Point-by-point responses to the comments

The manuscript has been carefully revised according to the reviewers' comments. Our point-by-point responses to the reviewers' comments and suggestions are listed below.

Reviewers' comments:

Reviewer #1: The authors used the CLEAN algorithm, a published contrastive learning method, to predict enzyme function from metagenome-assembled genomes that were recently published in the curatedFoodMetagenomicData. The authors use this information to comparatively describe the abundance of different enzyme clusters in different food categories, identify unknown enzymes, and trained a random forest classifier to predict food type based on enzyme cluster abundance. The approach is interesting and useful, and to my knowledge has not yet been applied to fermented foods at this scale. Although the results are mostly technical sound (see some minor comments below), the authors do not expand much beyond cataloging and comparing basic food categories — it is not clear what new knowledge or innovation is gained from this approach. Indeed, much could be, but those findings should be presented to make this storyline complete. Moreover, the results are taken as-is without any real validation. Given that the authors used a pre-existing, pre-compiled dataset, and a pre-existing contrastive learning model, it is unclear what new knowledge this study brings — if the dataset were published openly, then this would provide a useful dataset of predicted enzymes from different fermented foods but not even this is provided with this article.

The enzyme sequences, predicted annotations, and metadata should be made publicly available (perhaps even in the curatedFoodMetagenomicData database) for review and on publication.

Response: We sincerely thank the reviewer for their constructive feedback and recognition of the technical soundness and potential utility of our approach. In the revised manuscript, we have added a series of analyses and made substantial revisions to address your concerns.

First, we highlight the work that provides data resources—covering a collection of predicted enzymes from different fermented foods and committed to depositing all the data on public repository. All raw enzyme sequences (fasta files), CLEAN-predicted annotations (EC numbers), metadata (sample-food category mappings), and complete analysis codes (R scripts) are permanently hosted on Zenodo (https://zenodo.org/records/15665866).

In addition, in response to the reviewers' comments on improving the new knowledge or innovation of the research, we have focused on strengthening the analysis of the physicochemical properties of enzymes and niche breadth in the revised manuscript. In order to explore extremely environment stable biocatalysts with industrial potential, we used dedicated computational tools EpHod to predict the optimal pH and Seq2pHopt to predict the optimal temperature parameters of known and new enzyme clusters, analyzed the distribution of the optimal conditions of enzymes. Among them, enzymes with extreme physicochemical tolerance may provide valuable resources for industrial development. To show the environmental distribution patterns of enzyme clusters, we incorporated the analysis of Levins' niche breadth index of the enzyme and rigorously evaluated food-type specificity of enzyme clusters. About 31% of the enzyme clusters are food type specific, and exist in each functional enzyme groups (EC1-7). reflecting the existence of numerous rarely distributed yet functionally diverse enzyme cluster resources within fermented food ecosystems.

Finally, we fully acknowledge the reviewer' valuable insights regarding the validation scope of the CLEAN method. To minimize potential methodological biases, we employed two independent algorithms, FEDKEA and ProTrek, to validate CLEAN's predictions (see STAR Methods for details). For experimental verification of CLEAN's predictions, it is important to note that while initial validation of the method focused on halogenases, its algorithmic framework has since been successfully applied to the functional annotation of diverse enzyme classes, including plastic-degrading enzymes and methyltransferases (*PNAS*, 2024, 121: e2318522121; *Metab. Eng. Commun.*, 2024, 19: e00248).

Minor comments:

Line 75: What is the reasoning behind the 80% threshold?

Response: Thank you for raising this important methodological point. We believe that clustering sequences using MMseqs2 can greatly improve computational efficiency. By grouping highly similar sequences (identity ≥ 80%) into clusters, we significantly reduce redundancy. At the same time, we believe that at such a high clustering threshold, all enzyme sequences in this cluster can be considered to have the same EC number, which allows us to judge the EC number of all sequences in the cluster by predicting the function of a single representative sequence in each cluster instead of all individual sequences, thus significantly reducing computational costs. To verify the robustness of the method, we performed iterative clustering of reference sequences (identity 10-90%). The results show that the error rate (different EC numbers exist in one cluster) decreases to near zero when the identity is ≥ 80%, and the results are discussed in detail in the revised manuscript.

79: The CLEAN approach is an interesting one but it was only validated with halogenases. It would be important to see more validation with relevant enzyme categories for fermented foods.

Response: We appreciate this insightful observation regarding the scope of validation for the CLEAN approach. To address potential methodological bias, we supplemented CLEAN predictions with recently developed tools including FEDKEA and ProTrek, and compared the results across these independent frameworks (see STAR★Methods). The results (**Figure S1C**), it can be seen that a large proportion of clusters below the CLEAN confidence threshold (<0.2) were also classified as low-confidence or nonenzyme clusters by the other two methods (FEDKEA: ≈80%, ProTrek: ≈90%). While CLEAN's original validation focused on halogenases, its application has since been successfully extended to diverse enzyme categories including plastic-degrading enzymes and methyltransferases (*PNAS*, 2024, 121: e2318522121; *Metab. Eng. Commun.*, 2024, 19: e00248).

120: This was not really explained in the method.

Response: Thank you for pointing out the ambiguity in our writing. We have modified the method, specifically, we changed it to "We evaluated the known or novel nature of the entire enzyme cluster by comparing the enzyme sequences of known EC numbers with the members of the fermented food microbial enzyme cluster at the sequence level. We employed DIAMOND (parameters: --evalue 0.001 --sensitive --header-simple --max-target-seqs 1) to align predicted enzyme clusters against enzyme known sequences. Clusters were classified as "known clusters" if at least one sequence in clusters matched a known enzyme sequence under thresholds (sequence identity ≥ 80% and coverage ≥ 80%). The novel enzyme cluster is one that does not match any known sequence under this criterion (see STAR★Methods)." After the revision, the proportion of novel clusters remained high at 84.4%.

163: The metric for diversity per MAG is biased for eukaryotes since they have larger genomes. It should be per gene content.

Response: We thank the reviewer for highlighting this potential bias. To address genome size variation, we reanalyzed enzyme diversity using coding ORF count per genome as a normalization metric, alongside the original per-MAG counts (**Figure 2A, 2F, S3B**).

204: It is not explained in the method how habitat specific is defined and in any case it should be food type specific. Existing methods for identifying cosmopolitan species like the Levins Niche Breadth (NB) index could be applied here.

Response: Thank you. We have incorporated the reviewer's suggestion by adding the Levins' niche breadth index as a quantitative metric to define and rigorously evaluate the food-type specificity of enzyme clusters (**Figure 4A**). Relevant methodological descriptions and results interpretation of the results have been added to the revised manuscript.

231: The clusters don't seem to be super separated, and the matrix is most likely not as important as the substrate.

Response: Thank you for pointing this out. We have added 80 specific food types to the analysis in the revised manuscript in addition to the 10 food categories.

285: The raw materials should be described in more detail — how is this assessed?

Response: Thank you for your comments. We have modified the relevant statements.

313: The methods here are not clear: how were results from both tools merged?

Response: Thank you for pointing out the problems in our method writing. The potentially ambiguous sentence has been revised to: On the basis of taxonomic classifications provided by the database, bacterial-origin MAGs were analyzed using Prodigal (v2.6.3) for open reading frame (ORF) prediction, whereas fungal-origin MAGs were processed with TransDecoder (v5.7.1) to identify translationally active regions. Using the standardized FASTA format, the two outputs were integrated into a unified ORF dataset for subsequent analysis.

320: 276,345 sequences are not many, why was this selection made? Swissprot has twice as many that are well curated and trEMBL many more although most probably don't contain the valid EC

Response: Thank you for your insightful comment. We selected 276,345 sequences with well-defined functions and EC numbers from the reviewed Swiss-Prot subset of UniProt. Although Swiss-Prot and especially TrEMBL include many more sequences, most lack reliable EC annotations or are unreviewed, making them unsuitable for accurate functional analysis. Furthermore, our selection is consistent with the CLEAN method, which also relies on Swiss-Prot as the primary database for model development and evaluation (*Science*, 2023, 379: 1358-1363).

334: As only 20% of the clusters are kept, this is discarding most data and may significantly skew results (e.g., for classification); most of the novel clusters would be the ones discrded.

Response: Thank you for your valuable comment. In our results, about 20% of the enzymes (98,693 clusters) were identified from 472,428 protein clusters, and the rest (low confidence or not annotated to EC numbers) were classified as non-enzymatic proteins. Specifically, CLEAN assigns an EC number to each sequence along with a corresponding confidence score (*Science*, 2023, 379: 1358-1363). Following CLEAN's recommended threshold, we retained only sequences with confidence scores ≥ 0.2 as enzymes. Moreover, compared with other two recently developed annotation tools, FEDKEA and ProTrek, as presented in the revised manuscript. Most of the high-confidence enzyme sequences identified by CLEAN were labeled as enzymes in two independent annotation methods; At the same time, the low-confidence sequences determined by CLEAN were mainly classified as non-enzymatic proteins by the control method (**Figures S1C and S1D**). Finally, the proportion of enzyme clusters (EC1-7) that we identified was almost identical to that between classes in UniProt.

337: Were these sequences clustered with the Uniprot seqs? It should be clearly described.

Response: We have modified the relevant description in the method. Thank you for pointing this out.

342: This is not clear and could be rephrased for clarity. If I understand correctly, this criterium is not the best, should also consider if there is any annotation for any member of the cluster.

Response: Thank you for pointing this out. We have modified the technical approach for determining novel clusters in the revised manuscript. We evaluated the known or novel nature of the entire enzyme cluster by comparing the enzyme sequences of known EC numbers with those of the members of the fermented food microbial enzyme cluster at the sequence level. We employed DIAMOND (parameters: --evalue 0.001 --sensitive --header-simple --max-target-seqs 1) to align the predicted enzyme clusters against known enzyme sequences. Clusters were classified as "known clusters" if at least one sequence in the clusters matched a known enzyme sequence under thresholds (sequence identity ≥ 80% and coverage ≥ 80%). The novel enzyme cluster is one that does not match any known sequence under this criterion. After this revision, the proportion of novel clusters decreased from 88% in the first draft to 84.4%, but it still remained at a high level.

Reviewer #2: In this manuscript, Li et al. evaluate enzyme diversity in the fermented food microbiome using metagenome-assembled genomes (MAGs) and machine learning. While the study explores a timely and intriguing research area with a comprehensive analysis, it lacks a clear definition of the terms and methods used, a direction for further development and does not provide a concise summary of the findings.

Strengths:

The manuscript analyses comprehensive dataset, applies novel methodological approaches which result in important findings in enzyme diversity and food microbiomes. The identification of over 5 million enzyme sequences and 98,693 homologous clusters is impressive, especially highlighting the novelty of up to 88% of enzyme clusters. The use of the CLEAN tool for enzyme annotation is methodologically sound and innovative. The study's findings on habitat-specific enzymes and their link to food matrices provide valuable insights for food microbiology and enzyme discovery.

Response: Thank you for your thoughtful and encouraging feedback. In response, we have substantially revised the manuscript to improve clarity, focus, and conciseness. We clearly defined key terms and methodologies, restructured the results section for better flow, and added summary tables and figures to highlight key findings. Additional analyses were included, such as multi-tool validation of enzyme functions, physicochemical profiling of enzyme clusters, and ecological niche quantification to correct sampling bias. The methods section was thoroughly revised for clarity and reproducibility, and we expanded the discussion to outline future research directions. We believe these revisions significantly enhance our manuscript.

Major comments:

* Unclear Definition: The manuscript is titled "Assessment of Enzyme Diversity…", yet it lacks a clear definition of enzyme diversity and does not explicitly outline the methods used to define and compare it.

Response: Thank you for your insightful comment. Our study aims to elucidate the microbial-encoded enzyme resources in the microbiome of fermented foods, and to show the types of these enzymes, predict EC numbers, taxonomic sources, and environmental distribution. We 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. For these enzyme clusters, we analyze the distribution of their diversity, that is, the abundance or barrenness of enzyme resources, by comparing the types of enzyme clusters (or the number of member sequences) that occur in different functions (EC number or KEGG annotation), different classifications, and different environments.

* Unclear Data Processing: The manuscript often presents results without adequately detailing the data processing and analysis methods. For instance, when comparing MAG or cluster numbers across conditions, it is unclear whether and how these results were normalized based on the sample size of each condition.

Response: We appreciate the reviewer's critical comment regarding the clarity of the data processing methods. In response, we have thoroughly revised the STAR★Methods section to explicitly detail all the data analysis procedures. We used box plots and nonparametric Wilcoxon rank-sum tests to ensure that the sample values did not affect our conclusions. The full details of the normalization and statistical methods are now clearly documented within the revised STAR★Methods section and the relevant figure legends.

* The methods section lacks crucial information regarding the statistical analysis, with only the R packages listed without clearly stating which data was processed and how. For instance: Figure 4D, L227 "On the basis of the composition and diversity of the enzyme clusters" the PCA was done. How was the diversity calculated? Which distance matric was used and why? Was this plot based on abundances or distances?

Response: Thank you for pointing out the shortcomings in our method writing. We have revised the Methods and Results sections to address the unclear points you noted out. The R code for reproducing all our results and the raw data is available at Zenodo (https://zenodo.org/records/15665866). For example, we changed the method to principal coordinates analysis (PCoA), and the input data consisted of a Bray-Curtis distance matrix of nonredundant enzyme cluster species for each sample. PCoA analysis was performed based on 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 shows in 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").

* Insufficient assumptions and limitations: The authors should explicitly state their assumptions and limitations, particularly when interpreting the results. Some assumptions lack support from data or literature, such as in L 216, "number of reconstructed MAGs in the sample information as the alpha diversity of the community". Stating clearly the limitations of the methods is essential for understanding the context and potential biases of the analysis. Generally, the study lacks a clearer comparison with existing literature.

Response: We sincerely thank the reviewers for their important corrections to the research assumptions and limitations. To this end, we have reorganized the entire paper and systematically elaborated all the core assumptions in the results and methods. At the same time, we have deeply expanded the comparison dimension with cutting-edge research in the field in the discussion. These revisions have made the theoretical basis of the study more transparent and the scope of application of the conclusions clearer, fundamentally strengthening the rigor of the research and the interpretability of the conclusions.

* The introduction should provide a more structured background on the significance of enzymatic diversity in fermented foods and why previous studies have not adequately explored this topic.

Response: Thank you for pointing this out. The Introduction has been restructured to establish a stepwise conceptual foundation. This revision directly bridges fundamental knowledge gaps in our methodological approach.

* The Results section is highly data-heavy, which makes it difficult to follow. This section should hold the most pertinent results to the general idea of the paper.

Response: Thank you. We have reorganized the Results section to establish a clear hierarchical framework that prioritizes core biological insights rather than raw data outputs, and all supplementary analyses have been moved to dedicated subsections or figures.

* The discussion lacks depth in presenting key findings and, more importantly, their implications. A more detailed analysis of the novel enzymes identified and their potential applications in food biotechnology or health benefits is needed. Additionally, the discussion should clearly articulate how this study advances our understanding compared to previous research.

Response: We sincerely thank the reviewers for their guidance on the depth of the discussion. The revised manuscript has highlighted that: the industrial adaptability of novel enzyme clusters (e.g., outliers in temperature and pH prediction results of some enzyme clusters) provides a new way to save energy and reduce consumption in food bioprocessing, and its natural properties from the source of fermented food also provide health application scenarios (including the safety advantage of targeted degradation of harmful components); compared with traditional small-scale screening or insufficient functional annotation studies, this study provides the first large-scale, functionally characterized enzyme catalog, covering enzymes specifically from different fermented foods. It has made significant progress over previous studies that are usually small-scale or less functionally annotated.

Minor comments:

* L5 "contrastive learning" needs to be briefly explained.

Response: Thank you for your suggestion. We have added relevant explanations to the revised manuscript.

* L19 - The first sentence of the introduction is long, should be broken into two sentences.

Response: Modified, thanks for pointing this out.

* L24 - rephrase this confusing sentence

Response: Already edited, thank you.

* L71 - the full term for CDS should be written first.

Response: Thank you for pointing this out. We have revised the relevant expressions and ensured that the terminology is consistent throughout the text.

* L73-74 - EC numbers should be written in full here (not later in line 80) and the term should be briefly explained.

Response: Thank you for the comment. We have added the complete statement and explanation to the revised manuscript.

* L149 refer to the data and/or figure that support the result stated here?

Response: Thank you for pointing out that we have added the relevant data to the revised manuscript.

* L159 Is this normalized by the taxonomic diversity and/or genome size of bacteria and fungi?

Response: Thank you for your suggestion. We have significantly expanded our comparative analysis. Our revised results now explicitly account for genomic-scale differences: while the total enzyme cluster diversity remains higher in bacterial MAGs (81,848 clusters) than in fungal MAGs (25,233 clusters), per-genome normalization reveals that bacterial MAGs exhibit lower average enzyme cluster diversity per genome and reduced novelty rates compared to fungi (**Figure S3A**).

* L205 Is this finding controlled by sample number?

Response: Thank you for pointing this out. To rigorously address the reviewer's concern about sample number bias in identifying habitat-specific enzymes, we implemented Levins' niche breadth index as a normalized metric of environmental distribution specificity.

* L209 which statistical-significant tests were applied?

Response: Thank you. For comparisons between sample groups visualized using box plots, statistical significance was rigorously assessed through nonparametric Wilcoxon rank-sum tests.

* L210-215 Please add the supportive data and/or figure here.

Response: Edited, thanks for pointing this out.

* L241-253 How the RF was validated is not clear, in Method, authors mentioned 1:3 split of dataset, but later, the ten-fold cross-validation is mentioned, which is conflicting the initial dataset separation ratio.

Response: Thank you. We appreciate the reviewer's attention to the methodology and clarify that the 3:1 split (75% training, 25% completely held out of the test set) establishes independent evaluation data, whereas the three replications of ten-fold cross-validation occur exclusively within the training partition to tune hyperparameters and select model configurations without touching the test set. The cross-validation process involves repeated randomized partitioning of the training data into 10 folds (9 for training, 1 for validation per iteration) to mitigate overfitting and enhance reliability. Only after finalizing the model through this internal cross-validation do we perform a single evaluation on the untouched 25% test set to report unbiased generalization performance metrics, following standard machine learning practice.

Figures:

* The quality of the main figures is very low, making it difficult to simply view (the supplement figures are of good quality).

* The colour code is not consistent across all figures and panels, making the inference of the plots very confusing.

* The figure legend is unclear, such as Figure 1B displays no percentage above 50%, despite the legend stating otherwise. Furthermore, the percentages in the Venn diagram appear to be nonsensical.

Response: Thank you for pointing out. We have made the corresponding modifications to the figure as suggested. The quality of the main figures may be reduced when the submission system is merged. It is recommended that you download the original figures of the main image in the system.

Reviewer #3: The manuscript entitled "Assessment of enzyme diversity in the fermented food microbiome" by Peng Li et al. presents a comprehensive and well-structured study on the diversity, novelty, and distribution of enzymatic resources in fermented foods using metagenome-assembled genomes (MAGs) and machine learning approaches. The reported work emphasizes the untapped potential of fermented food environments for enzyme resource exploration, offering valuable insights into microbial functions for future food research. In this point of view, this work is very interesting and meaningful, I would say. The study is clearly within the scope of the Cell Systems, and I overall recommend publication of the manuscript. Here are a few comments.

Comments:

1. All sequences in the article were clustered based on 80% amino acid identity and 80% minimum sequence coverage, and protein clusters (a collection of proteins with similar functions) were obtained through quality control. However, there are no literature was cited to support this method and threshold. It is recommended to supplement the parameter selection basis in similar studies. I think these should be provided and discussed in the paper.

Response: We sincerely thank the reviewers for their important suggestions on methodological rigor. In response to this issue, the revised manuscript has added. Using MMseq2, we performed multi-threshold clustering analysis (sequence identity thresholds: 10%–90%) on enzyme sequences from the UniProt database. After excluding multifunctional enzymes and EC numbers with ambiguous functional annotations, cluster members exhibited nearly 100% EC number consistency when the sequence identity threshold reached 80%.

2. The author found that up to 88% of the identified enzyme clusters were not annotated in existing databases, particularly those involved in terpenoid and polyketide metabolism, indicating a high degree of novelty. This high novelty is a key highlight of

the study. It would be beneficial to briefly analyze or discuss why certain enzymes exhibit higher novelty rates. Could this be attributed to their functional diversity, or is it due to insufficient representation in current databases? Are specific ecological or evolutionary factors contributing to this phenomenon? Furthermore, what are the biological implications of these variations in novelty rates? Please discuss.

Response: We thank the reviewer for highlighting this critical observation and appreciate the opportunity to deepen our analysis. In the revised discussion, we addressed the high novelty of terpenoid/polyketide enzymes through an integrated perspective

3. Line 69: Change "metagenome-assembled genomes" to "MAGs".

Response: Thank you. We have corrected this error and ensured consistent terminology throughout the manuscript.

4. Line109-112: The statements of "long-term anthropogenic selection" and "metabolic preference" are not precise enough. Please refine these expressions.

Response: We have rewritten this sentence in the revised manuscript; thank you for pointing it out.

5. Line 149: It is recommended to replace "enzyme systems" with "enzyme clusters", as the term "systems" can lead to ambiguity.

Response: Already edited, thank you.

6. Line149-151: The authors believe that the phenomenon of "large number of enzyme systems related to lipid metabolism and amino acid metabolism" can bring applications in the direction of "improving the nutrient composition of fermented foods". It is

recommended to cite empirical studies on lipid metabolism-related accumulation in fermented foods to support the inference.

Response: Thank you for your valuable suggestions. We have rewritten this sentence in the revised manuscript.

7. Lines 202-204: A detailed analysis of the environmental distribution of enzyme clusters reveals that 39% of these clusters (38,723 out of 98,693) are habitat-specific and originate from various food matrix types. Further analysis of the functional composition of these habitat-specific enzyme clusters is recommended, along with a discussion of their association with environmental differences, particularly in terms of food matrices.

Response: We have addressed the reviewer's insightful suggestion by conducting a further functional and ecological analysis of habitat-specific enzyme clusters. We supplemented the calculations on Levins niche breadth and re-analyzed the food type specificity of enzyme clusters in combination with sample source information. The relevant content has been added in the Results and Discussion sections of the revised manuscript.

8. Line 299: The statement that "eukaryotic genome characterization from MAGs remains underestimated" is made, but the impact of this underestimation on the research findings is not discussed in detail. It would be beneficial to explain how this limitation affects the assessment of enzyme diversity.

Response: Thank you for your valuable suggestions. In the revised manuscript, we have supplemented the calculation of the density of ORF sequence-encoding enzymes on eukaryotic and prokaryotic MAGs, re-performed related analysis, and expanded the discussion section.

9. Figure 1 only shows the first-level EC classification may ignore key enzyme differences. It is recommended to split the second or third-level EC classification in the supplementary materials.

Response: Thank you for pointing out. We have made the corresponding modifications to the figure as suggested so that they do not cause reading difficulties again.

10. In the Figure 4, the color indicates the continent where the sample came from, but there may be significant environmental heterogeneity within the same continent. It is recommended to draw a supplementary map based on environmental type as the classification basis.

Response: Thank you. We have made the corresponding modifications to the figure as suggested so that they do not cause reading difficulties again.