomics

Traditionally, the search for bioactive natural products in microorganisms relied largely on activity-based screening approaches [88], which in turn necessitate the isolation and pure culture of the source microorganism. However, recent advances in culture-independent metagenomic and bioinformatic analyses have made it possible to search for novel natural products in microorganisms without the need for their pure culture. This approach offers the exciting potential to delve into the enzymatic mechanisms involved in the biosynthesis and modification of these natural medicinally important compounds. Despite the structural complexity of natural products, their biosynthetic pathways and the enzymes involved in their bioconversion exhibit a remarkable degree of conservation across diverse microbial lineages [23]. This conservation facilitates the discovery, annotation, and characterization of novel natural product biosynthetic enzymes and pathways through sequence homology searches and structural predictions [89]. The combined application of advanced bioinformatic tools and high-throughput screening methodologies offers a powerful approach for targeted mining of metagenomic data, with the potential to significantly accelerate the discovery of novel natural product biosynthesis pathways and subsequent characterization of valuable therapeutic agents and bioactive compounds.

The majority of bacterial natural products fall into the category of secondary metabolites that are encoded by conserved biosynthetic gene clusters (BGCs), a group of two or more closely linked genes that encode enzymes of the biosynthetic pathway for a specific metabolite or natural product [90]. This genomic organization facilitates the identification of natural products through genome mining approaches. Genome mining tools such as antiSMASH [91], PRISM [92], CLUSEAN [93], NP.searcher [94], and NRPminer [95] have been developed to identify putative BGCs in genome or metagenome datasets. AntiSMASH stands out among other tools by offering a

Table 2 The list of candidate enzymes discovered through integrated sequence and structure analyses

Novel metagenomic enzyme	Environment	Computational multi-stage pipeline						
		S1	S2	S3	S4	S5	S6	Refs.
Alkali-thermostable xylanases	*Aspergillus fumigatus*	✓	✓	✓	✓	✓	✓	[103]
Xylanase (XynNTU)	*Paenibacillus campinasensis NTU-11*	–	✓	✓	–	–	–	[104]
Extreme halophilic xylanase	Camel rumen microbiome	✓	–	✓	–	✓	✓	[105]
Thermostable xylanase	Hot spring microbiome	✓	✓	✓	–	✓	✓	[106]
Alkali-thermostable xylanase	Termite gut microbiome	✓	✓	✓	–	–	✓	[107]
Thermostable xylanase	Hot sediment microbiome	✓	✓	✓	–	–	–	[108]
Thermostable xylanase	Hot spring sediment microbiome	✓	✓	✓	–	–	✓	[109]
Thermostable xylanase	Camel rumen microbiome	✓	–	–	–	–	✓	[110]
Thermostable xylanase	Cattle rumen microbiome	–	✓	✓	✓	–	✓	[111]
Thermostable xylanase	Pulp and paper wastewater microbiome	–	✓	✓	✓	–	–	[112]
Alkali-thermostable xylanases *(PersiXyn1)*	Camel rumen microbiome	✓	–	✓	✓	–	✓	[113]
Alkali-thermostable xylanases *(PersiXyn2)*	Camel rumen microbiome	✓	✓	✓	✓	–	✓	[114]
Alkali-thermostable Xylanase *(PersiXyn3,4)*	Cattle rumen microbiome	✓	–	✓	✓	✓	✓	[115]
Thermal dependent xylanases *(PersiXyn5,6,7)*	Sheep and cattle rumen microbiome	✓	–	✓	✓	✓	✓	[60]
Thermostable xylanase *(PersiXyn8)*	Cattle rumen microbiome	✓	–	✓	✓	✓	✓	[116]
Hyperthermostable xylanase *(PersiXyn10)*	Camel rumen microbiome	✓	–	✓	✓	–	✓	[117]
xylanase/ esterase	Cattle rumen microbiome	✓	–	✓	✓	–	–	[118]
Bifunctional mannanase/xylanase *(PersiManXyn1)*	Sheep rumen microbiome	✓	–	✓	✓	–	✓	[119]
Xylanase/β-glucosidase *(PersiBGLXyn1)*	Cattle rumen microbiome	✓	–	✓	✓	–	✓	[120]
Thermostable cellulase	Soil microbiome	–	✓	✓	✓	✓	✓	[121]
Thermostable cellulase	Cattle rumen microbiome	✓	–	✓	✓	–	✓	[122]
Hyperthermophilic cellulase	Arctic Mid-Ocean Ridge vent field microbiome	✓	–	✓	✓	–	✓	[123]
Acidic cellulase	Buffalo rumen microbiome	✓	✓	✓	✓	–	✓	[124]
Alkaline-thermostable cellulase	Goat rumen microbiome	✓	–	✓	✓	–	✓	[125]
Thermostable endoglucanase	Termite gut microbiome	✓	✓	✓	✓	–	✓	[126]
Alkalic and thermostable cellulase *(PersiCel1,2)*	Camel rumen microbiome	✓	–	✓	✓	✓	✓	[127]
Thermostable and halotolerant cellulase *(PersiCel3)*	Sheep rumen microbiome	✓	–	✓	✓	✓	✓	[128]
Alkali-thermostable endo-β-1,4-glucanase *(PersiCel4)*	Sheep rumen microbiome	✓	✓	✓	✓	✓	✓	[129]
Cellulase/Hemicellulase	Soil microbiome	✓	✓	✓	✓	✓	–	[130]
Alkalophilic, thermophilic carboxylesterase	Soil microbiome	✓	✓	✓	✓	–	–	[131]
Carboxylesterase	Soil microbiome	✓	✓	✓	✓	–	–	[132]
Carboxylesterase	Compost microbiome	✓	✓	✓	✓	–	✓	[133]
Carboxylesterase	Sediment microbiome	✓	✓	✓	✓	–	–	[33]
Thermostable bifunctional cellulase/xylanase *(PersiCelXyn1)*	Cattle rumen microbiome	✓	–	✓	✓	✓	✓	[134]
Glucose and ethanol tolerant β-Glucosidase	Hot spring microbiome	✓	✓	✓	✓	✓	✓	[135]
β-glucosidase, α-L-arabinofuranosidase, β-xylosidase, and endo-1,4-β-xylanase	Porcupine microbiome	✓	✓	✓	✓	✓	–	[136]
Homologue of human α-glucosidase *(PersiAlpha-GL1)*	In vitro gastrointestinal digestion	✓	–	✓	✓	✓	✓	[137]
α-amylase *(PersiAmy1)*	Sheep rumen microbiome	✓	–	✓	✓	✓	✓	[138]

Table 2 (continued)

Novel metagenomic enzyme	Environment	Computational multi-stage pipeline						
		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**Refs.**
Acidic-thermostable α-amylase *(PersiAmy2)*	Sheep rumen microbiome	✓	–	✓	✓	✓	✓	[139]
Acidic-Thermostable *α*-amylase *(PersiAmy3)*	Sheep rumen microbiome	✓	✓	✓	✓	✓	✓	[139]
Cold-active pullulanase	Hot spring microbiome	✓	✓	✓	✓	–	✓	[140]
Thermostable pullulanase *(PersiPul1)*	Cattle rumen microbiome	✓	–	✓	✓	✓	✓	[141]
Laccase	Soil microbiome	✓	–	✓	✓	–	–	[142]
Stable laccase *(PersiLac1)*	Tannery wastewater microbiome	✓	–	✓	✓	✓	✓	[143]
Thermo-halotolerant laccase *(PersiLac2)*	Tannery wastewater microbiome	✓	–	✓	✓	✓	✓	[144]
Protease	Solid tannery waste microbiome	✓	✓	✓	✓	–	–	[145]
Protease	Solid tannery waste microbiome	✓	✓	✓	✓	–	✓	[146]
Thermo-halo-alkali-stable protease *(PersiProtease1)*	Tannery wastewater microbiome	✓	–	✓	✓	✓	✓	[147]
Feruloyl esterase	Soil microbiome	✓	✓	✓	✓	–	–	[148]
Solvent-tolerant esterase	Compost microbiome	✓	✓	✓	✓	–	✓	[35]
Esterase	Wastewater sediments microbiome	✓	✓	✓	✓	–	–	[149]
Esterase	Soil microbiome	✓	✓	✓	✓	–	✓	[150]
Lipid hydrolyzing enzyme	Hot spring microbiome	✓	–	✓	✓	–	✓	[151]
Tyrosine Phosphatase	Soil microbiome	✓	✓	✓	✓	✓	✓	[152]
PETase	Environmental metagenome	✓	✓	✓	✓	–	✓	[153]
PETase	Human saliva microbiome	✓	✓	✓	✓	✓	–	[154]
β-galactosidase	Marine microbiome	✓	✓	✓	✓	–	–	[155]
β -Glucuronidase	Mouse gut microbiome	✓	✓	✓	✓	–	–	[156]
β -Glucanase	Soil microbiome	✓	✓	✓	✓	–	✓	[157]
β -Glucanase	Vermicompost	✓	✓	✓	✓	–	–	[158]
ferulic acid esterase, α-L-arabinofuranosidase, GH10 β-D-1,4-xylanase	Wastewater treatment sludge	✓	–	–	–	–	–	[159]

The table includes information about the enzyme family, metagenome source, and the *in-silico* analyses conducted during the bioprospecting processes. Detailed analysis steps are outlined below

S1: BLAST alignment of metagenome sequences against a curated list of experimentally validated enzymes with desired properties obtained from a literature review. Selection of the most similar sequences, determined by their E-value and alignment score, for further refinement

S2: Analysis of the selected sequences to determine their phylogenetic positions among the related sequences obtained from the literature search. Focus on closely related sequences after removing distant relatives

S3: Determining the frequency and position of important amino acids in the candidate metagenome sequences. Assessment of statistical compatibility of these amino acids with literature and experimentally characterized enzymes possessing desired properties. This stage requires in-depth and comprehensive review of literature

S4: Comparing the candidate sequences for their active sites and other key amino acids with enzymes possessing the desired properties

S5: At this stage, the presence of conserved domains in the candidate enzymes is confirmed by utilizing tools such as CDD [141], Position Specific Scoring Matrices (PSSMs) or Hidden Markov Models (HMMs) or other motif modeling strategies

S6: Predicting the 3D structure of the candidate sequences and filtering for less related sequences

comprehensive suite of tools and databases for automated genome mining of a wide array of secondary metabolites. By combining genome mining for BGCs and chemical structure prediction for the encoded secondary metabolites, PRISM significantly improves the detection of genetically encoded nonribosomal peptides and polyketides [92]. While these tools facilitate the identification of genomic loci responsible for natural product biosynthesis, challenges arise in connecting these loci to the specific chemical structures of the encoded products [96]. Genomic analysis has revealed that bacterial genomes house numerous orphan BGCs, which are clusters not yet associated with the natural products they encode. There are also numerous examples of isolated natural products that have not been linked to their corresponding BGCs [97]. ML approaches have shown potential in genome mining for natural biological products, predicting the structure of natural products, and inferring biological activity from BGCs or the chemical structure of the respective secondary metabolite. Recently Prihoda et al. [98] showed that ML can be used in several steps to find bioactive natural products in genome sequences, including genome annotations, feature representation, BGC detection, structure prediction, and activity profiling. Another study developed a comprehensive ML method to predict the structures and biological activity of secondary metabolites from microbial genome sequences [99]. This approach can be used to predict the structures of natural products encoded by orphan BGCs.

In light of widespread metagenomic explorations of diverse microbial niches, huge amounts of genomic data are now at our fingertips. This genetic bounty holds immense potential for bioprospecting, offering novel microbial secondary metabolites, with a spectrum of promising medical and biotechnological applications. Numerous attempts have been made to explore metagenome data for novel natural products. In a study by Nayfach et al. [9], over 100,000 BGCs were predicted in 52,515 metagenome-assembled genomes, which were cataloged from diverse microbial communities representing the Earth's microbiome. This antiSMASH-based BGC discovery yielded up to 54 times more BGCs than manually curated entries in the MIBiG dataset, highlighting a vast reservoir of unexplored microbial natural products. In a comprehensive computational and experimental study, a probabilistic algorithm named MetaBGC was developed and applied to identify potential BGCs in complex metagenomic sequences from various regions of the human microbiome (gut, mouth, skin, and vagina) [100]. Out of the 13 BGCs encoding type II polyketides, two were successfully cloned and expressed in a heterologous system, revealing their potent antibacterial activities against gut microbes and suggesting a potential role in microbial interactions within the gut environment. These findings underscore the urgent need for the development of advanced tools and pipelines for targeted mining of metagenomes for novel, game-changing microbial secondary metabolites with biotechnological and medicinal potential.

4 Future directions

An extensive literature review highlights that functional screening is, in fact, a major source of currently characterized enzymes from environmental samples. However, there are instances where the integration of FBS and SBS methods has proven to be successful. For example, in a study on the pre-screening of clone libraries using functional screening followed by insert sequencing, a remarkable 106-fold increase in the success rate was achieved in identifying genes encoding desired enzymes compared to direct sequencing approaches [101]. Both FBS and SBS methods offer distinct advantages and disadvantages. SBS approaches may have limitations in terms of sequencing cost and errors. Furthermore, uncertainty in functional annotations and their limitations in discovering novel enzymes pose challenges to their widespread applications. FBS approaches can be used to identify novel enzymes and facilitate the direct determination of gene functions. However, the FBS methods also suffer from higher costs, the lack of effective screening methods for certain enzyme activities, and the challenges associated with heterologous expression systems.

In the past decade, significant improvements have been made in the computational modeling of 3D structures of proteins. These advancements have made it possible to take advantage of protein structure modeling in screening for novel enzymes from metagenomic sequences. Structural modeling can be used to evaluate enzymes for substrate specificity, enantioselectivity, metal ion specificity, pH and temperature dependence, as well as stability and secondary catalytic function.

In the era of a rapid expansion in enzyme-related biological databases as repositories for genome sequences, enzymes, tertiary structures, active sites, as well as metabolic pathways and reactions, there is an increased demand for the development of functional and computational screening tools. It is evident that the integration of SBS and FBS methods, coupled with the utilization of structural modeling, paves the way toward efficient exploration of novel enzymes from high throughput metagenomic data. This combination of approaches presents a promising roadmap for effective enzyme and natural product mining in the future.

Acknowledgements
Funding was provided by the Agricultural Biotechnology Research Institute of Iran (ABRII), Swedish Research Council (Vetenskapsrådet grant no.: 2017-05019), and the BBSRC Institute Strategic Programme Gut Microbes and Health (BB/r012490/1, its constituent project BBS/ e/F/000Pr10355).

Author Contribution
SA, KK, and GHS conceptualized and designed the review. SA, JG, MFS, and FFA conducted the literature review, analyzed the data, and drafted the manuscript. J-LH, X-Z H, FH, MB, KK, GHS critically reviewed and edited the manuscript. All authors contributed to the intellectual content and approved the final version for publication.

Declarations

Competing interests
There is no potential conflict of interest to disclose.

Received: 25 October 2023 Accepted: 19 December 2023
Published online: 11 January 2024

References

1. Gurung N, Ray S, Bose S, Rai V. A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. Biomed Res Int. 2013;2013: 329121. https://doi.org/10.1155/2013/329121.
2. Guazzaroni ME, Beloqui A, Vieites JM, Al-ramahi Y, Cortés NL, Ghazi A, et al. Metagenomic mining of enzyme diversity. Handbook of hydrocarbon and lipid microbiology. 2010. p. 2911–27.
3. Liu X, Kokare C. Microbial enzymes of use in industry. Biotechnology of microbial enzymes. 2017. p. 267–98.
4. Singh RS, Singh T, Pandey A. microbial enzymes—an overview. Advances in Enzyme Technology. 2019. p. 1–40.
5. Lammle K, Zipper H, Breuer M, Hauer B, Buta C, Brunner H, et al. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J Biotechnol. 2007;127(4):575–92. https://doi.org/10.1016/j.jbiotec.2006.07.036.
6. Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl Environ Microbiol. 1990;56(6):1919–25. https://doi.org/10.1128/aem.56.6.1919-1925.1990.
7. Glogauer A, Martini VP, Faoro H, Couto GH, Muller-Santos M, Monteiro RA, et al. Identification and characterization of a new true lipase isolated through metagenomic approach. Microb Cell Fact. 2011;10(1):54. https://doi.org/10.1186/1475-2859-10-54.
8. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Corrigendum: shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(12):1211. https://doi.org/10.1038/nbt1217-1211b.
9. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2021;39(4):499–509. https://doi.org/10.1038/s41587-020-0718-6.
10. Berini F, Casciello C, Marcone GL, Marinelli F. Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett. 2017. https://doi.org/10.1093/femsle/fnx211.
11. Itoh N. Metagenomics for improved biocatalysis. Future directions in biocatalysis. 2017. p. 375–84.
12. Colin PY, Kintses B, Gielen F, Miton CM, Fischer G, Mohamed MF, et al. Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. Nat Commun. 2015;6(1):10008. https://doi.org/10.1038/ncomms10008.
13. Arnold FH. Combinatorial and computational challenges for biocatalyst design. Nature. 2001;409(6817):253–7. https://doi.org/10.1038/35051731.
14. Robinson SL, Piel J, Sunagawa S. A roadmap for metagenomic enzyme discovery. Nat Prod Rep. 2021;38(11):1994–2023. https://doi.org/10.1039/d1np00006c.
15. Hou Q, Pucci F, Pan F, Xue F, Rooman M, Feng Q. Using metagenomic data to boost protein structure prediction and discovery. Comput Struct Biotechnol J. 2022;20:434–42. https://doi.org/10.1016/j.csbj.2021.12.030.
16. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2019;47(D1):D542–9. https://doi.org/10.1093/nar/gky1048.
17. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS ONE. 2009;4(2): e4345. https://doi.org/10.1371/journal.pone.0004345.
18. Lapebie P, Lombard V, Drula E, Terrapon N, Henrissat B. Bacteroidetes use thousands of enzyme combinations to break down glycans. Nat Commun. 2019;10(1):2043. https://doi.org/10.1038/s41467-019-10068-5.
19. Gharechahi J, Vahidi MF, Sharifi G, Ariaeenejad S, Ding XZ, Han JL, et al. Lignocellulose degradation by rumen bacterial communities: new insights from metagenome analyses. Environ Res. 2023;229: 115925. https://doi.org/10.1016/j.envres.2023.115925.
20. Ngara TR, Zhang H. Recent advances in function-based metagenomic screening. Genom Proteom Bioinform. 2018;16(6):405–15. https://doi.org/10.1016/j.gpb.2018.01.002.
21. Patel T, Chaudhari HG, Prajapati V, Patel S, Mehta V, Soni N. A brief account on enzyme mining using metagenomic approach. Front Syst Biol. 2022. https://doi.org/10.3389/fsysb.2022.1046230.
22. Sampaio PS, Fernandes P. Machine learning: a suitable method for biocatalysis. Catalysts. 2023. https://doi.org/10.3390/catal13060961.
23. Scherlach K, Hertweck C. Mining and unearthing hidden biosynthetic potential. Nat Commun. 2021. https://doi.org/10.1038/s41467-021-24133-5.
24. Zaparucha A, de Berardinis V, Vaxelaire-Vergne C. Chapter 1. Genome Mining for Enzyme Discovery. Modern Biocatalysis. Catalysis Series, 2018. p. 1–27.
25. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol. 1985;39(1):321–46. https://doi.org/10.1146/annurev.mi.39.100185.001541.
26. Uchiyama T, Miyazaki K. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol. 2009;20(6):616–22. https://doi.org/10.1016/j.copbio.2009.09.010.
27. Daniel R. The soil metagenome–a rich resource for the discovery of novel natural products. Curr Opin Biotechnol. 2004;15(3):199–204. https://doi.org/10.1016/j.copbio.2004.04.005.
28. Yun J, Ryu S. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. Microb Cell Fact. 2005;4(1):8. https://doi.org/10.1186/1475-2859-4-8.
29. Madhavan A, Sindhu R, Parameswaran B, Sukumaran RK, Pandey A. Metagenome analysis: a powerful tool for enzyme bioprospecting. Appl Biochem Biotechnol. 2017;183(2):636–51. https://doi.org/10.1007/s12010-017-2568-3.
30. Dadheech T, Shah R, Pandit R, Hinsu A, Chauhan PS, Jakhesara S, et al. Cloning, molecular modeling and characterization of acidic cellulase from buffalo rumen and its applicability in saccharification of lignocellulosic biomass. Int J Biol Macromol. 2018;113:73–81. https://doi.org/10.1016/j.ijbiomac.2018.02.100.
31. De Santi C, Altermark B, Pierechod MM, Ambrosino L, de Pascale D, Willassen NP. Characterization of a cold-active and salt tolerant esterase identified by functional screening of Arctic metagenomic libraries. BMC Biochem. 2016;17(1):1. https://doi.org/10.1186/s12858-016-0057-x.
32. Pereira MR, Maester TC, Mercaldi GF, de Macedo Lemos EG, Hyvonen M, Balan A. From a metagenomic source to a high-resolution structure of a novel alkaline esterase. Appl Microbiol Biotechnol. 2017;101(12):4935–49. https://doi.org/10.1007/s00253-017-8226-4.
33. Araujo FJ, Hissa DC, Silva GO, Antunes A, Nogueira VLR, Goncalves LRB, et al. A novel bacterial carboxylesterase identified in a metagenome derived-clone from Brazilian mangrove sediments. Mol Biol Rep. 2020;47(5):3919–28. https://doi.org/10.1007/s11033-020-05484-6.
34. Istvan P, Souza AA, Garay AV, Dos Santos DFK, de Oliveira GM, Santana RH, et al. Structural and functional characterization of a novel lipolytic enzyme from a Brazilian Cerrado soil metagenomic library. Biotechnol Lett. 2018;40(9–10):1395–406. https://doi.org/10.1007/s10529-018-2598-0.
35. Park JM, Kang CH, Won SM, Oh KH, Yoon JH. Characterization of a novel moderately thermophilic solvent-tolerant esterase isolated from a compost metagenome library. Front Microbiol. 2019;10:3069. https://doi.org/10.3389/fmicb.2019.03069.
36. Maruthamuthu M, van Elsas JD. Molecular cloning, expression, and characterization of four novel thermo-alkaliphilic enzymes retrieved from a metagenomic library. Biotechnol Biofuels. 2017;10(1):142. https://doi.org/10.1186/s13068-017-0808-y.
37. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012;2(1):3. https://doi.org/10.1186/2042-5783-2-3.
38. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017;355(6322):294–8. https://doi.org/10.1126/science.aah4043.
39. Wang Y, Shi Q, Yang P, Zhang C, Mortuza SM, Xue Z, et al. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. Genome Biol. 2019;20(1):229. https://doi.org/10.1186/s13059-019-1823-z.
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

41. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176.
42. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1. https://doi.org/10.1093/bioinformatics/btq461.
43. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63. https://doi.org/10.1093/bioinformatics/14.9.755.
44. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389.
45. Szalkai B, Grolmusz V. MetaHMM: a webserver for identifying novel genes with specified functions in metagenomic samples. Genomics. 2019;111(4):883–5. https://doi.org/10.1016/j.ygeno.2018.05.016.
46. Koutsandreas T, Ladoukakis E, Pilalis E, Zarafeta D, Kolisis FN, Skretas G, et al. ANASTASIA: an automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data. Front Genet. 2019;10:469. https://doi.org/10.3389/fgene.2019.00469.
47. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotechnol. 2009;27(2):157–67. https://doi.org/10.1038/nbt1519.
48. Elbehery AH, Leak DJ, Siam R. Novel thermostable antibiotic resistance enzymes from the Atlantis II Deep Red Sea brine pool. Microb Biotechnol. 2017;10(1):189–202. https://doi.org/10.1111/1751-7915.12468.
49. Garg R, Srivastava R, Brahma V, Verma L, Karthikeyan S, Sahni G. Biochemical and structural characterization of a novel halotolerant cellulase from soil metagenome. Sci Rep. 2016;6(1):39634. https://doi.org/10.1038/srep39634.
50. Al-Shahib A, Breitling R, Gilbert DR. Predicting protein function by machine learning on amino acid sequences–a critical evaluation. BMC Genomics. 2007;8(1):78. https://doi.org/10.1186/1471-2164-8-78.
51. Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins. 2020;88(3):397–413. https://doi.org/10.1002/prot.25832.
52. Zou Z, Tian S, Gao X, Li Y. mlDEEPre: multi-functional enzyme function prediction with hierarchical multi-label deep learning. Front Genet. 2018;9:714. https://doi.org/10.3389/fgene.2018.00714.
53. Shen HB, Chou KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun. 2007;364(1):53–9. https://doi.org/10.1016/j.bbrc.2007.09.098.
54. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, et al. SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PLoS ONE. 2016;11(8): e0155290. https://doi.org/10.1371/journal.pone.0155290.
55. Li Y, Wang S, Umarov R, Xie B, Fan M, Li L, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics. 2018;34(5):760–9. https://doi.org/10.1093/bioinformatics/btx680.
56. Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Dogan T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics. 2018;19(1):334. https://doi.org/10.1186/s12859-018-2368-y.
57. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. Science. 2023;379(6639):1358–63. https://doi.org/10.1126/science.adf2465.
58. Shi Z, Deng R, Yuan Q, Mao Z, Wang R, Li H, et al. Enzyme commission number prediction and benchmarking with hierarchical dual-core multitask learning framework. Research. 2023;6:0153. https://doi.org/10.34133/research.0153.
59. Buton N. Datasets and models for EnzBert. Zenodo; 2023.
60. Foroozandeh Shahraki M, Farhadyar K, Kavousi K, Azarabad MH, Boroomand A, Ariaeenejad S, et al. A generalized machine-learning aided method for targeted identification of industrial enzymes from metagenome: a xylanase temperature dependence case study. Biotechnol Bioeng. 2021;118(2):759–69. https://doi.org/10.1002/bit.27608.
61. Foroozandeh Shahraki M, Ariaeenejad S, Fallah Atanaki F, Zolfaghari B, Koshiba T, Kavousi K, et al. MCIC: automated identification of cellulases from metagenomic data and characterization based on temperature and pH dependence. Front Microbiol. 2020;11: 567863. https://doi.org/10.3389/fmicb.2020.567863.
62. Shahraki MF, Atanaki FF, Ariaeenejad S, Ghaffari MR, Norouzi-Beirami MH, Maleki M, et al. A computational learning paradigm to targeted discovery of biocatalysts from metagenomic data: a case study of lipase identification. Biotechnol Bioeng. 2022;119(4):1115–28. https://doi.org/10.1002/bit.28037.
63. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins. 2008;71(2):891–902. https://doi.org/10.1002/prot.21770.
64. Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. Protein Sci. 2002;11(2):350–60. https://doi.org/10.1110/ps.18602.
65. Littlechild JA. Protein structure and function. Introduction to biological and small molecule drug research and development. 2013. p. 57–79.
66. Petrey D, Honig B. Protein structure prediction: inroads to biology. Mol Cell. 2005;20(6):811–9. https://doi.org/10.1016/j.molcel.2005.12.005.
67. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins. 1995;23(3):356–69. https://doi.org/10.1002/prot.340230309.
68. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: an overview of protein structure prediction. Front Bioinform. 2023;3:1120370. https://doi.org/10.3389/fbinf.2023.1120370.
69. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12(1):7–8. https://doi.org/10.1038/nmeth.3213.
70. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–58. https://doi.org/10.1038/nprot.2015.053.
71. Webb B, Sali A. Comparative protein structure modeling using modeller. Curr Protoc Bioinformatics. 2016;54(1):561–5637. https://doi.org/10.1002/cpbi.3.
72. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42(Web Server issue):W252–8. https://doi.org/10.1093/nar/gku340.
73. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2.
74. Baltzer L, Nilsson H, Nilsson J. De novo design of proteins–what are the rules? Chem Rev. 2001;101(10):3153–63. https://doi.org/10.1021/cr0000473.
75. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun. 1995;91(1–3):43–56. https://doi.org/10.1016/0010-4655(95)00042-e.
76. Nelson MT, Humphrey W, Gursoy A, Dalke A, Kalé LV, Skeel RD, et al. NAMD: a parallel, object-oriented molecular dynamics program. Int J High Perform Comput Appl. 2016;10(4):251–68. https://doi.org/10.1177/109434209601000401.
77. Seritan S, Bannwarth C, Fales BS, Hohenstein EG, Isborn CM, Kokkila-Schumacher SIL, et al. TeraChem: a graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics. WIREs Comput Mol Sci. 2020. https://doi.org/10.1002/wcms.1494.
78. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins. 1999;36(3):340–6.
79. Ban X, Lahiri P, Dhoble AS, Li D, Gu Z, Li C, et al. Evolutionary stability of salt bridges hints its contribution to stability of proteins. Comput Struct Biotechnol J. 2019;17:895–903. https://doi.org/10.1016/j.csbj.2019.06.022.
80. Ahmed MC, Papaleo E, Lindorff-Larsen K. How well do force fields capture the strength of salt bridges in proteins? PeerJ. 2018;6: e4967. https://doi.org/10.7717/peerj.4967.
81. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011;12(1):151. https://doi.org/10.1186/1471-2105-12-151.
82. Pucci F, Kwasigroch JM, Rooman M. SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. Bioinformatics. 2017;33(21):3415–22. https://doi.org/10.1093/bioinformatics/btx417.

83. Hubbard RE, Kamran Haider M. Hydrogen bonds in proteins: role and strength. eLS. 2010.
84. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol. 1994;238(5):777–93. https://doi.org/10.1006/jmbi.1994.1334.
85. odinger L. The PyMOL molecular graphics system, version 2.0 Schrödinger, LLC. 2015.
86. Li Y, Roy A, Zhang Y. HAAD: A quick algorithm for accurate prediction of hydrogen atoms in protein structures. PLoS ONE. 2009;4(8): e6701. https://doi.org/10.1371/journal.pone.0006701.
87. Salam NK, Adzhigirey M, Sherman W, Pearlman DA. Structure-based approach to the prediction of disulfide bonds in proteins. Protein Eng Des Sel. 2014;27(10):365–74. https://doi.org/10.1093/protein/gzu017.
88. Katz L, Baltz RH. Natural product discovery: past, present, and future. J Ind Microbiol Biotechnol. 2016;43(2–3):155–76. https://doi.org/10.1007/s10295-015-1723-5.
89. Scott TA, Piel J. The hidden enzymology of bacterial natural product biosynthesis. Nat Rev Chem. 2019;3(7):404–25. https://doi.org/10.1038/s41570-019-0107-1.
90. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum information about a biosynthetic gene cluster. Nat Chem Biol. 2015;11(9):625–31. https://doi.org/10.1038/nchembio.1890.
91. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023;51(W1):W46–50. https://doi.org/10.1093/nar/gkad344.
92. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017;45(W1):W49–54. https://doi.org/10.1093/nar/gkx320.
93. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol. 2009;140(1–2):13–7. https://doi.org/10.1016/j.jbiotec.2009.01.007.
94. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 2009;10:185. https://doi.org/10.1186/1471-2105-10-185.
95. Behsaz B, Bode E, Gurevich A, Shi YN, Grundmann F, Acharya D, et al. Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. Nat Commun. 2021;12(1):3225. https://doi.org/10.1038/s41467-021-23502-4.
96. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 2020;11(1):6058. https://doi.org/10.1038/s41467-020-19986-1.
97. Jensen PR. Natural products and the gene cluster revolution. Trends Microbiol. 2016;24(12):968–77. https://doi.org/10.1016/j.tim.2016.07.006.
98. Prihoda D, Maritz JM, Klempir O, Dzamba D, Woelk CH, Hazuda DJ, et al. The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. Nat Prod Rep. 2021;38(6):1100–8. https://doi.org/10.1039/d0np00055h.
99. Walker AS, Clardy J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J Chem Inf Model. 2021;61(6):2560–71. https://doi.org/10.1021/acs.jcim.0c01304.
100. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, et al. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science. 2019. https://doi.org/10.1126/science.aax9176.
101. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, et al. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. Genome Res. 2010;20(11):1605–12. https://doi.org/10.1101/gr.108332.110.
102. Buton N, Coste F, Le Cunff Y, Valencia A. Predicting enzymatic function of protein sequences with attention. Bioinformatics. 2023. https://doi.org/10.1093/bioinformatics/btad620.
103. Dodda SR, Hossain M, Kapoor BS, Dasgupta S, Aikat K, et al. Computational approach for identification, characterization, three-dimensional structure modelling and machine learning-based thermostability prediction of xylanases from the genome of *Aspergillus fumigatus*. Comput Biol Chem. 2021;91:107451. https://doi.org/10.1016/j.compbiolchem.2021.107451.
104. Wang L, Wang Y, Chang S, Gao Z, Ma J, Wu B, et al. Identification and characterization of a thermostable GH11 xylanase from Paenibacillus campinasensis NTU-11 and the distinct roles of its carbohydrate-binding domain and linker sequence. Colloids Surf B Biointerfaces. 2022;209(Pt 1): 112167. https://doi.org/10.1016/j.colsurfb.2021.112167.
105. Ghadikolaei KK, Sangachini ED, Vahdatirad V, Noghabi KA, Zahiri HS. An extreme halophilic xylanase from camel rumen metagenome with elevated catalytic activity in high salt concentrations. AMB Express. 2019;9(1):86. https://doi.org/10.1186/s13568-019-0809-2.
106. Joshi N, Sharma M, Singh SP. Characterization of a novel xylanase from an extreme temperature hot spring metagenome for xylooligosaccharide production. Appl Microbiol Biotechnol. 2020;104(11):4889–901. https://doi.org/10.1007/s00253-020-10562-7.
107. Mon ML, Marrero Díaz de Villegas R, Campos E, Soria MA, Talia PM. Characterization of a novel GH10 alkali-thermostable xylanase from a termite microbiome. Bioresour Bioprocess. 2022. https://doi.org/10.1186/s40643-022-00572-w.
108. Fredriksen L, Stokke R, Jensen MS, Westereng B, Jameson JK, Steen IH, et al. Discovery of a Thermostable GH10 Xylanase with Broad Substrate Specificity from the Arctic Mid-Ocean Ridge Vent System. Appl Environ Microbiol. 2019. https://doi.org/10.1128/AEM.02970-18.
109. Knapik K, Becerra M, Gonzalez-Siso MI. Microbial diversity analysis and screening for novel xylanase enzymes from the sediment of the Lobios Hot Spring in Spain. Sci Rep. 2019;9(1):11195. https://doi.org/10.1038/s41598-019-47637-z.
110. Rajabi M, Nourisanami F, Ghadikolaei KK, Changizian M, Noghabi KA, Zahiri HS. Metagenomic psychrohalophilic xylanase from camel rumen investigated for bioethanol production from wheat bran using Bacillus subtilis AP. Sci Rep. 2022;12(1):8152. https://doi.org/10.1038/s41598-022-11412-4.
111. Hu D, Zhao X. Characterization of a new xylanase found in the rumen metagenome and its effects on the hydrolysis of wheat. J Agric Food Chem. 2022;70(21):6493–502. https://doi.org/10.1021/acs.jafc.2c00827.
112. Wang J, Liang J, Li Y, Tian L, Wei Y. Characterization of efficient xylanases from industrial-scale pulp and paper wastewater treatment microbiota. AMB Express. 2021;11(1):19. https://doi.org/10.1186/s13568-020-01178-1.
113. Ariaeenejad S, Hosseini E, Maleki M, Kavousi K, Moosavi-Movahedi AA, Salekdeh GH. Identification and characterization of a novel thermostable xylanase from camel rumen metagenome. Int J Biol Macromol. 2019;126:1295–302. https://doi.org/10.1016/j.ijbiomac.2018.12.041.
114. Ariaeenejad S, Maleki M, Hosseini E, Kavousi K, Moosavi-Movahedi AA, Salekdeh GH. Mining of camel rumen metagenome to identify novel alkali-thermostable xylanase capable of enhancing the recalcitrant lignocellulosic biomass conversion. Bioresour Technol. 2019;281:343–50. https://doi.org/10.1016/j.biortech.2019.02.059.
115. Ariaeenejad S, Lanjanian H, Motamedi E, Kavousi K, Moosavi-Movahedi AA, Hosseini SG. The stabilizing mechanism of immobilized metagenomic xylanases on bio-based hydrogels to improve utilization performance: computational and functional perspectives. Bioconjug Chem. 2020;31(9):2158–71. https://doi.org/10.1021/acs.bioconjchem.0c00361.
116. Mousavi SH, Sadeghian Motahar SF, Salami M, Kavousi K, Sheykh Abdollahzadeh Mamaghani A, Ariaeenejad S, et al. Invitro bioprocessing of corn as poultry feed additive by the influence of carbohydrate hydrolyzing metagenome derived enzyme cocktail. Sci Rep. 2022;12(1):405. https://doi.org/10.1038/s41598-021-04103-z.
117. Ariaeenejad S, Kavousi K, Zolfaghari B, Roy S, Koshiba T, Hosseini SG. Efficient bioconversion of lignocellulosic waste by a novel computationally screened hyperthermostable enzyme from a specialized microbiota. Ecotoxicol Environ Saf. 2023;252: 114587. https://doi.org/10.1016/j.ecoenv.2023.114587.
118. Pavarina GC, Lemos EGM, Lima NSM, Pizauro JM Jr. Characterization of a new bifunctional endo-1,4-beta-xylanase/esterase found in the rumen metagenome. Sci Rep. 2021;11(1):10440. https://doi.org/10.1038/s41598-021-89916-8.
119. Ariaeenejad S, Kavousi K, Maleki M, Motamedi E, Moosavi-Movahedi AA, Hosseini SG. Application of free and immobilized novel bifunctional biocatalyst in biotransformation of recalcitrant lignocellulosic biomass.

Chemosphere. 2021;285: 131412. https://doi.org/10.1016/j.chemosphere.2021.131412.
120. Ariaeenejad S, Motamedi E, Kavousi K, Ghasemitabesh R, Goudarzi R, Salekdeh GH, et al. Enhancing the ethanol production by exploiting a novel metagenomic-derived bifunctional xylanase/beta-glucosidase enzyme with improved beta-glucosidase activity by a nanocellulose carrier. Front Microbiol. 2022;13:1056364. https://doi.org/10.3389/fmicb.2022.1056364.
121. Sanjaya RE, Putri KDA, Kurniati A, Rohman A, Puspaningsih NNT. In silico characterization of the GH5-cellulase family from uncultured microorganisms: physicochemical and structural studies. J Genet Eng Biotechnol. 2021;19(1):143. https://doi.org/10.1186/s43141-021-00236-w.
122. Patel M, Patel HM, Dave S. Determination of bioethanol production potential from lignocellulosic biomass using novel Cel-5m isolated from cow rumen metagenome. Int J Biol Macromol. 2020;153:1099–106. https://doi.org/10.1016/j.ijbiomac.2019.10.240.
123. Stepnov AA, Fredriksen L, Steen IH, Stokke R, Eijsink VGH. Identification and characterization of a hyperthermophilic GH9 cellulase from the Arctic Mid-Ocean Ridge vent field. PLoS ONE. 2019;14(9): e0222216. https://doi.org/10.1371/journal.pone.0222216.
124. Hammami A, Fakhfakh N, Abdelhedi O, Nasri M, Bayoudh A. Proteolytic and amylolytic enzymes from a newly isolated *Bacillus mojavensis* SA: Characterization and applications as laundry detergent additive and in leather processing. Int J Biol Macromol. 2018;108:56–68. https://doi.org/10.1016/j.ijbiomac.2017.11.148.
125. Nguyen KHV, Dao TK, Nguyen HD, Nguyen KH, Nguyen TQ, Nguyen TT, et al. Some characters of bacterial cellulases in goats' rumen elucidated by metagenomic DNA analysis and the role of fibronectin 3 module for endoglucanase function. Anim Biosci. 2021;34(5):867–79. https://doi.org/10.5713/ajas.20.0115.
126. Guerrero EB, de Villegas RMD, Soria MA, Santangelo MP, Campos E, Talia PM. Characterization of two GH5 endoglucanases from termite microbiome using synthetic metagenomics. Appl Microbiol Biotechnol. 2020;104(19):8351–66. https://doi.org/10.1007/s00253-020-10831-5.
127. Maleki M, Shahraki MF, Kavousi K, Ariaeenejad S, Hosseini SG. A novel thermostable cellulase cocktail enhances lignocellulosic bioconversion and biorefining in a broad range of pH. Int J Biol Macromol. 2020;154:349–60. https://doi.org/10.1016/j.ijbiomac.2020.03.100.
128. Motamedi E, Sadeghian Motahar SF, Maleki M, Kavousi K, Ariaeenejad S, Moosavi-Movahedi AA, et al. Upgrading the enzymatic hydrolysis of lignocellulosic biomass by immobilization of metagenome-derived novel halotolerant cellulase on the carboxymethyl cellulose-based hydrogel. Cellulose. 2021;28(6):3485–503. https://doi.org/10.1007/s10570-021-03727-8.
129. Ariaeenejad S, Sheykh Abdollahzadeh Mamaghani A, Maleki M, Kavousi K, Foroozandeh Shahraki M, Hosseini Salekdeh G. A novel high performance in-silico screened metagenome-derived alkali-thermostable endo-beta-1,4-glucanase for lignocellulosic biomass hydrolysis in the harsh conditions. BMC Biotechnol. 2020;20(1):56. doi: https://doi.org/10.1186/s12896-020-00647-6.
130. Chai S, Zhang X, Jia Z, Xu X, Zhang Y, Wang S, et al. Identification and characterization of a novel bifunctional cellulase/hemicellulase from a soil metagenomic library. Appl Microbiol Biotechnol. 2020;104(17):7563–72. https://doi.org/10.1007/s00253-020-10766-x.
131. Yan Z, Ding L, Zou D, Wang L, Tan Y, Guo S, et al. Identification and characterization of a novel carboxylesterase EstQ7 from a soil metagenomic library. Arch Microbiol. 2021;203(7):4113–25. https://doi.org/10.1007/s00203-021-02398-0.
132. Zhang Y, Ding L, Yan Z, Zhou D, Jiang J, Qiu J, et al. Identification and characterization of a novel carboxylesterase belonging to family VIII with promiscuous acyltransferase activity toward cyanidin-3-O-glucoside from a soil metagenomic library. Appl Biochem Biotechnol. 2023;195(4):2432–50. https://doi.org/10.1007/s12010-021-03614-9.
133. Lu M, Daniel R. A novel carboxylesterase derived from a compost metagenome exhibiting high stability and activity towards high salinity. Genes (Basel). 2021. https://doi.org/10.3390/genes12010122.
134. Ariaeenejad S, Kavousi K, Mamaghani ASA, Motahar SFS, Nedaei H, Salekdeh GH. In-silico discovery of bifunctional enzymes with enhanced lignocellulose hydrolysis from microbiota big data. Int J Biol Macromol. 2021;177:211–20. https://doi.org/10.1016/j.ijbiomac.2021.02.014.
135. Kaushal G, Rai AK, Singh SP. A novel beta-glucosidase from a hot-spring metagenome shows elevated thermal stability and tolerance to glucose and ethanol. Enzyme Microb Technol. 2021;145: 109764. https://doi.org/10.1016/j.enzmictec.2021.109764.
136. Thornbury M, Sicheri J, Slaine P, Getz LJ, Finlayson-Trick E, Cook J, et al. Characterization of novel lignocellulose-degrading enzymes from the porcupine microbiome using synthetic metagenomics. PLoS ONE. 2019;14(1): e0209221. https://doi.org/10.1371/journal.pone.0209221.
137. Salami M, Sadeghian Motahar SF, Ariaeenejad S, Sheykh Abdollahzadeh Mamaghani A, Kavousi K, Moosavi-Movahedi AA, et al. The novel homologue of the human alpha-glucosidase inhibited by the non-germinated and germinated quinoa protein hydrolysates after in vitro gastrointestinal digestion. J Food Biochem. 2022;46(1):e14030. https://doi.org/10.1111/jfbc.14030.
138. Ariaeenejad S, Zolfaghari B, Sadeghian Motahar SF, Kavousi K, Maleki M, Roy S, et al. Highly efficient computationally derived novel metagenome alpha-amylase with robust stability under extreme denaturing conditions. Front Microbiol. 2021;12: 713125. https://doi.org/10.3389/fmicb.2021.713125.
139. Sadeghian Motahar SF, Ariaeenejad S, Salami M, Emam-Djomeh Z, Sheykh Abdollahzadeh Mamaghani A. Improving the quality of gluten-free bread by a novel acidic thermostable alpha-amylase from metagenomics data. Food Chem. 2021;352:129307. https://doi.org/10.1016/j.foodchem.2021.129307.
140. Thakur M, Sharma N, Rai AK, Singh SP. A novel cold-active type I pullulanase from a hot-spring metagenome for effective debranching and production of resistant starch. Bioresour Technol. 2021;320(Pt A): 124288. https://doi.org/10.1016/j.biortech.2020.124288.
141. Sadeghian Motahar SF, Salami M, Ariaeenejad S, Emam-Djomeh Z, Sheykh Abdollahzadeh Mamaghani A, Kavousi K, et al. Synergistic Effect of metagenome-derived starch-degrading enzymes on quality of functional bread with antioxidant activity. Starch Stärke. 2021; doi: https://doi.org/10.1002/star.202100098.
142. Itoh N, Hayashi Y, Honda S, Yamamoto Y, Tanaka D, Toda H. Construction and characterization of a functional chimeric laccase from metagenomes suitable as a biocatalyst. AMB Express. 2021;11(1):90. https://doi.org/10.1186/s13568-021-01248-y.
143. Ariaeenejad S, Kavousi K, Afshar Jahanshahi D, Sheykh Abdollahzadeh Mamaghani A, Ghasemitabesh R, Moosavi-Movahedi AA, et al. Enzymatically triggered delignification through a novel stable laccase: a mixed in-silico /in-vitro exploration of a complex environmental microbiota. Int J Biol Macromol. 2022;211:328–41. https://doi.org/10.1016/j.ijbiomac.2022.05.039.
144. Motamedi E, Kavousi K, Sadeghian Motahar SF, Reza Ghaffari M, Sheykh Abdollahzadeh Mamaghani A, Hosseini Salekdeh G, et al. Efficient removal of various textile dyes from wastewater by novel thermo-halotolerant laccase. Bioresour Technol. 2021;337:125468. https://doi.org/10.1016/j.biortech.2021.125468.
145. Verma SK, Sharma PC. Isolation and biochemical characterization of a novel serine protease identified from solid tannery waste metagenome. Biologia. 2021;76(10):3163–74. https://doi.org/10.1007/s11756-021-00832-8.
146. Verma SK, Kaur S, Tevetia A, Chatterjee S, Sharma PC. Structural characterization and functional annotation of microbial proteases mined from solid tannery waste metagenome. Biologia. 2021;76(6):1829–42. https://doi.org/10.1007/s11756-021-00727-8.
147. Ariaeenejad S, Kavousi K, Mamaghani ASA, Ghasemitabesh R, Hosseini SG. Simultaneous hydrolysis of various protein-rich industrial wastes by a naturally evolved protease from tannery wastewater microbiota. Sci Total Environ. 2022;815: 152796. https://doi.org/10.1016/j.scitotenv.2021.152796.
148. Wu S, Nan F, Jiang J, Qiu J, Zhang Y, Qiao B, et al. Molecular cloning, expression and characterization of a novel feruloyl esterase from a soil metagenomic library with phthalate-degrading activity. Biotechnol Lett. 2019;41(8–9):995–1006. https://doi.org/10.1007/s10529-019-02693-3.
149. Cavello IA, Hours RA, Cavalitto SF. Enzymatic hydrolysis of gelatin layers of X-Ray films and release of silver particles using keratinolytic serine proteases from *Purpureocillium lilacinum* LPS # 876. J Microbiol Biotechnol. 2013;23(8):1133–9. https://doi.org/10.4014/jmb.1302.02038.

150. Sarkar J, Dutta A, Pal Chowdhury P, Chakraborty J, Dutta TK. Characterization of a novel family VIII esterase EstM2 from soil metagenome capable of hydrolyzing estrogenic phthalates. Microb Cell Fact. 2020;19(1):77. https://doi.org/10.1186/s12934-020-01336-x.
151. Kaur R, Kumar R, Verma S, Kumar A, Rajesh C, Sharma PK. Structural and functional insights about unique extremophilic bacterial lipolytic enzyme from metagenome source. Int J Biol Macromol. 2020;152:593–604. https://doi.org/10.1016/j.ijbiomac.2020.02.210.
152. Castillo Villamizar GA, Nacke H, Griese L, Tabernero L, Funkner K, Daniel R. Characteristics of the first protein tyrosine phosphatase with phytase activity from a soil metagenome. Genes (Basel). 2019. https://doi.org/10.3390/genes10020101.
153. Karunatillaka I, Jaroszewski L, Godzik A. Novel putative polyethylene terephthalate (PET) plastic degrading enzymes from the environmental metagenome. Proteins. 2022;90(2):504–11. https://doi.org/10.1002/prot.26245.
154. Goncalves TA, Sodre V, da Silva SN, Vilela N, Tomazetto G, Araujo JN, et al. Applying biochemical and structural characterization of hydroxycinnamate catabolic enzymes from soil metagenome for lignin valorization strategies. Appl Microbiol Biotechnol. 2022;106(7):2503–16. https://doi.org/10.1007/s00253-022-11885-3.
155. Sun J, Yao C, Li Y, Wang W, Hao J, Yu Y. A novel salt-tolerant GH42 beta-galactosidase with transglycosylation activity from deep-sea metagenome. World J Microbiol Biotechnol. 2022;38(9):154. https://doi.org/10.1007/s11274-022-03348-8.
156. Creekmore BC, Gray JH, Walton WG, Biernat KA, Little MS, Xu Y, et al. Mouse gut microbiome-encoded beta-glucuronidases identified using metagenome analysis guided by protein structure. mSystems. 2019. https://doi.org/10.1128/mSystems.00452-19.
157. Wierzbicka-Wos A, Henneberger R, Batista-Garcia RA, Martinez-Avila L, Jackson SA, Kennedy J, et al. Biochemical characterization of a novel monospecific endo-beta-1,4-glucanase belonging to GH family 5 from a rhizosphere metagenomic library. Front Microbiol. 2019;10:1342. https://doi.org/10.3389/fmicb.2019.01342.
158. Yasir M, Khan H, Azam SS, Telke A, Kim SW, Chung YR. Cloning and functional characterization of endo-beta-1,4-glucanase gene from metagenomic library of vermicompost. J Microbiol. 2013;51(3):329–35. https://doi.org/10.1007/s12275-013-2697-5.
159. Holck J, Djajadi DT, Brask J, Pilgaard B, Krogh K, Meyer AS, et al. Novel xylanolytic triple domain enzyme targeted at feruloylated arabinoxylan degradation. Enzyme Microb Technol. 2019;129: 109353. https://doi.org/10.1016/j.enzmictec.2019.05.010.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.