Assessment of enzyme diversity in the fermented food microbiome

Peng Li[1, #], Jingyu Sun[1, #], Yu Geng[1], Yiru Jiang[1], Yue-zhong Li[1, *], Zheng Zhang[1, *]

[1] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China

[#] These authors contributed equally: Peng Li, Jingyu Sun.

[*] Address correspondence to Yue-zhong Li (E-mail: lilab@sdu.edu.cn, ORCID: 0000-0001-8336-6638) or Zheng Zhang (E-mail: zhangzheng@sdu.edu.cn, ORCID: 0000-0001-9971-6006)

Summary

Microbial bioactivity is essential for the flavor, appearance, quality, and safety of fermented foods. However, the diversity and distribution of enzymatic resources in fermentation remain poorly understood. This study explored 10,202 metagenome-assembled genomes from global fermented foods using artificial intelligence, identifying over 5 million enzyme sequences grouped into 98,693 homologous clusters, representing over 3,000 enzyme types. Functional analysis revealed that 84.4% of these clusters were unannotated in current databases, with high novelty in terpenoid and polyketide metabolism enzymes. Peptide-bound hydrolases have the potential to adapt to a wider range of environmental conditions, as predicted by the optimal temperature and pH of the enzyme cluster. We calculated the niche breadth on the basis of the distribution of enzymes in different food types, and 31.3% of the enzyme clusters were food type specific. Additionally, we developed a machine learning model to classify fermented food sources by enzyme clusters, highlighting key enzymes differentiating habitats. Our findings emphasize the untapped potential of fermented food environments for enzyme resource exploration, offering valuable insights into microbial functions for future food research.

Keywords: Food microbiome; Food; Enzymes; Machine learning; Metagenomics; Metagenome-assembled genomes; Large-scale microbiome analysis; Contrastive learning.

Introduction

Fermentation was originally employed as a means of food preservation. Over the centuries, it has remained a vital component of the human diet, owing to the distinctive textures and flavors generated by the natural activities of microorganisms and enzymes.[1,2] Fermented foods are recently defined as foods or beverages produced through the growth of specific microorganisms and the enzymatic transformation of food components.[3] The microorganisms found in fermented foods come from the food matrix, the production environment, or are introduced by humans. These microorganisms produce enzymes, volatile compounds, and antimicrobial substances—such as organic acids, hydrogen peroxide, and bacteriocins—that help slow or prevent spoilage and inhibit the growth of harmful pathogens.[4] In addition, fermented foods are rich in beneficial microorganisms and complex metabolites that can positively affect human health through dietary intake.[5-7] In terms of food–microbe interactions, thousands of fermented foods and beverages exist globally.[8] Fermented products from various food matrices (e.g., dairy, vegetable, and meat) exhibit diverse microbiota compositions and functions, and the composition of the microbiota in a specific fermented food type may vary over time and space,[9] thus affecting the quality of the product.[10,11] Therefore, characterizing the potential traits and functions of microbial communities associated with fermented foods has been the focus of extensive research. Here, the functional core of fermentation lies in the metabolic activity of microbial enzymes. During in situ fermentation, endogenous enzymes play an indispensable catalytic role: hydrolyzing proteins, polysaccharides and lipids to improve the bioavailability of nutrients[12,13] and producing volatile substances that impart flavor.[14] For example, certain proteases from *Lactobacillus* can catalyze the production of high concentrations of various peptides and amino acids during grain fermentation, which have demonstrated potential antioxidant, antihypertensive, or cancer-preventive properties *in vitro* and in animal models.[15,16] In addition, certain enzymes in fermented foods, such as serine fibrinolytic enzyme (nattokinase), which possesses cardiovascular health-related properties, including antithrombotic, anticoagulant and fibrinolytic properties, have been shown to provide potential health benefits.[5] This fully demonstrates that the diverse enzyme species encoded in the fermented food microbiome are the basis of its sensory characteristics, nutritional quality, health benefits and industrial value.

Despite the crucial roles played by enzymes, comprehensive exploration of enzyme diversity in the fermented food microbiome has lagged far behind taxonomic characterization. While high-throughput sequencing has revolutionized our understanding of the complex taxonomic diversity in fermented foods, the vast majority of research efforts have focused on cataloging the microbial species composition, while the full range and functional landscape of encoded enzymes remain largely unknown.[17,18] Recent advances have demonstrated the power of metagenomics beyond reference genomes: studies have revealed new species that influence fermented food properties such as flavor and color,[19] revealed a large amount of "functional dark matter" (uncharacterized protein families)[20] and identified the fermentation environment as an untapped reservoir of secondary metabolite biosynthesis-related gene clusters.[21] The application of omics technologies to the study of fermentation systems promises to provide new perspectives for the industrialization of traditional fermentation processes. Compared with culture-based methods, culture-independent sequencing-based methods enable more rapid analysis of microbial ecology and link it to changes in metabolites formed during fermentation.[22] However, the analysis of core enzymatic substrates is still lacking, which limits our ability to rationally design fermentation processes or fully explore the biocatalytic potential of these special sources of enzyme resources.

In this study, we collected metagenome-assembled genomes representing typical fermented food types from around the globe and systematically mined enzymes using machine learning. The diversity of enzyme resources in fermented food microbiomes is comprehensively presented by elucidating the distribution of enzymes across various food matrices and microorganisms and differences in functional composition.

Results

A catalog of 98,693 enzyme sequence clusters from microbial fermented foods

We collected metagenome-assembled genomes (MAGs) from fermented foods and analyzed the composition and distribution of enzymes within these food microbiomes. We first predicted all open reading frames (ORFs) in 10,202 MAGs (from 2,101 fermented food samples; **Figure S1A**), yielding a total of 30,924,933 sequences. In addition, we selected 276,345 functionally annotated sequences with known Enzyme

Commission (EC) numbers from public databases as reference sequences. All sequences were clustered using an 80% amino acid identity threshold and 80% sequence coverage. This consistency threshold was determined through iterative clustering of reference sequences on the basis of intracluster functional consistency (**Figure S1B**). Following quality control, which requires clusters to contain ≥ 10 members and representative sequences to exceed 100 amino acids in length, we obtained 472,428 protein clusters (**Figure 1A**). To identify potential enzymes, we annotated representative sequences of protein clusters with EC numbers via the CLEAN tool (comparative learning-based enzyme annotation algorithm). Among these, 98,693 protein clusters received high-confidence annotations (CLEAN confidence score ≥ 0.2) and were designated enzyme clusters (**Data S1**).[23] These clusters represent collections of enzymes with identical functions, with prediction reliability quantitatively compared against that of two comparable tools (**Figures S1C and S1D**).

These enzyme clusters were assigned a total of 3,017 distinct EC numbers (**Data S2**). Among these annotated enzyme clusters (**Figure 1B**), the three most abundant enzyme classes by quantity and proportion were transferases (EC2; 34,476 clusters, 34.9%), hydrolases (EC3; 24,514 clusters, 24.8%), and oxidoreductases (EC1; 15,333 clusters, 15.5%). This distribution indicates that microorganisms in fermented foods predominantly encode these enzyme classes (**Figure 1B**). However, although ligases (EC6) and translocases (EC7) have relatively few overall enzyme clusters, their average number of enzyme clusters that can correspond to each EC number is much greater than that of other categories. These findings suggest that these sequences may serve as key resources for the discovery of potential isoenzymes (**Figure S2A**).

Using reference sequences from the clustering process (i.e., public database proteins with experimentally validated enzymatic activity), we assessed the novelty of the enzyme clusters identified in fermented food microbiomes at the sequence level. Specifically, we aligned all reference sequences against each enzyme cluster's member sequences via DIAMOND. An enzyme cluster was classified as known if ≥ 80% amino acid identity and ≥ 80% sequence coverage was achieved between any member and a reference sequence. The remaining clusters were designated novel (**Figure 1A**). Based on this classification, we assessed the novelty of all EC-number-annotated enzyme clusters (**Figure 1C**).

A total of 83,254 clusters were predicted as potential new enzymes, accounting for 84.4% of the total, indicating substantial unexplored enzymatic diversity in food microbiomes. Significant differences in cluster size were observed between known and novel clusters (Wilcoxon rank-sum test, $P < 0.001$), with novel clusters typically containing fewer members than known clusters (**Figures S2B and S2C**). Analysis of novelty rates across enzyme classes revealed consistently high values (73.0–89.0%; **Figure 1C**), with hydrolases (EC3) exhibiting the highest novelty rate (89.0%), making them primary contributors to novel enzyme diversity. This was followed by translocases (EC7, 87.1%) and oxidoreductases (EC1, 86.9%).

Taxonomic distribution of enzymes in the fermented food microbiome

By tracing the genomic information of the enzyme clusters, we combined the taxonomy and sequence correspondence of the fermented food microbiome MAGs to determine the taxonomic distribution of enzymes in the fermented food microbiome. We first compared enzyme cluster diversity between bacterial and fungal MAGs. The diversity of both known and novel enzyme clusters originated predominantly from bacteria (**Figure S3A**). Fungal MAGs harbored significantly more enzyme clusters per genome than bacterial MAGs did (median: 1,401 vs. 499 clusters; Wilcoxon rank-sum test, $P < 0.001$; **Figure S3B**). However, analysis of the ORF content revealed that bacterial genomes encoded significantly more enzyme clusters per 100 ORFs than did fungal genomes (median: 24.3 vs. 16.2; $P < 0.001$), a pattern consistent with both known and novel clusters (**Figure 2A**). These findings collectively indicate that although fungal genomes are typically larger and harbor more enzyme-encoding genes, bacterial genomes allocate a higher proportion of ORFs to enzymatic functions, demonstrating greater coding density for enzyme production.

We examined the distribution of novelty rates and the diversity of enzyme clusters within genomes across different phylum levels. Diversity was concentrated in phyla with abundant MAGs: *Firmicutes*, *Proteobacteria*, *Actinobacteria*, and *Ascomycota*-core components of fermented food microbiomes (**Figures 2B and 2C**). Analysis of enzyme clusters per MAG revealed greater diversity and novelty proportions in fungal phyla than in bacterial phyla (**Figures 2D and 2E**). Among bacteria, *Proteobacteria* exhibited the highest enzyme cluster diversity, with a single MAG encoding up to 1,394 clusters. This was followed by *Actinobacteria* and *Firmicutes*. These phylum-level patterns systematically reveal taxonomic functional profiles and bioprospecting potential in fermented food microbiomes. Significant positive correlations emerged between MAG genome size (ORFs and base counts) and encoded enzyme clusters in both bacteria (Spearman's $\rho = 0.620$, $P < 0.001$ [ORFs]; $\rho = 0.658$, $P < 0.001$ [bases]) and fungi ($\rho = 0.250$, $P < 0.001$ [ORFs]; $\rho = 0.504$, $P < 0.001$ [bases]) (**Figures 2F and S3C**), confirming that larger genomes encode greater enzyme cluster diversity. Finally, known enzyme clusters originated from significantly more MAGs than novel clusters across most primary EC categories (all except lyases; $P < 0.05$), with broader taxonomic distributions (**Figures S3D and S3E**).

Functional diversity and enzymatic properties of enzymes in fermented food microbiomes

To elucidate the functional composition of microbial enzymes in fermented foods, we performed KEGG annotation on representative sequences of 98,693 enzyme clusters and mapped them to KEGG pathways. The distribution of enzyme clusters across pathways (allowing duplicate counting of the same cluster in multiple pathways; **Figures 3A and 3B**) revealed that at the first pathway level, the majority of the functionally annotated enzyme clusters were concentrated in metabolism (46,888; 47.5%), with a substantial proportion remaining unannotated (22,817; 23.1%). This observation aligns with the sensitivity of CLEAN toward understudied enzymes. Notably, 37,552 (80.1%) of the novel enzyme clusters belonged to metabolic pathways. Among the second-level KEGG pathways under metabolism, carbohydrate metabolism (11,041), amino acid metabolism (9,528), and metabolism of cofactors and vitamins (8,788) presented relatively high counts of enzyme clusters. Certain pathways, despite having lower absolute numbers of novel enzyme clusters, demonstrated high proportions of novelty within their respective pathways, including lipid metabolism (83.7%), biosynthesis of other secondary metabolites (82.8%), glycan biosynthesis and metabolism (82.6%), metabolism of other amino acids (82.4%), and metabolism of terpenoids and polyketides (82.2%) (**Figure 3B**). Furthermore, we examined the diversity of novel enzyme clusters within selected third-level metabolic pathways (**Figure S4A**). Particularly noteworthy was terpenoid backbone biosynthesis, where novel enzyme clusters contributed to more than half of the diversity within this second-level pathway. Overall, our mining results revealed abundant enzyme systems related to lipid and amino acid metabolism, which may serve as a foundation for improving nutritional components in fermented foods. Simultaneously, we identified a rich repertoire of novel enzyme systems associated with secondary metabolite metabolism in fermented food microbiomes, especially in terpenoid backbone biosynthesis, offering new resources for secondary metabolite exploration.

Physicochemical conditions critically regulate enzymatic activity, allowing enzymes adapted to extreme industrial temperatures and pH to be valuable developmental resources. To assess the biotechnological potential of fermented food-derived enzymes, we predicted the optimal pH for known and novel enzyme clusters via EpHod and the optimal temperature via Seq2pHopt (**Figure 3C**). The results revealed that the optimal pH and optimal temperature distributions for the known and novel enzyme clusters overlapped closely, with the optimal pH in the range of 7-8.5 and the optimal temperature in the range of 30-40°C. Furthermore, we analyzed the enzymatic conditions of several types of enzymes that play important roles in the food fermentation process, including starch and sugar degradation enzymes (α-amylase-EC 3.2.1.1, glucoamylase-EC 3.2.1.3, β-glucosidase-EC 3.2.1.21), protein degradation and modification enzymes (protease-EC 3.4.x.x, transglutaminase-EC 2.3.2.13), lipid metabolism enzymes (esterase-EC 3.1.1.x, lipase-EC 3.1.1.3), and peroxidase (EC 1.11.x.x). The results (**Figures 3D and 3E**) revealed that, with the exception of proteases and esterases, enzyme clusters associated with all other functions presented no significant differences (Wilcoxon rank-sum test, $P > 0.05$) in the distributions of optimal temperature and pH between novel and known enzymes. Notably, however, comparisons of β-glucosidases, proteases, esterases, and peroxidases consistently revealed that novel enzyme clusters appeared to harbor a certain number of outliers; these uncommon optimal temperatures or pH values could be explored as potential enzyme sources for extreme industrial environments. Furthermore, significant differences were observed in the predicted optimal pH between the overall novel and known enzyme clusters within each hydrolase category that cleave peptide bonds, glycosidic bonds, and ester bonds (**Figure S4B**, Wilcoxon rank-sum test, $P < 0.001$). For the optimal temperature, a significant difference was detected only between the two groups for peptide-bound hydrolases (**Figure S4C**), suggesting that this hydrolase class may possess broader adaptive potential to environmental conditions.

Distribution of enzymes in various fermented foods and machine learning classification

Significant differences exist in the composition of microbial communities between different fermented foods, which are grouped by food type (e.g., dairy, vegetables, and meat).[10] Even within the same type of fermented food, the microbial community may vary over time or space, thereby affecting the fermentation process or product quality.[11] Understanding the variable composition of species or functions in fermented food microbiomes has long been a key focus.[17] In our analysis, the MAGs of microorganisms in fermented foods were derived from 10 food categories and 80 specific food types. On the basis of the sample distribution data of 98,693 enzyme cluster member sequences, we calculated the sequence diversity of each enzyme within individual samples (**Data S3**). This diversity was normalized against the total number of enzyme cluster sequences within the respective sample to yield relative diversity. The samples were subsequently pooled by food type, and the distribution specificity of enzyme clusters across food types was quantified via Levins' niche breadth index.

On the basis of their niche breadth values, the enzyme clusters were categorized into five groups (ranging from narrow to broad niche distributions) to characterize their distribution breadth across food types (**Figure 4A**). Analysis of the distribution patterns of enzyme clusters among food categories revealed 38,723 clusters that occurred exclusively within a single food category. Among these, 30,926 clusters exhibited even narrower food type specificity (niche breadth = 1, calculated based on food type), accounting for 31.3% of the total enzyme clusters (**Figure S5A**). Notably, a significant difference in niche breadth was observed between novel and known enzyme clusters across different EC functional classes (Wilcoxon rank-sum test, P < 0.001). This manifested as novel enzyme clusters having significantly lower niche breadth than known clusters and being narrowly distributed across fewer food categories (**Figure 4B**). Furthermore, the proportions of narrowly distributed versus broadly distributed enzyme sequences appeared relatively conserved among enzymes of different functions (i.e., sequences grouped under different secondary EC numbers) (**Figure 4C**). Additionally, we compared the diversity and proportion of novel enzyme clusters across different fermented food categories (**Figures 5A and 5B**). Compared with other food categories, fermented beverages (food types, including water kefir, pulque, and lemon) presented relatively greater enzyme cluster diversity and a greater proportion of novel clusters. Fermented fish showed the lowest enzyme cluster diversity among all the food categories, whereas alcohol had the lowest proportion of novel clusters, suggesting that microbial enzymes associated with alcoholic fermentation have been extensively characterized.

Further analysis explored the associations between sample species diversity and the rates of encoded enzyme clusters across different fermented food substrate types. The results revealed that for most substrate types (excluding fermented fish), a significant positive correlation consistently existed between the within-sample number of MAGs and the diversity of enzyme clusters encoded by the entire microbial community (Spearman $\rho > 0$, $P < 0.001$). Furthermore, the rate of change in enzyme cluster diversity with increasing species diversity varied across environments (**Figure S5B**; Shannon index: **Figure S5C**). On the basis of the composition of enzyme sequences with specific EC numbers within the communities, we calculated a distance matrix between samples and performed principal coordinate analysis (**Figure 5C**). This revealed significant differences in enzyme cluster composition among communities from different environmental food categories (PERMANOVA, $P < 0.001$). Therefore, food category not only is a key determinant of species composition within food microbiomes but also significantly shapes both the composition and diversity of enzyme clusters within these communities.

The differential enzymes or functions present among communities from different food matrices may serve as important materials for optimizing and modifying fermented foods or exploring new functionalities. Like other studies identifying key genes in communities, we employed machine learning methods to construct a classifier for predicting food matrix types via a random forest model on the diversity of different enzyme clusters in the samples, aiming to identify key enzymes distinguishing various food matrices. We assessed the classifier's predictive ability for food matrix categories (**Figures S6A and S6B**), where the area under the ROC curve (AUC) for multiclass classification was 0.973 (hand-till method; test dataset), indicating that the model is relatively reliable. Next, we extracted the variable importance from the model (EC numbers, **Figures S6C and Data S4**), which represents the list of enzyme clusters that can differentiate food matrix types. We also evaluated the relative importance of different functions (EC number categories) for the overall model (**Figure S6C**) and found that transferases, hydrolases, and oxidoreductases were the primary types influencing food matrix classification. Furthermore, by correlating enzyme clusters with KEGG metabolic pathway information, we analyzed the importance of these three types of enzyme clusters across different pathways. The results indicated that carbon metabolism, amino acid metabolism, and cofactor and vitamin metabolism are key functions contributing to differences among communities (**Figures 5D and 5E**).

Discussion

Food fermentation results from the biological activity of microorganisms in food categories, and most research has focused on the taxonomic characterization of these microorganisms, particularly to reveal the contributions of specific taxa to the fermentation process.[17] With the use of the latest sequencing technologies and bioinformatics tools, we can recover genomes from the metagenome of the environment to reveal many taxa that cannot be found by current culture-based techniques, thereby revealing hidden taxa that cannot be captured by traditional culture techniques.[19,24] The rapid accumulation of massive amounts of metagenome sequencing data and MAG data enables the characterization of the functional composition of fermented food microbiomes.[18] With the aid of advanced machine learning methods, it is now possible to predict new functions and enzyme constants independently of sequence similarity.[23,25]

Based on microbial metagenome and MAG data from a diverse range of representative food matrices worldwide, we mined a large number of microbially encoded enzyme clusters from fermented foods and assessed their novelty for the first time. We identified a total of 98,693 enzyme clusters with potential functions (assigned EC numbers), of which novel clusters accounted for up to 84.4%. These findings suggest the potential for discovering new enzymes in the fermented food microbiome. We provide the species origin and functional characterization of all the enzyme clusters, with the highest proportion of novel enzyme clusters in the hydrolase class (EC3), which is a key process for the hydrolysis of compounds such as proteins that impact the flavor of the product.[11,26] These novel enzyme clusters are promising candidates for flavor enhancement and the diversification of fermented foods.

Novel enzymes are often associated with unique properties or functions, such as low-temperature activity, thermostability, pH adaptability, and other properties or functions, such as tolerance to high salinity, pressure, solvents, metal ions, and inhibitors. These novel enzymes are valuable in food bioprocessing.[27] In our results, the novel enzyme clusters as a whole have a high degree of overlap with known enzyme clusters in terms of their optimum temperature and pH distribution, and a certain number of outliers (points that deviate from the conventional optimum temperature and pH) were detected in several key functional enzyme categories (β-glucosidases, proteases, and peroxidases) that affect food fermentation. Among them, thermophilic β-glucosidases are particularly promising for transformation—recent studies have verified that novel β-glucosidases from thermophilic microorganisms can efficiently catalyze the synthesis of prebiotic trisaccharides at pH 6.5 and 75°C,[28] which provides a molecular template for the development of thermostable glycoside hydrolases. More significantly, the temperature and pH of the novel clusters in the peptide bond hydrolase class were significantly different from those of the known clusters. The expansion of these adaptive boundaries indicates that the fermented food microbiome contains untapped resources that break through the application limits of existing industrial enzymes and is particularly valuable for the development of efficient biocatalysts that require extreme conditions (such as high-temperature sterilization for soy sauce production[29] and acidic environment biocatalysis).[30] These findings not only reveal the industrialization potential of fermentation-derived extreme enzymes but also provide a sustainable path for the use of environmentally adaptive enzyme resources produced by the natural evolution of food-grade microbiomes, it may systematically replace the current chemical treatment processes that rely on high temperatures or strong acids and alkalis and provide a new generation of catalysts for biomanufacturing that are both efficient and ecologically safe.[27]

Environmental factors can substantially influence the microbial species composition of a given habitat. Indeed, the community structure of fermented foods is highly variable over time and space,[11,31] with raw materials and processing methods considered important factors driving the microbiota in food fermentation.[10] We show that the diversity of enzyme clusters encoded by communities varies between different food categories, with liquid matrices (fermented beverages such as water kefir) hosting particularly diverse enzyme clusters, which is consistent with previous knowledge that environmental factors dominate microbial communities.[10,32] Revealing the functional characteristics of the microbiome is a prerequisite for optimizing the various attributes of fermented foods, and a recent study revealed the habitat specificity of secondary metabolite biosynthesis potential in fermented foods.[21] In our work, the environmental distribution characteristics of enzyme clusters were demonstrated through the calculation of niche width, among which about 31.3% of the enzyme clusters were food type specific, and a certain specific enzyme cluster proportion was also maintained in each EC category enzyme cluster, reflecting the existence of many rarely distributed enzyme cluster resources with various functions in the fermented food environment. We also analyzed the diversity of enzyme clusters in different habitats and used machine learning to identify potential key functional enzyme clusters that differ between habitats. Our study helps elucidate the functional composition of fermentation communities and contributes to the understanding and exploitation of previously unused or underutilized properties and bioactivities of fermented foods.

The analyses presented in this work are based on MAGs reconstructed from metagenomic sequencing data to predict potential new enzymes. Although MAG-based approaches are effective tools for exploring the microbiome in fermented foods,[26,33] eukaryotic genome characterization from MAGs remains underestimated owing to multiple limitations of possible biases in the metagenome sampling process.[34] Furthermore, although the samples we used cover the main types of fermented foods, there are currently over 200 fermented foods worldwide, each with different origins and processing methods, which means that our analyses do not fully represent the entire fermentation environment. It can be inferred that the enzyme diversity and novelty rate of fermented foods as a whole may exceed the current observation range, indicating that there is still a wider enzyme resource library to explore in the fermentation environment.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zheng Zhang (E-mail: zhangzheng@sdu.edu.cn).

Materials availability

This study did not generate new materials.

Data and code availability

All sequenced genomes are available in the curatedFoodMetagenomicData (cFMD) database (https://zenodo.org/records/13285428) and UniProt database (https://www.uniprot.org/).

All original code has been deposited at Zenodo: https://zenodo.org/records/15665866.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China

(32270073).

AUTHOR CONTRIBUTIONS

P.L., J.S., and Z.Z. conceived and developed the study. Z.Z., J.S., and P.L. gathered the data and conducted the analyses. J.S., P.L., and Z.Z. led the writing of the manuscript. Y.G. and Y.J. contributed critically to the analyses and writing. Y.-Z.L. directed the study and critically revised the manuscript for important intellectual content.

DECLARATION OF INTERESTS

The authors declare no competing interests.

FIGURES

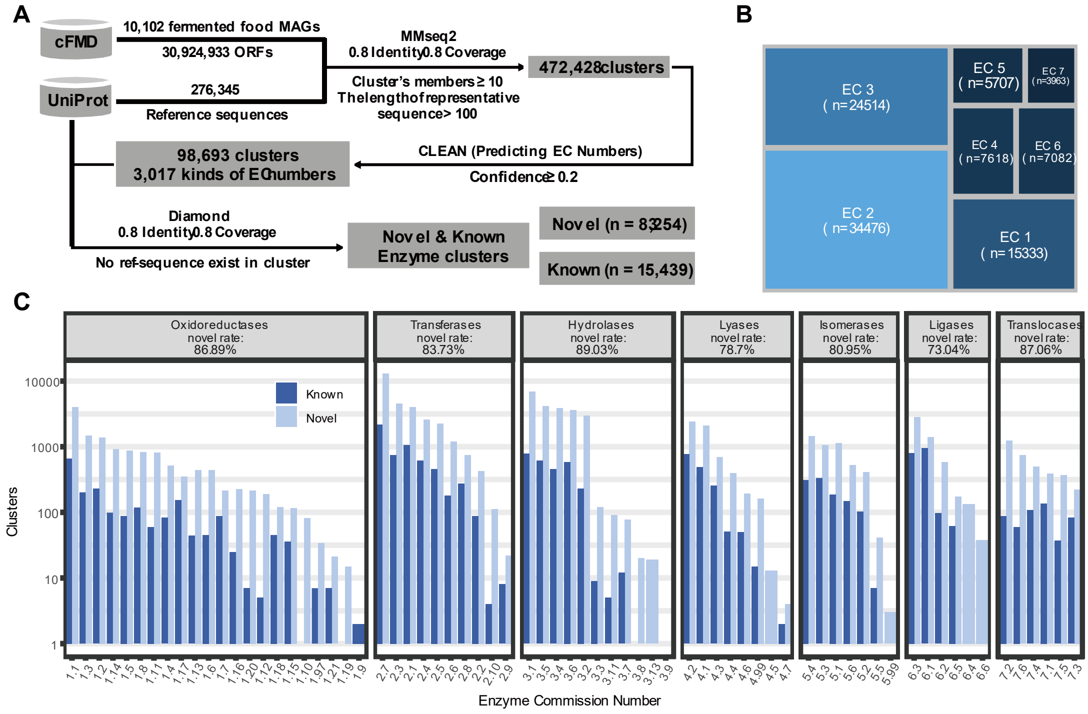

Figure 1: Catalog of enzymes in fermented food microbiomes.

(A) Overview of the workflow used to identify enzymes in the fermented food microbiome.

(B) Distribution of known versus novel enzyme clusters across major enzyme classes (n = number of enzyme clusters contained in the modified enzyme class).

(C) Distribution of known versus novel enzyme clusters across major enzyme classes. The Y axis is displayed in logarithmic values. The header values indicate the proportion of novel clusters within each primary EC category.

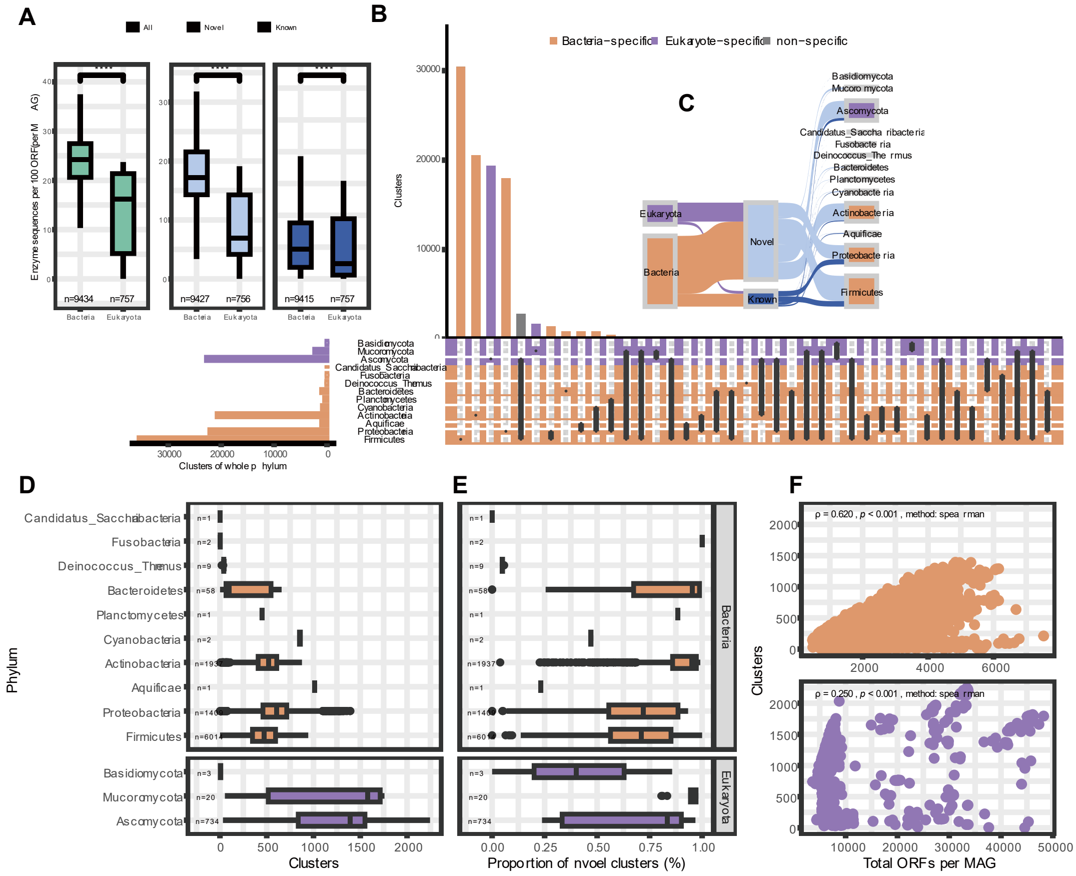

Figure 2: Taxonomic origins of enzymes in fermented food microbiomes.

(A) Enzyme coding density comparison between bacteria and eukaryote. Bar plots show the average number of enzyme sequences per 100 ORFs for novel versus known clusters (n = number of MAGs from bacteria or eukaryote). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, *P* > 0.05; * *P* ≤ 0.05; ** *P* ≤ 0.01; *** *P* ≤ 0.001; **** *P* ≤ 0.0001.

(B) Phylum-level distribution of all the enzyme clusters. The UpSet plot displays the composition and overlaps of 98,693 clusters across phyla.

(C) Sankey diagram depicting phylum-level origins of novel versus known enzyme clusters in bacterial and eukaryotic MAGs.

(D) Novelty rates of enzyme clusters across phyla (n = number of MAGs per phylum).

(E) Enzyme cluster diversity across phyla (n = number of MAGs per phylum).

(F) Correlation between genome size (total ORFs per MAG) and enzyme cluster diversity. The bacterial and eukaryotic MAGs are represented by orange and purple points, respectively.

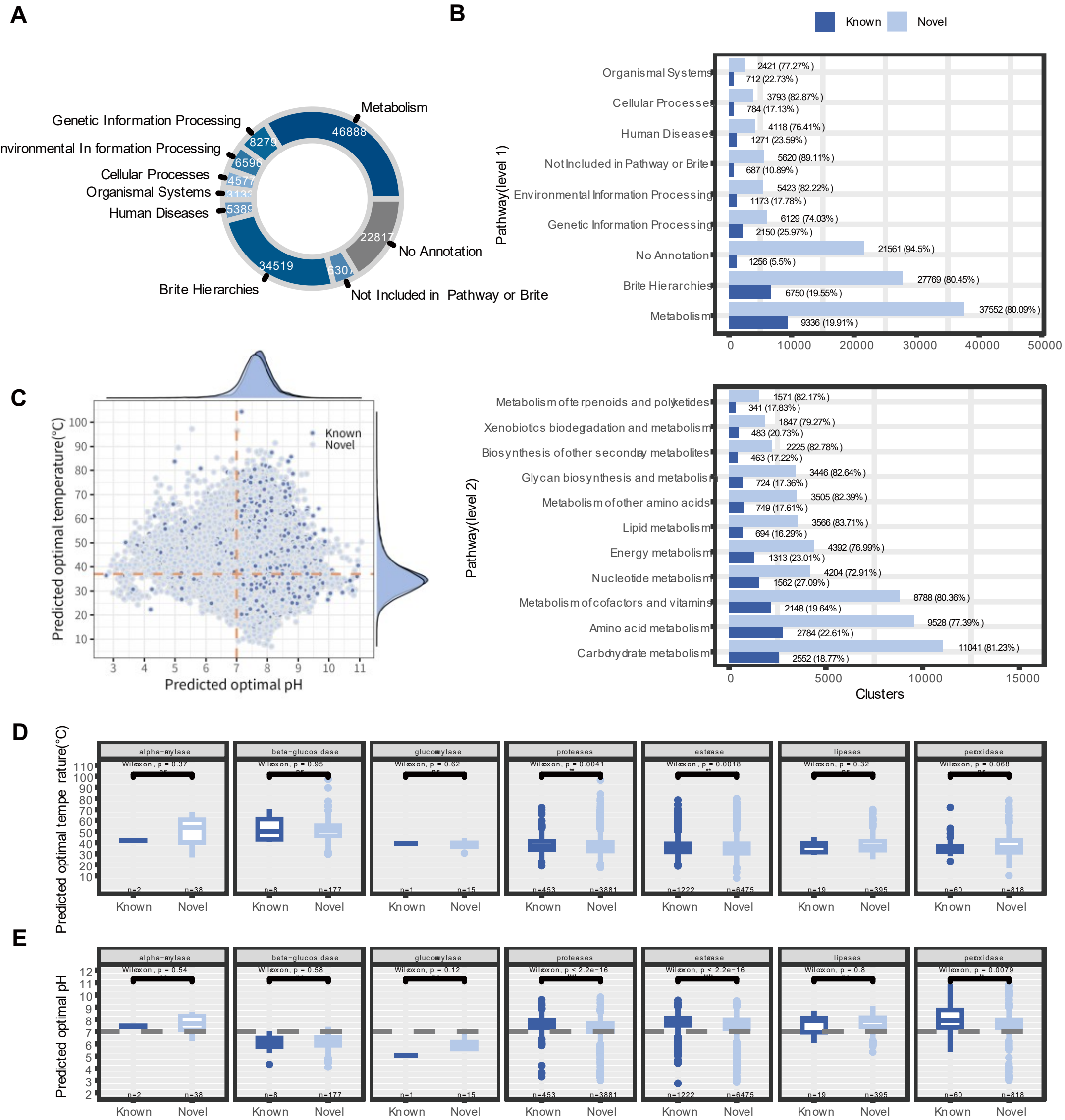

Figure 3: Functional diversity and enzymatic properties of fermented food microbiome enzymes.

(A) KEGG primary pathway annotation results for 98,693 enzyme clusters (n = number of enzyme clusters annotated to this pathway).

(B) The enzyme cluster diversity of the KEGG level 1 pathway and the level 2 pathway whose KEGG level 1 pathway is "Metabolism" enzyme cluster diversity of the cluster are shown respectively. Different shades of blue represent known enzyme clusters and novel enzyme clusters. The numerical value represents the number of enzyme clusters, and the percentage represents the proportion of enzyme clusters.

(C) The distribution of optimal temperature and optimal pH. The orange dotted line marks 37℃ and pH 7.0.

(D) Box plots characterizing the optimal temperatures of key industrial enzymes (n = number of enzyme clusters).

(E) Box plots characterizing the optimal pH of key industrial enzymes (n = number of enzyme clusters). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$.

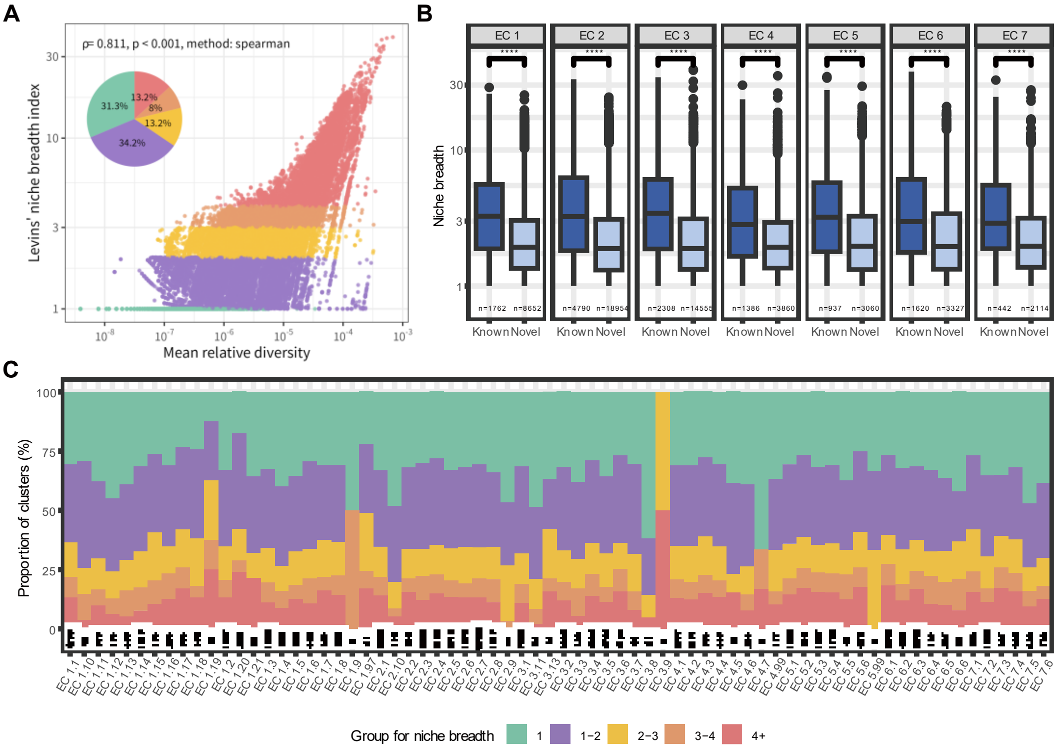

Figure 4: Distribution of microbial enzymes across food categories.

(A) Distribution of enzyme clusters between the average relative diversity of the samples and Levins' niche breadth.

(B) Distribution of the niche breadth of known or novel enzyme clusters classified by primary EC number (n = number of enzyme clusters). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$.

(C) Proportions of different niche breadth groups of enzyme clusters classified by secondary EC number.

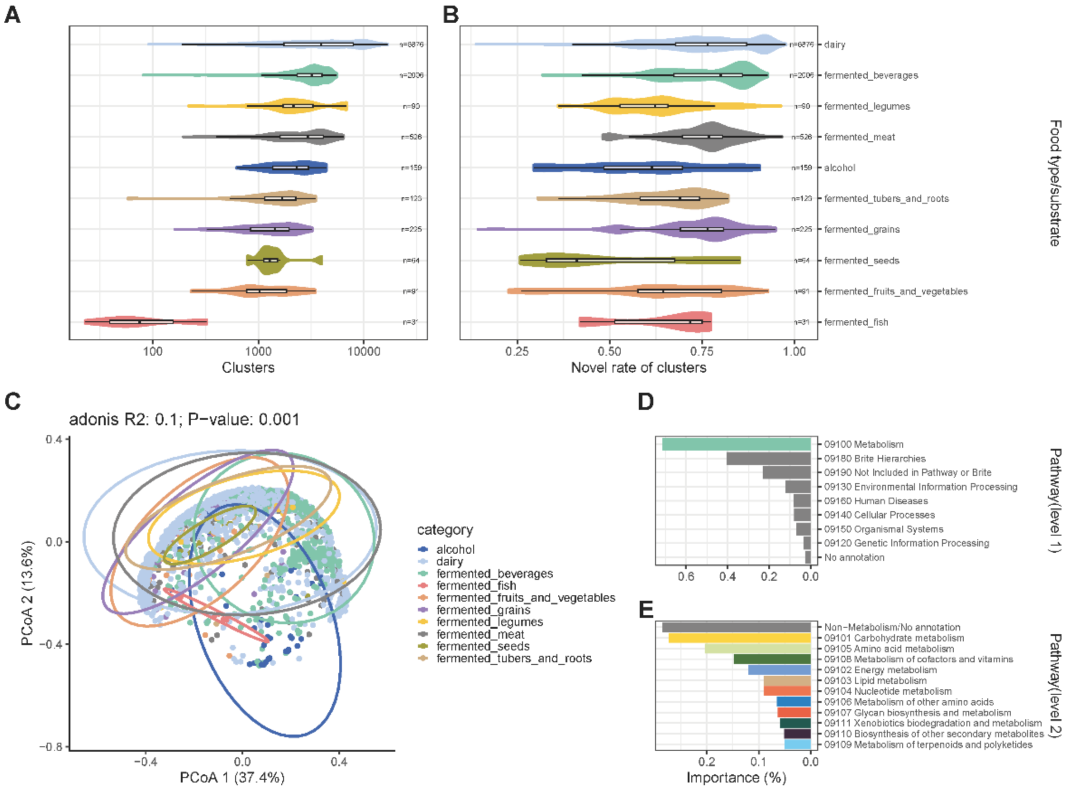

Figure 5: Distribution of microbial enzyme diversity across food categories and machine learning predictions in different fermented foods.

(A) Statistics of the number of enzyme clusters in different habitat types (n = number of samples in the habitat). The box plot represents the distribution of enzyme cluster diversity among different food categories.

(B) Statistics of enzyme cluster novelty across different habitat types (n = number of samples in the habitat). The box plot represents the distribution of enzyme cluster novelty among different food categories.

(C) Principal coordinate analysis based on the composition of enzyme clusters with different EC numbers in samples.

(D, E) Sum of feature importance of machine learning under different KEGG metabolic pathways.

STAR★Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
10,202 MAGs from fermented foods	(Carlino et al., 2024)[18]	https://github.com/SegataLab/cFMD/tree/main?tab=readme-ov-file
276,345 sequences with well-defined functions and EC numbers	(Bateman et al., 2025)[35]	https://www.uniprot.org/
Enzyme resources derived from fermented foods	This paper	https://zenodo.org/records/15665866
Software and algorithms		
CLEAN	(Yu et al., 2023)[23]	https://github.com/tttianhao/CLEAN
FEDKEA	(Zheng et al., 2024)[36]	https://github.com/Stevenleizheng/FEDKEA
ProTrek	(Su et al., 2024)[37]	https://github.com/westlake-repl/ProTrek
prodigal	(Hyatt et al., 2010)[38]	http://compbio.ornl.gov/prodigal/
TransDecoder	(Haas)[39]	https://github.com/TransDecoder/TransDecoder
MetaPhlAn	(Blanco et al., 2023) [40]	https://github.com/biobakery/MetaPhlAn
Kofamscan	(Aramaki et al., 2020)[41]	https://www.genome.jp/tools/kofamkoala/
MMseq2	(Steinegger et al., 2017)[42]	https://github.com/soedinglab/mmseqs2

diamond	(Buchfink et al., 2015)[43]	https://github.com/bbuchfink/diamond
EpHod	(Gado et al., 2025)[44]	https://github.com/beckham-lab/EpHod
Seq2pHopt	(Qiu et al., 2025)[45]	https://github.com/SizheQiu/Seq2Topt
R version 4.1.2 and 4.4.0	(R Core Team, 2013)[46]	https://www.r-project.org/
R Studio	(R Team, 2020)[47]	https://posit.co/
Other		
Zenodo repository with custom code and deposited data to reproduce analyses	This paper	https://zenodo.org/records/15665866

Method details

Data collection and processing

In this study, we processed and compiled metagenome-assembled genomes (MAGs), taxonomic annotations, and sample metadata of fermentation-associated microorganisms from the curatedFoodMetagenomicData (cFMD) database.[18] On the basis of the taxonomic classifications provided by the database, bacterial-origin MAGs were analyzed via Prodigal (v2.6.3) for open reading frames (ORFs) prediction,[38] whereas fungal-origin MAGs were processed via TransDecoder (v5.7.1) to identify ORFs.[39] Both tools were run with default parameters. We performed taxonomic annotation of the MAGs via MetaPhlAn (v4).[40]

For enzyme function annotation, all the predicted ORFs were clustered on the basis of sequences using MMseq2 (v15.6f452), and the clustering threshold was set to sequence identity ≥ 80% and coverage ≥ 80% to ensure strict consistency.[42] During the clustering process, 276,345 reviewed enzyme sequences with well-defined functional annotations and Enzyme Commission (EC) numbers from the UniProt/Swiss-Prot database were incorporated as internal references for coclustering.[35] We chose "80% identity + 80% coverage" as the clustering standard based on a multi-threshold preliminary clustering analysis of the UniProt enzyme sequence dataset. This analysis tested the clustering performance with identity thresholds ranging from 10% to 90%. After excluding multifunctional enzymes and sequences with unclear EC numbers, the results showed that when the identity threshold reached 80%, the resulting clusters were almost completely consistent in EC numbers. Compared with the 70% identity standard commonly used by most current enzyme function prediction tools,[23,48] the 80% threshold we set performed better in terms of functional consistency and provided a more reliable and homogeneous clustering basis (**Figure S1B**). With this strategy, any sequences with different EC numbers are not classified into the same cluster. Therefore, we can use the EC number of the representative sequence in each cluster as the basis for functional annotation of all members in the entire cluster.

Prediction of enzyme functions

CLEAN is a machine learning algorithm based on comparative learning that can perform enzyme function prediction on the basis of amino acid sequences of proteins with high accuracy, reliability, and sensitivity and can predict new enzymes by learning the embedding space of the enzyme, reacting the functional similarity with the Euclidean distance, and outputting a list of enzyme functions sorted by likelihood.[23] To reduce the effect of sequencing errors, we used the representative sequences of 472,428 clusters with more than 10 cluster members and representative sequence lengths of at least 100 amino acids for enzyme function prediction via CLEAN to obtain data on enzyme resources of microbial origin in fermented foods. We identified a total of 98,693 clusters, encompassing 3,017 distinct EC numbers, for downstream analyses. These clusters were selected based on representative sequence prediction results with medium to high confidence (confidence ≥ 0.2) and the criterion that at least one member within each cluster originated from fermented food sources. The annotation of the representative sequence was used to infer the EC numbers of all other members in each cluster. We used Kofamscan 1.3.0 to perform KEGG functional annotation of representative sequences of enzyme clusters on the basis of default parameters.[41]

To improve the accuracy of the enzyme function prediction results, we also used FEDKEA and ProTrek to predict these enzyme clusters.[36] [37] For FEDKEA, we utilized this enzyme function prediction method, which combines a pretrained protein language model with a distance-weighted k-nearest neighbor (k-NN) algorithm. In the screening results, predictions with a FirstProbability value ≥ 0.95 were identified as potential enzyme sequences. For ProTrek, we installed the basic environment of the software according to the guidelines (https://github.com/westlake-repl/ProTrek) and downloaded the pre-trained model weights and pre-computed faiss index (ProTrek_650M_UniRef50). We deployed the server locally and predicted the protein sequence by calling the API. Specific parameters: input = "protein aa sequences"; nprobe = 1000, opk = 5, input_type = "sequence", query_type = "text", subsection_type

= "Enzyme commission number", db = "Swiss-Prot", api_name = "/search". For the returned score, 15 and above are considered to be high-quality annotations.

Identification of known enzyme clusters

On the basis of the 276,345 enzyme sequences we obtained from Swiss-Prot, we used DIAMOND (parameters: --evalue 0.001 --sensitive --header-simple --max-target-seqs 1) to align the predicted enzyme sequences with known enzyme sequences.[43] If at least one sequence in an enzyme cluster matched a known enzyme sequence and the predicted and reference EC numbers were the same, it was classified as a "known cluster". All sequences within such a cluster were further designated "known sequences". The novelty rate of the enzyme cluster was calculated as the ratio of known clusters to total clusters in each sample.

Prediction of enzymatic properties

Given the extreme environments that may exist in fermentation systems (e.g., extreme acidity, alkalinity, or temperature), we predicted the enzymatic properties of 98,693 representative sequences that were identified as enzyme clusters. Specifically, we used EpHod (parameters: --verbose 1 --save_attention_weights 0 --save_embeddings 0) to calculate the optimal pH of the enzymes,[44] and the average of the results output by two machine learning models, support vector regression (SVR) and the transfer learning approach (ESM-1v–RLATtr), was used as the optimal pH predicted for each sequence, thereby reducing the bias of a single model and outputting more robust results. The model adjusts the training process through a sample-weighted loss function and uses a sample-weighted metric to evaluate the model, which mitigates the bias toward the general neutral pH value and ensures that a model is developed that is good at predicting extreme pH values, thereby being able to identify enzymes with high acid or alkalinity tolerance in fermented food systems. We calculated the optimal temperature of the enzymes using a deep learning model provided by Seq2pHopt (parameters: python code/seq2topt.py –query data.csv –output data_result.csv),[45] which improves prediction accuracy through PLM embedding of protein sequences, multihead attention, and residual dense neural networks.

Niche breadth calculation and principal coordinate analysis

We obtained food classification information from the cFMD database, which included a total of 10 food categories and 80 food types.[18] On the basis of the sample distribution data of 98,693 enzyme cluster member sequences, the sequence diversity of each enzyme cluster within individual samples was first calculated. This diversity was then normalized against the total number of enzyme cluster sequences within the sample to obtain the relative diversity. Subsequently, the sample data were pooled by food type, and Levins' niche breadth index was computed via the niche.width function from the ‘spaa’ package to quantify the distribution of enzyme clusters across food types.

To assess overall differences in enzyme functional profiles between samples and food types (beta diversity), we performed principal coordinates analysis (PCoA). The input data consisted of a Bray-Curtis distance matrix of nonredundant enzyme cluster species for each sample. PCoA was performed via the cmdscale function of the ‘stat’ package in R (with parameter eig = TRUE). The ‘ggplot2’ package was used for data visualization, and the variance explained is shown on the axis. PERMANOVA analysis (Permutational multivariate analysis of variance) was used to verify the significance of the PCoA grouping, and the adonis2 function of the ‘vegan’ package was used for calculation based on the Bray-Curtis distance (with parameters permutations = 999, method = "bray").

Development of the machine learning classifier

To construct a classifier capable of predicting strain taxonomy on the basis of key enzyme clusters across different food matrices, a random forest model was selected for data classification. The dataset was derived from the EC number-based species statistics of all the enzyme clusters in the sample where the MAG is located. Multiclassification random forest models were constructed for different food matrix types based on the framework of the R package ‘tidymodels’. On the basis of a 3:1 division of the training set test set, the preprocessing process (Recipe Steps) included the removal of zero variance variables, the removal of highly autocorrelated variables, and the resampling of unbalanced sets. Three replications of ten-fold cross-validation were used to train the model and adjust the parameters to minimize the likelihood of model overfitting. The final model was determined on ROC_AUC values, and the relevant evaluation parameters are presented in the Supplemental information. Random forest significance was used to estimate key enzyme clusters that differed between habitats.

Statistical analysis

Statistical analyses were performed in RStudio via R4.1.2 and 4.4.0. Radar plots were drawn using the ‘ggradar’ package (v.0.2), UpSet plots were drawn via the ‘UpSetR’ package (v.1.4.0), and box-and-line, bar, and scatter plots were drawn via the ‘ggplot2’ package. The ‘spaa’ package was used to calculate the Levins' niche breadth index, and the ‘stat’ package was used to perform PCoA analysis. Each box-and-line plot presents the data distribution in the following way: the box represents the interquartile spacing (IQR), and the horizontal line inside the box marks the median. The maximum and minimum values within 1.5 times the IQR from the edge of the box must be extended, and outliers beyond the required range are plotted separately. All other plotting codes involved and their related dependencies are described in the Code Availability section.

References

1. Gänzle, M.G., Monnin, L., Zheng, J., Zhang, L., Coton, M., Sicard, D., and Walter, J. (2024). Starter Culture Development and Innovation for Novel Fermented Foods. Annu. Rev. Food Sci. Technol. *15*, 211-239. https://doi.org/10.1146/annurev-food-072023-034207.
2. Gadaga, T.H., Mutukumira, A.N., Narvhus, J.A., and Feresu, S.B. (1999). A review of traditional fermented foods and beverages of Zimbabwe. Int. J. Food Microbiol. *53*, 1-11. https://doi.org/10.1016/s0168-1605(99)00154-3.
3. Marco, M.L., Sanders, M.E., Gänzle, M., Arrieta, M.C., Cotter, P.D., De Vuyst, L., Hill, C., Holzapfel, W., Lebeer, S., Merenstein, D., et al. (2021). The International

Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on fermented foods. Nat. Rev. Gastroenterol. Hepatol. *18*, 196-208. https://doi.org/10.1038/s41575-020-00390-5.

4. Reis, J.A., Paula, A.T., Casarotti, S.N., and Penna, A.L.B. (2012). Lactic acid bacteria antimicrobial compounds: characteristics and applications. Food Eng. Rev. *4*, 124-140. https://doi.org/10.1007/s12393-012-9051-2.
5. Mukherjee, A., Breselge, S., Dimidi, E., Marco, M.L., and Cotter, P.D. (2024). Fermented foods and gastrointestinal health: underlying mechanisms. Nat. Rev. Gastroenterol. Hepatol. *21*, 248-266. https://doi.org/10.1038/s41575-023-00869-x.
6. Kolodziejczyk, A.A., Zheng, D., and Elinav, E. (2019). Diet-microbiota interactions and personalized nutrition. Nat. Rev. Microbiol. *17*, 742-753. https://doi.org/10.1038/s41579-019-0256-8.
7. Wastyk, H.C., Fragiadakis, G.K., Perelman, D., Dahan, D., Merrill, B.D., Yu, F.B., Topf, M., Gonzalez, C.G., Van Treuren, W., Han, S., et al. (2021). Gut-microbiota-targeted diets modulate human immune status. Cell *184*, 4137-4153.e14. https://doi.org/10.1016/j.cell.2021.06.019.
8. Marco, M.L., Heeney, D., Binda, S., Cifelli, C.J., Cotter, P.D., Foligné, B., Gänzle, M., Kort, R., Pasin, G., Pihlanto, A., et al. (2017). Health benefits of fermented foods: microbiota and beyond. Curr. Opin. Biotechnol. *44*, 94-102. https://doi.org/10.1016/j.copbio.2016.11.010.
9. Blasche, S., Kim, Y., Mars, R.A.T., Machado, D., Maansson, M., Kafkia, E., Milanese, A., Zeller, G., Teusink, B., Nielsen, J., et al. (2021). Metabolic

cooperation and spatiotemporal niche partitioning in a kefir microbial community. Nat. Microbiol. *6*, 196-208. https://doi.org/10.1038/s41564-020-00816-5.

10. Leech, J., Cabrera-Rubio, R., Walsh, A.M., Macori, G., Walsh, C.J., Barton, W., Finnegan, L., Crispie, F., O'Sullivan, O., Claesson, M.J., and Cotter, P.D. (2020). Fermented-food metagenomics reveals substrate-associated differences in taxonomy and health-associated and antibiotic resistance determinants. mSystems *5*, 00522-20. https://doi.org/10.1128/mSystems.00522-20.

11. Walsh, A.M., Crispie, F., Kilcawley, K., O'Sullivan, O., O'Sullivan, M.G., Claesson, M.J., and Cotter, P.D. (2016). Microbial succession and flavor production in the fermented dairy beverage kefir. mSystems *1*, 00052-16. https://doi.org/10.1128/mSystems.00052-16.

12. Yao, G.Q., Yu, J., Hou, Q.C., Hui, W.Y., Liu, W.J., Kwok, L.Y., Menghe, B., Sun, T.S., Zhang, H.P., and Zhang, W.Y. (2017). A perspective study of koumiss microbiome by metagenomics analysis based on single-cell amplification technique. Front. Microbiol. *8*, 165. https://doi.org/10.3389/fmicb.2017.00165.

13. Escobar-Zepeda, A., Sanchez-Flores, A., and Quirasco Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiol. *57*, 116-127. https://doi.org/10.1016/j.fm.2016.02.004.

14. Gänzle, M.G. (2014). Enzymatic and bacterial conversions during sourdough fermentation. Food Microbiol. *37*, 2-10. https://doi.org/10.1016/j.fm.2013.04.007.

15. Gobbetti, M., Cagno, R.D., and De Angelis, M. (2010). Functional microorganisms

for functional food quality. Crit. Rev. Food Sci. Nutr. *50*, 716-727. https://doi.org/10.1080/10408398.2010.499770.

16. Gobbetti, M., Rizzello, C.G., Di Cagno, R., and De Angelis, M. (2014). How the sourdough may affect the functional features of leavened baked goods. Food Microbiol. *37*, 30-40. https://doi.org/10.1016/j.fm.2013.04.012.

17. Walsh, A.M., Leech, J., Huttenhower, C., Delhomme-Nguyen, H., Crispie, F., Chervaux, C., and Cotter, P.D. (2023). Integrated molecular approaches for fermented food microbiome research. FEMS Microbiol. Rev. *47*, fuad001. https://doi.org/10.1093/femsre/fuad001.

18. Carlino, N., Blanco-Míguez, A., Punčochář, M., Mengoni, C., Pinto, F., Tatti, A., Manghi, P., Armanini, F., Avagliano, M., Barcenilla, C., et al. (2024). Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. Cell *187*, 5775-5795. https://doi.org/10.1016/j.cell.2024.07.039.

19. Walsh, A.M., Macori, G., Kilcawley, K.N., and Cotter, P.D. (2020). Meta-analysis of cheese microbiomes highlights contributions to multiple aspects of quality. Nat. Food *1*, 500-510. https://doi.org/10.1038/s43016-020-0129-3.

20. Pavlopoulos, G.A., Baltoumas, F.A., Liu, S., Selvitopi, O., Camargo, A.P., Nayfach, S., Azad, A., Roux, S., Call, L., Ivanova, N.N., et al. (2023). Unraveling the functional dark matter through global metagenomics. Nature *622*, 594-602. https://doi.org/10.1038/s41586-023-06583-7.

21. Du, R., Xiong, W., Xu, L., Xu, Y., and Wu, Q. (2023). Metagenomics reveals the habitat specificity of biosynthetic potential of secondary metabolites in global food

fermentations. Microbiome *11*, 115. https://doi.org/10.1186/s40168-023-01536-8.

22. Cocolin, L., and Ercolini, D. (2015). Zooming into food-associated microbial consortia: a 'cultural' evolution. Curr. Opin. Food Sci. *2*, 43-50. https://doi.org/10.1016/j.cofs.2015.01.003.

23. Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. Science *379*, 1358-1363. https://doi.org/10.1126/science.adf2465.

24. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833-844. https://doi.org/10.1038/nbt.3935.

25. Yu, H., Deng, H., He, J., Keasling, J.D., and Luo, X. (2023). UniKP: a unified framework for the prediction of enzyme kinetic parameters. Nat. Commun. *14*, 8211. https://doi.org/10.1038/s41467-023-44113-1.

26. Smit, G., Smit, B.A., and Engels, W.J. (2005). Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products. FEMS Microbiol. Rev. *29*, 591-610. https://doi.org/10.1016/j.femsre.2005.04.002.

27. Zhang, Y., He, S.D., and Simpson, B.K. (2018). Enzymes in food bioprocessing - novel food enzymes, applications, and related techniques. Curr. Opin. Food Sci. *19*, 30-35. https://doi.org/10.1016/j.cofs.2017.12.007.

28. Yang, J.W., Gao, R.J., Zhou, Y., Anankanbil, S., Li, J.B., Xie, G.Q., and Guo, Z. (2018). β-Glucosidase from *Thermotoga naphthophila* RKU-10 for exclusive synthesis of galactotrisaccharides: Kinetics and thermodynamics insight into

reaction mechanism. Food Chem. *240*, 422-429. https://doi.org/10.1016/j.foodchem.2017.07.155.

29. Feng, J., Huang, Z., Huang, M., Cui, C., Zhao, M., and Feng, Y. (2025). Revealing the Microbial Origins of N-Lactoyl Amino Acids in Soy Sauce: Synthesis Conditions, Potential Enzymes, and Utilization Preference. J. Agric. Food Chem. *73*, 3008-3015. https://doi.org/10.1021/acs.jafc.4c04907.
30. Mageswari, A., Subramanian, P., Chandrasekaran, S., Karthikeyan, S., and Gothandam, K.M. (2017). Systematic functional analysis and application of a cold-active serine protease from a novel *Chryseobacterium* sp. Food Chem. *217*, 18-27. https://doi.org/10.1016/j.foodchem.2016.08.064.
31. van de Wouw, M., Walsh, A.M., Crispie, F., van Leuven, L., Lyte, J.M., Boehme, M., Clarke, G., Dinan, T.G., Cotter, P.D., and Cryan, J.F. (2020). Distinct actions of the fermented beverage kefir on host behaviour, immunity and microbiome gut-brain modules in the mouse. Microbiome *8*, 67. https://doi.org/10.1186/s40168-020-00846-5.
32. Landis, E.A., Oliverio, A.M., McKenney, E.A., Nichols, L.M., Kfoury, N., Biango-Daniels, M., Shell, L.K., Madden, A.A., Shapiro, L., Sakunala, S., et al. (2021). The diversity and function of sourdough starter microbiomes. eLife *10*, e61644. https://doi.org/10.7554/eLife.61644.
33. Rizo, J., Guillén, D., Farrés, A., Díaz-Ruiz, G., Sánchez, S., Wacher, C., and Rodríguez-Sanoja, R. (2020). Omics in traditional vegetable fermented foods and beverages. Crit. Rev. Food Sci. Nutr. *60*, 791-809.

https://doi.org/10.1080/10408398.2018.1551189.

34. Brauer, A., and Bengtsson, M.M. (2022). DNA extraction bias is more pronounced for microbial eukaryotes than for prokaryotes. Microbiologyopen *11*, e1323. https://doi.org/10.1002/mbo3.1323.

35. Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E.H., Bye-A-Jee, H., Carpentier, D., Denny, P., et al. (2025). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. *53*, D609-D617. https://doi.org/10.1093/nar/gkae1010.

36. Zheng, L., Li, B., Xu, S., Chen, J., and Liang, G. (2024). FEDKEA: Enzyme function prediction with a large pretrained protein language model and distance-weighted k-nearest neighbor. bioRxiv, 2024.08.12.604109. https://doi.org/10.1101/2024.08.12.604109.

37. Su, J., Zhou, X., Zhang, X., and Yuan, F. (2024). ProTrek: Navigating the Protein Universe through Tri-Modal Contrastive Learning. bioRxiv, 2024.05.30.596740. https://doi.org/10.1101/2024.05.30.596740.

38. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

39. Haas, B., Papanicolaou, A., and Yassour, M. (2017). TransDecoder. https://github.com/TransDecoder/TransDecoder.

40. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo,

M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. *41*, 1633-1644. https://doi.org/10.1038/s41587-023-01688-w.

41. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251-2252. https://doi.org/10.1093/bioinformatics/btz859.

42. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026-1028. https://doi.org/10.1038/nbt.3988.

43. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59-60. https://doi.org/10.1038/nmeth.3176.

44. Gado, J.E., Knotts, M., Shaw, A.Y., Marks, D., Gauthier, N.P., Sander, C., and Beckham, G.T.J.N.M.I. (2025). Machine learning prediction of enzyme optimum pH. Nat. Mach. Intell. *7*, 716-729. https://doi.org/10.1038/s42256-025-01026-6.

45. Qiu, S., Hu, B., Zhao, J., Xu, W., and Yang, A. (2025). Seq2Topt: a sequence-based deep learning predictor of enzyme optimal temperature. Brief. Bioinform. *26*, bbaf114. https://doi.org/10.1093/bib/bbaf114.

46. Team, R.C. (2013). A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

47. Team, R. (2020). RStudio: Integrated Development for R. R Studio (PBC).

48. Song, Y., Yuan, Q., Chen, S., Zeng, Y., Zhao, H., and Yang, Y. (2024). Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. Nat. Commun. *15*, 8180. https://doi.org/10.1038/s41467-024-52533-w.