frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 20 September 2023
DOI 10.3389/fmicb.2023.1147007

OPEN ACCESS

EDITED BY
Zhipeng Li,
Jilin Agriculture University, China

REVIEWED BY
Limei Lin,
Nanjing Agricultural University, China
Fei He,
Northeast Normal University, China

*CORRESPONDENCE
Qiushi Li
✉ qiushili_jlu@126.com

RECEIVED 18 January 2023
ACCEPTED 28 August 2023
PUBLISHED 20 September 2023

CITATION
Yan Y, Shi T, Bao X, Gai Y, Liang X, Jiang Y and Li Q (2023) Combined network analysis and interpretable machine learning reveals the environmental adaptations of more than 10,000 ruminant microbial genomes. *Front. Microbiol.* 14:1147007. doi: 10.3389/fmicb.2023.1147007

Combined network analysis and interpretable machine learning reveals the environmental adaptations of more than 10,000 ruminant microbial genomes

Yueyang Yan[1], Tao Shi[2], Xin Bao[3], Yunpeng Gai[4], Xingxing Liang[4], Yu Jiang[2] and Qiushi Li[1,5]*

[1]Key Laboratory for Zoonoses Research of the Ministry of Education, Institute of Zoonosis, College of Veterinary Medicine, Jilin University, Changchun, China, [2]College of Animal Science and Technology, Northwest A&F University, Yangling, China, [3]Department of Stomatology, Taian Central Hospital, Tai'an, Shandong, China, [4]School of Grassland Science, Beijing Forestry University, Beijing, China, [5]Department of Stomatology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, Guangdong, China

Background: The ruminant gastrointestinal contains numerous microbiomes that serve a crucial role in sustaining the host's productivity and health. In recent times, numerous studies have revealed that variations in influencing factors, including the environment, diet, and host, contribute to the shaping of gastrointestinal microbial adaptation to specific states. Therefore, understanding how host and environmental factors affect gastrointestinal microbes will help to improve the sustainability of ruminant production systems.

Results: Based on a graphical analysis perspective, this study elucidates the microbial topology and robustness of the gastrointestinal of different ruminant species, showing that the microbial network is more resistant to random attacks. The risk of transmission of high-risk metagenome-assembled genome (MAG) was also demonstrated based on a large-scale survey of the distribution of antibiotic resistance genes (ARG) in the microbiota of most types of ecosystems. In addition, an interpretable machine learning framework was developed to study the complex, high-dimensional data of the gastrointestinal microbial genome. The evolution of gastrointestinal microbial adaptations to the environment in ruminants were analyzed and the adaptability changes of microorganisms to different altitudes were identified, including microbial transcriptional repair.

Conclusion: Our findings indicate that the environment has an impact on the functional features of microbiomes in ruminant. The findings provide a new insight for the future development of microbial resources for the sustainable development in agriculture.

KEYWORDS

ruminants, metagenomics, machine learning, network, metagenome-assembled genome

1. Introduction

Ruminants, as ancient animals, exhibit a wide range of morphological and ecological diversity (Mennecart et al., 2021). They have adapted to diverse habitats, from tropical jungles (Díaz-Céspedes et al., 2021) to the plateau (Guo et al., 2020); range in size from 2 kg (Pickford, 2001) to 1.5 tons (Sauer et al., 2016); show great variations in diet, feeding on objects ranging from moss (Ihl and Barboza, 2007) to ordinary standard feed (Li et al., 2019); and have adapted to almost all ecosystems on Earth. Ruminants are distinguished by their plant digestion patterns and have evolved the rumen. As one of the most vital organs, the rumen allows partial microbial digestion of feed before it enters the true stomach (van Lingen et al., 2017). The rumen is a crucial factor underlying the domestication of ruminants. The productivity of ruminant livestock depends on their gastrointestinal microbiota, which can transform plant material that humans cannot digest into easily accessible animal products (Shabat et al., 2016). The gastrointestinal microbiota of ruminants is characterized by its diversity and dynamic nature, making it prone to alterations due to changes in diet (Malmuthuge and Guan, 2016), environmental factors (Cholewinska et al., 2021), and the presence of enteric pathogens (Cortés et al., 2020). These perturbations play an integral role in host nutritional intake, behavior, metabolism, immunological function, and development. Natural selection has allowed hosts and symbiotic microbes to evolve as integrated systems.

In both the ecological and social realms, ruminants are immensely valuable. Due to rising consumer demand for animal products resulting from population growth (Seshadri et al., 2018), ruminants play an increasingly vital role in ensuring agricultural security. They generate a significant amount of the meat and milk that are the primary sources of protein in the human diet (Stewart et al., 2019). Nonetheless, sustainable manufacturing confronts significant obstacles due to the depletion of natural resources and the resulting rise in production costs.

The ability of ruminants to utilize microorganisms is one of their key traits. Microorganisms bring significant benefits to ruminant animals. However, due to the diverse functionalities and species diversity of microorganisms, they exhibit intricate physiological and biochemical characteristics, making their in-depth analysis quite challenging (Ban and Guan, 2021). We have uncovered inconsistencies in predicting microbial community structures, primarily stemming from a limited grasp of the mechanisms governing microbial community assembly. In order to mitigate this unpredictability, it is imperative to comprehensively understand the microbiome as a cohesive entity.

With the ongoing increase in the depth of metagenomic sequencing, the range of sequencing is progressively expanding (Vestergaard et al., 2017), while the cost of the technology is decreasing. Thus, large amounts of data can be generated for analysis. Consequently, substantial volumes of data can be generated for analysis. However, despite genomics being inherently data-driven, the resultant datasets are becoming both exceedingly large and complex, thereby giving rise to technological challenges. Recent publication of the most recent collection of gut microbial genomes includes a ruminant whole gastrointestinal tract microbial gene set and the reconstruction of over 10,000 nonredundant ruminant gastrointestinal microbial genomes (Xie et al., 2021). This represents a significant change in the ability to understand the ruminant microbiome. This study used public database collections to characterize the microbiomes and functional groups and applied a metagenomics approach to achieve the following objectives: (1) building microbial cooccurrence networks for exploring linkages in microbial communities, (2) the influence of the network's special microstructure on the survivability of gastrointestinal networks in ruminant was explored, (3) detecting antibiotic resistance in different ruminants on a large scale, and (4) exploring the adaptation of ruminant microbes to their environment.

2. Results

2.1. Microbiological characteristics of the gastrointestinal microbiota of ruminants

The microbes serve as an often overlooked yet independent source of data for understanding host evolution and ecological shifts. Based on Spearman's rank correlation coefficient matrix of the relative abundance of 488 collected microorganisms, a microbial network was constructed using seven microbiomes representing distinct ruminant gastrointestinal environments (Figure 1). To eliminate noise and false positive predictions, a conservative statistical cutoff was adopted to reject points with Spearman coefficients less than 0.85 and p-values less than 0.01 The number of modules in this global network was distributed with a long tail, with an average of 32 modules per network, while 80% of the vertices were concentrated in only 8 modules.

The organization of ruminant gastrointestinal microbial networks varies greatly, and there is minimal relationship between the various species (Figure 2A). Dairy cattle had the highest mean clustering coefficients, which indicated substantial clustering and indicated that the network's nodes tended to form relationships over shorter distances. Sheep, dairy cattle and water buffaloes all had more edges than the others. The quantity of edges varied notably among networks, with dairy cattle networks exhibiting three times the number of edges as compared to goat networks. Additionally, we found that all networks had a modular structure (modularity > 0.3) and dense connections between nodes within modules.

Figure 2B demonstrates that the random attack is less effective than the purposeful approach, showing that the microbial network is more resistant to random attacks. During simulations of purposeful attacks, the network stability rapidly degrades and shows a clear declining trend, illustrating the superiority of the DG (max degree graph node) and PG (max pagerank graph node) importance ranking-based techniques. Sheep had the most resilient microbial network regardless of the technique of hitting, while water deer and yak networks had the most susceptible microbial networks. The efficiency of network attacks using the same method varied in a multimethod robustness assessment of individual networks, with PG attacks performing better in elk and less successfully in deer.

2.2. The K-shell decomposition in gastrointestinal microbiota of ruminants

The K-shell technique is utilized to examine the network's hierarchical structure. The procedure is run repeatedly until the most concentrated core is discovered (Figure 3A). Nodes dispersed at the

10.3389/fmicb.2023.1147007

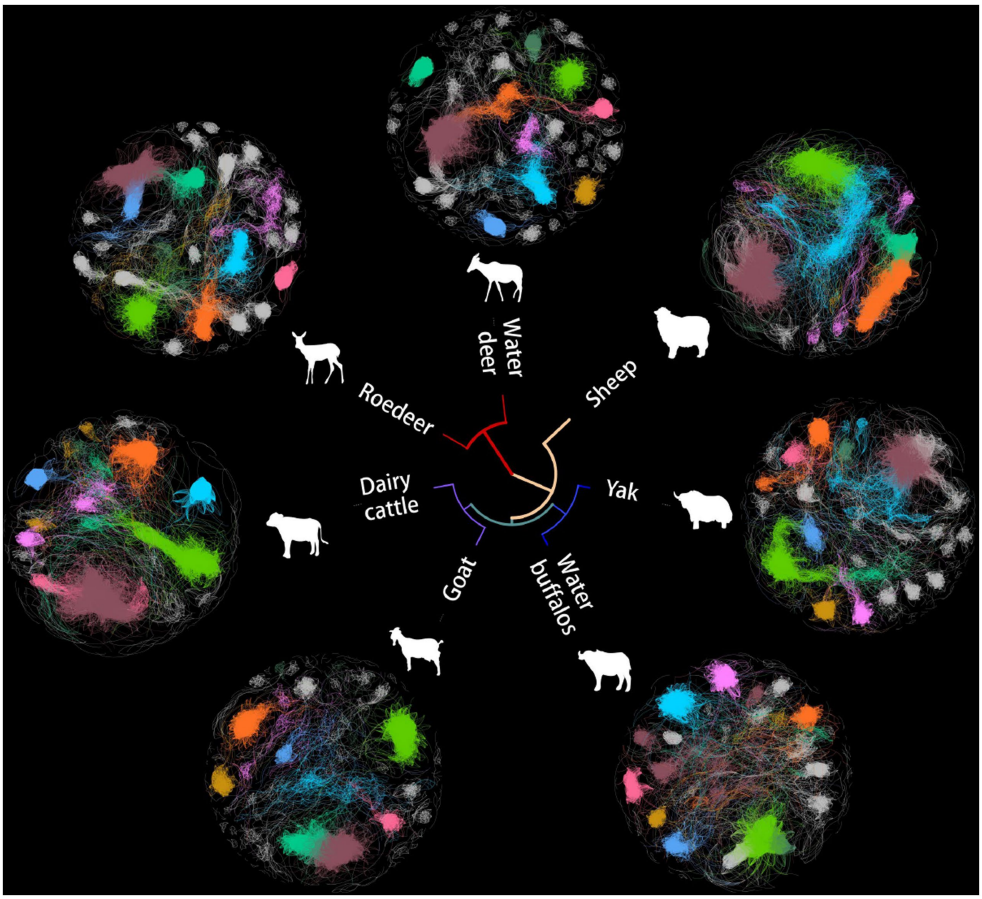

FIGURE 1
Networks of co-occurrence among seven ruminant microbes. Unrooted tree exhibiting a Bray-Curtis clustering tree of microbial genera. The first eight domain modules in the cooccurrence network are displayed in various colors in the outer circle, which displays various animal microbial cooccurrence networks.

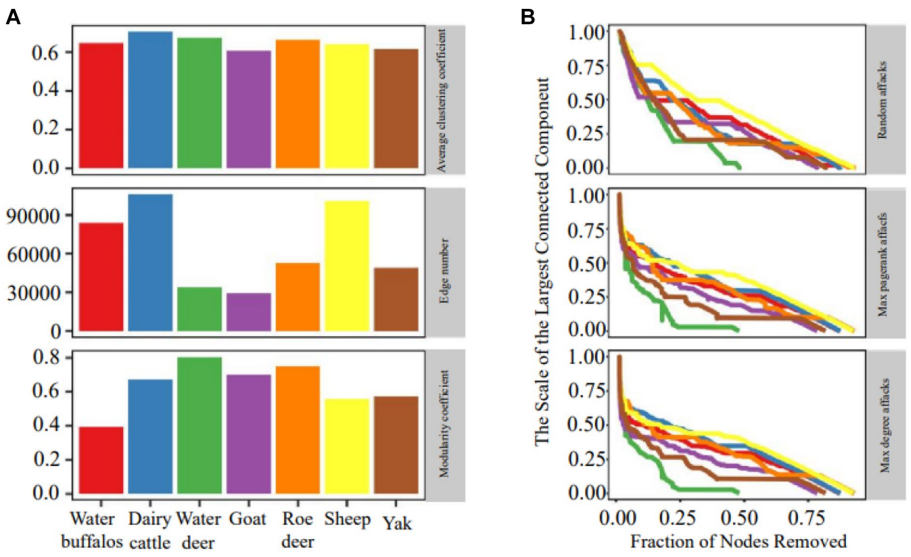

FIGURE 2
Network topology and robustness. **(A)** Microbial network topology inferred from a dataset of microbiome abundance of seven ruminant species. **(B)** The relationship between the proportion of removed nodes and the scale of the largest connected component.

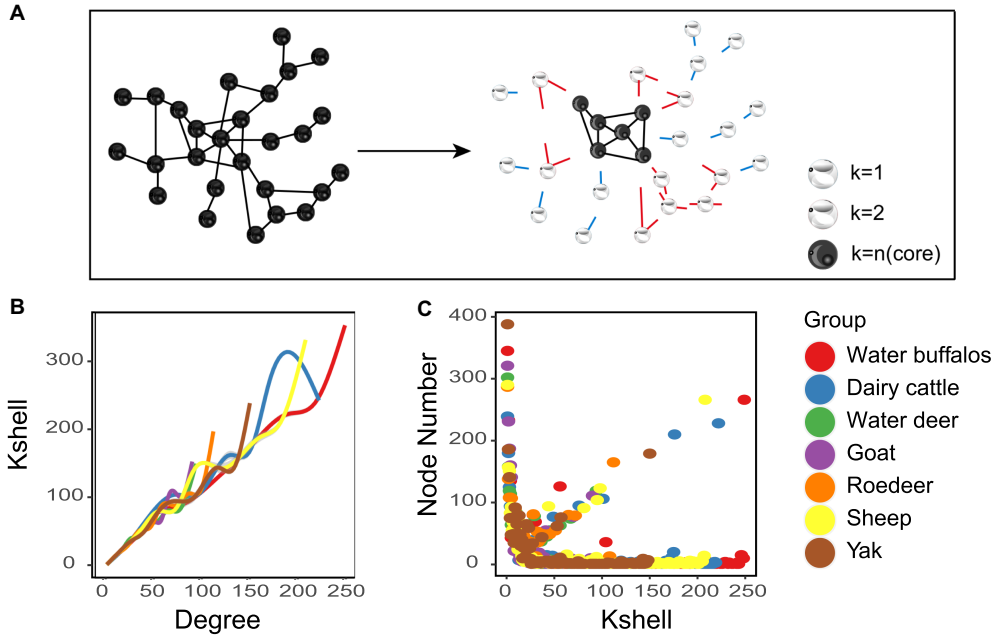

FIGURE 3
Network analyses of K-shell. **(A)** Schematic illustration of network layering using the k-shell decomposition method. **(B)** Distribution of the average degree of each shell. **(C)** each shell number.

network's periphery have a much lower average than nodes positioned close to its core (Figure 3B). The statistics suggest that inner layer nodes may be more efficient information emitters (Figure 3C). We also found that the overall structure of the microbial network topology is reminiscent of the internet network. While merely 0.5% of the nodes in the internet network constitute the nucleus, a mere 0.1% of the nodes comprise the core of the microbiological network.

3. Broad-spectrum profile of ARG abundance in ruminant gastrointestinal

Antibiotics stand as influential elements in microbial networks, exerting substantial impacts. Unveiling antibiotic resistance genes' presence holds the potential to influence the operational dynamics of microbial communities and the overall stability of ecosystems. We discovered a significant prevalence of antibiotic resistance genes (ARGs) in ruminant gastrointestinal microbes, with 6,268 (62%) of the 10,073 genomes tested in the microbial resistance gene analysis yielding positive results. Multiple ARGs demonstrated a high degree of variability within the ruminant resistance group (Figure 4). We also quantified the possible mechanisms of the identified ARGs. Multiple ARGs demonstrated a high degree of variability within the ruminant resistance group. We found that yaks had the fewest ARGs, with an average of 181 ARGs per sample. We hypothesize that this is because the Qinghai–Tibet Plateau is less contaminated as a result of human activities.

We further analyzed the detailed composition of ARGs in the gastrointestinal of seven ruminants. Among all the types of ARGs, uppP, tufab, tetW, tetT, tet37, rpsL, rpsJ, rpoC, rpoB, parE, nimJ, macB, and gyrAAPH (2″)-Ig, which had the highest prevalence, were found in the gastrointestinal tracts of all ruminants. We quantified the risk index of species containing risk genes using quantitative methods, and we characterized the risk level of MAG based on the risk score quartile. The risk level is divided down into 4 levels: 2611 (risk index 1), 1,491 (1 = risk index 10), 657 (10 = risk index 100), and 62 (risk index ≥100). The total number of level 1 microorganisms was 62, or 1.2%. We displayed the tick microbial symbiont risk indicator top 50 in. We discovered that among the top 10 MAGs of all hazards, there were members of Pseudomonas, Escherichia, and Acinetobacter.

Explanation of the machine learning-based analysis of the evolution of gastrointestinal microbial environmental suitability in ruminants.

Many creative strategies are supported by machine learning that can detect patterns and trends in huge data that cannot be identified using traditional analysis-based methods. To this end, we further investigated the differences between rumen microbes at different altitudes using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database based on feature engineering. The 330 yak rumen microbial genomes were subjected to KEGG annotation with 1992 other ruminant rumen microbial genomes to generate a KEGG microbial function matrix (Figure 5) that indicated whether each genomic source was from a high altitude. Thus, the raw data were transformed into features that were more representative of potential problems with the prediction model. The ratio of the number of high-and low-elevation samples was approximately 1: 6. The significant imbalance between these two categories was evident. In order to balance the two sets of sample data, our study employed the synthetic minority oversampling technique (SMOTE) algorithm, which is a completely sampled synthetic data algorithm.

We developed four models of machine learning. The accuracy and applicability of the various techniques were assessed using

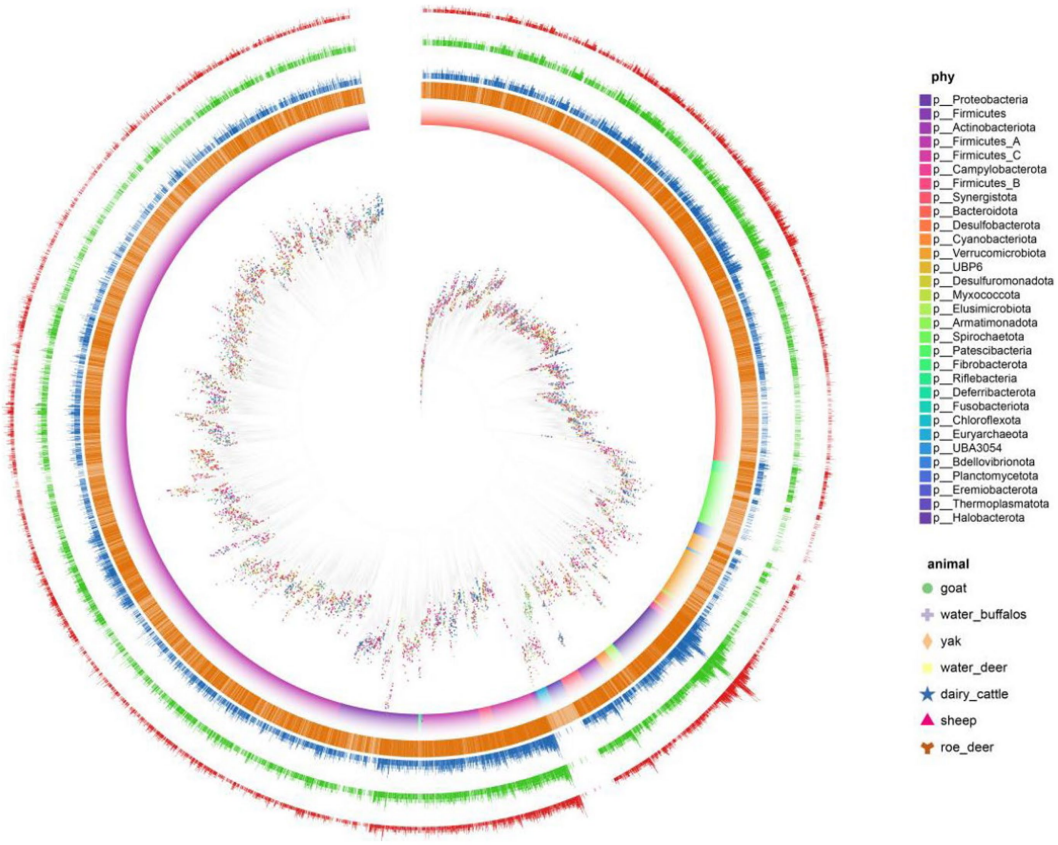

FIGURE 4
Antibiotic-resistance distribution in gastrointestinal microbes. Clades are 146 colored according to phyla and SHAP represents different hosts. The first layer uses 147 orange represents the genome of antibiotics. Second, three, and the four layers represent 148 the number of ARG, how many types and the ARG risk index.

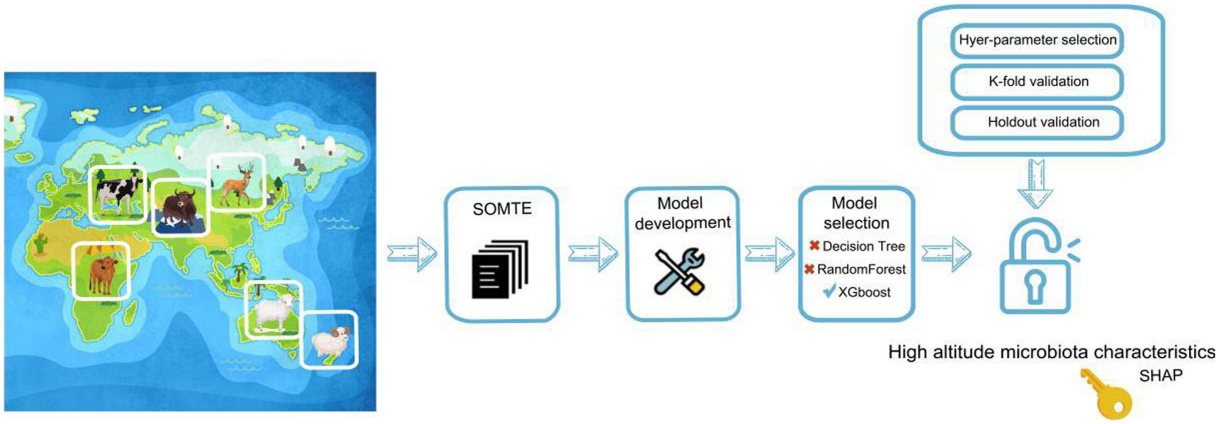

FIGURE 5
Flow chart of machine learning implementation for predicting the adaptive function of microorganisms.

cross-validation score, incorporating the evaluation of 10 distinct sets of cross-validation (Figure 6A). In the 10 cross-validations, the effect of the training and test sets became less effective as more tests were conducted, and the green line (lightGBM) outperformed all the other algorithms. Because of this, we ultimately chose lightGBM as the algorithm for the model. After choosing the LightGBM algorithm, hyperparameter training was performed. The input data were first divided into a test set and a training set, and then a search for hyperparameters was carried out. The appropriate parameters were selected using a plotted learning curve. By aggregating multiple

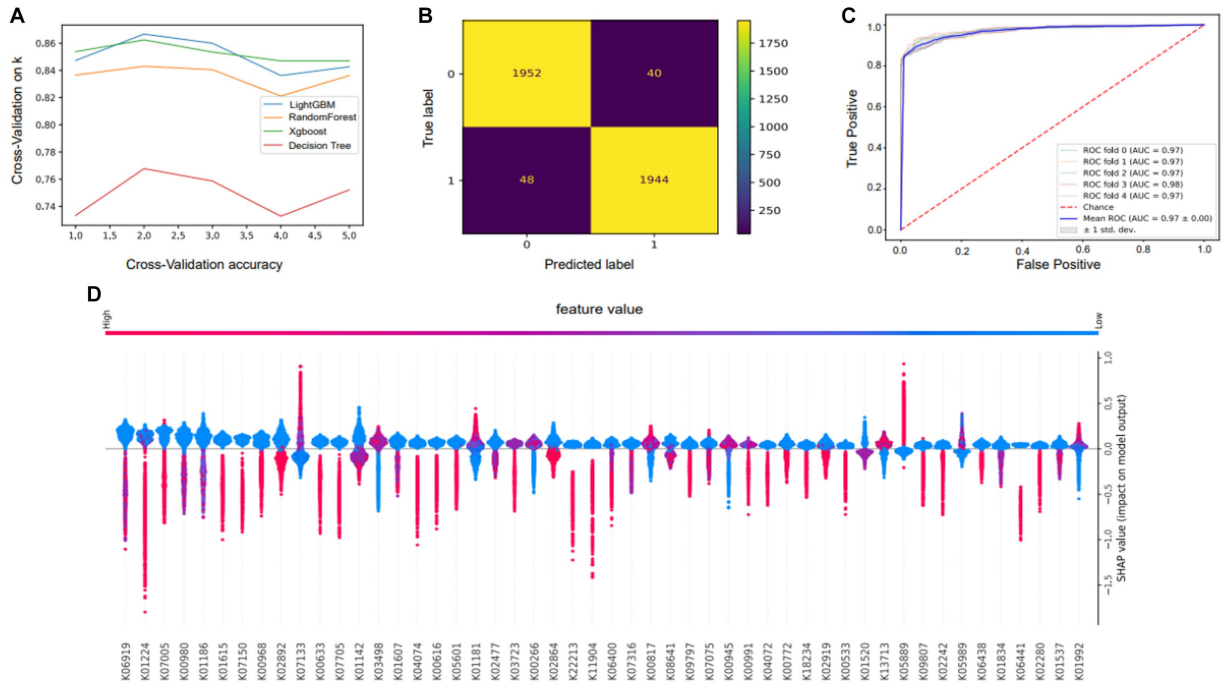

FIGURE 6
Performance and characteristics of the machine learning model. **(A)** Selection of model algorithms based on 10 cross-validation sets: lightGBM is the blue line, XGBoost is the green line, random forest is the yellow line and the decision tree is the red line. **(B)** Confusion matrix showing the performance of the LightGBM model. **(C)** Receiver operating characteristic (ROC) curve of the RF model. The horizontal coordinate of this curve is the false-positive rate (FPR), and the vertical coordinate is the true-positive rate (TPR). **(D)** Plots summarizing the SHAP values of all the samples were used for analysis to interpret key features.

hyperparameters, this model achieved an accuracy of 92.2%, representing an improvement of 13.68% over the lightGBM before tuning. To evaluate the efficacy of the model, a receiver operating characteristic (ROC) curve was constructed (Figure 6C). The closer the ROC curve is to the upper left corner, the better the performance of the classifier; the point on the ROC curve closest to the upper left corner is the best threshold for the lowest number of classification errors and the lowest number of false positives and false negatives. The ROC of the model is 0.97, which shows that the model has good prediction performance and is accurate and reliable and that there is no overfitting.

The mean absolute value of a feature's degree of influence on the target variable was utilized to determine its significance. Based on the Shapley additive explanation (SHAP) values (Figure 6D), the importance of features depicts the mean absolute SHAP values to illustrate the importance of global features. The abundance of genes in the KEGG orthology gene function category provides insight into the genetic underpinnings of adaptive phenotypic variation. The bee swarm plot is intended to show a summary of how the top characteristics of a data set influence the model's output. In each instance, the given explanation is represented by a single dot on each feature row. The x coordinate of the dot is determined by the SHAP value of that feature, and dots "pile up" along each feature row to represent density. Color is used to display the original value of a feature. Our investigation yielded insights into the genetic factors essential for the environmental adaptation of the yak gastrointestinal microbiota, highlighting the importance of the transcription-repair coupling factor (K03723), trk/ktr system potassium uptake protein (K03498), amino acid metabolism (K00817), and genes associated with polysaccharide and lipid metabolism (K05989, K01181). These identified genetic elements play pivotal roles in enabling the yak gut microbiota to effectively adapt to its surroundings. For example, the transcription-repair coupling factor likely contributes to genetic stability and maintenance in dynamic environments. Similarly, the trk/ktr system potassium uptake protein may support the maintenance of ionic balance critical for physiological functions. Amino acid metabolism, as well as polysaccharide and lipid metabolism genes, suggest the microbiota's ability to efficiently extract nutrients from its environment.

Enrichment analysis based on these gene revealed that the microbial community of yaks exhibits more robust specific pathways compared to other plains ruminants (Supplementary Figure S1). These pathways encompass carbohydrate metabolism, nucleotide metabolism, transport systems, and amino acid metabolism, potentially playing a pivotal role in their adaptation and survival within their specific environment. The heightened activity of carbohydrate metabolism suggests efficient energy extraction from their diet, supporting energy-intensive processes. Emphasis on nucleotide metabolism might indicate enhanced DNA and RNA synthesis, potentially aiding cellular growth and repair. The well-developed transport system could facilitate nutrient absorption and intercellular communication, thereby optimizing resource utilization. Overall, these findings shed light on the mechanisms by which the yak gut microbiota adapts to its habitat and offers valuable insights into the genetic underpinnings of its environmental resilience.

4. Discussion

Ruminants are among the most successful herbivorous mammals (Decker et al., 2009). In this study, we took full advantage of the largest and most comprehensive database of ruminant gastrointestinal microbes available. We employed a range of approaches, encompassing complex networks and interpretable machine learning, to characterize the state of environmental microbial populations.

In this study, we employ network science theory to examine the properties of gastrointestinal networks and evaluate the robustness of these networks by examining these properties in more detail. Diverse profiles of topological features in diverse environmental networks reveal the unique co-occurrence patterns of microorganisms in ruminant species. The prevalence of cross-feeding relationships in the network may be indicated by the high clustering coefficient, which suggests that these settings have abundant degradation pathways, niche filtering, or environmental unpredictability. Different models built to test the resilience of networks reveal that they are more resilient to random faults and more vulnerable to deliberate attacks. Understanding the resilience of networks and the various approaches to averting catastrophic failures of these networks is essential. Only a tiny number of significant species have been eliminated, which may have an effect on the general structure of the network of microbes in a healthy microbiome. This finding highlights the importance of using antibiotics sparingly once more.

By investigating antibiotic resistance genes, we gain insights into how microbial communities respond to antibiotics. This understanding is pivotal, as it unravels the intricate interplay among microorganisms and is fundamental to comprehending the intricate web of interactions that shape ecosystems (Sharland et al., 2015). The presence of antibiotic-resistant microbiomes (ARBs) and ARGs in supermarket meat and dairy products suggests that ARBs/ARGs from ruminants can penetrate the food system. It would be helpful to make an effort to compile a list of significant ARG-carrying species for monitoring and control based on the assessment of the total antibiotic resistance risk at the species level. It is alarming that the farming environment contains high-risk MAGs. Additionally, microbiomes communities frequently experience an increase or decrease in ARGs as a result of genetic changes or HGT. Human health would be seriously endangered by these high-risk MAGs.

The Qinghai–Tibet Plateau, sometimes known as the "Third Pole," is a huge, high-altitude region with a unique and fragile ecological environment (Yao et al., 2012). The region is characterized by a harsh climate of extreme cold, drought, high ultraviolet radiation and a lack of oxygen, making it a challenging living environment for humans and other mammals (Zhu et al., 2018; Pan et al., 2021; Shen et al., 2021). It is essential to determine precisely which genetic features give ruminants their exceptional digestive capacity and ability to live in harsh conditions (Zhu et al., 2020). To answer this question, we developed interpretable machine learning methods to deeply mine complex, high-dimensional metagenomic data. We found a significant increase in the transcription repair coupling factor (K03723) in the yak gastrointestinal microbiota. K03723 regulates transcriptional processes and recognizes DNA damage. In addition, such phenomena were also found in samples from plateau-based animals for *Rhodobacter* sp. (Pérez et al., 2018) and nitrogen-fixing microbiomes (Suyal et al., 2018). We hypothesize that this may be a common measure adopted by microorganisms facing extreme environments. In addition, we found that the K03498 trk/ktr system potassium uptake protein contributes significantly to plateau acclimatization. We speculate that the numerous high-salinity sites in the plateau region have resulted in a microbial response to salt stress (Li et al., 2021). Similar to previous studies (Guo et al., 2021),the results indicate a preference for an amino acid metabolism gene (K00817) and polysaccharide and lipid metabolism genes (K05989 and K01181) in the yak gastrointestinal microbiota. These pathways provide additional adaptive responses to the lack of energy intake in yaks. In conclusion, our study provides important insights into ruminant plateau adaptation and highlights the key role of the microbial genome as a "second genome" for adaptation, contributing to a more comprehensive understanding of mammals living in extreme environments.

5. Conclusion

In-depth exploration of ruminant gastrointestinal microbes is necessary to understand the function of the microbiome and its interactions with the host animal. This study enhances our comprehension of both the structure and function of the ruminant gastrointestinal microbiota, a critical aspect for investigating microbial-host symbiotic functional dynamics. Furthermore, it advances our understanding of the gastrointestinal microbiota adaptations necessary for herbivores. In addition, it informs strategies to decrease contamination and increase the robustness and efficiency of ruminants.

6. Methods

6.1. Data used in this study

The sequence files of 10,373 gastrointestinal microbial genomes of ruminants were downloaded in FASTA format from Figshare (DOI: 10.6084/m9.figshare.14176574). All the gene catalogs, annotation information, abundance profiles, assemblies, and predicted open reading frames (ORFs) from this study are available at https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-021-01078-x. The details of all samples used in this study are provided in Supplementary Table S1.

6.2. Network analysis

A Spearman correlation matrix was calculated based on the relative abundance of genera in each sample, and networks were graphed using Gephi (Bastian et al., 2009). Topological features were estimated with the igraph package (Csardi and Nepusz, 2006) (v1.4.1) in R 3.6.0.

The robustness of gastrointestinal networks can be regarded as the ability of the entire network to maintain the same performance when nodes in the network are affected by random factors and intentional damage. The survivability of networks can be measured by the change in network topological structure characteristics. We selected the relative size of the largest connected subgraph (RS) and relative connectivity efficiency (RE) after node failure as the measurement indexes of network survivability (Wei et al., 2021).

$$RS = \frac{N^t}{N}$$

where N is the number of nodes in the largest connected subgraph and N is the total number of nodes in the airline network. As the nodes in the network are attacked, the network splits into several subgraphs, and the relative size of the largest connected subgraph gradually decreases until the nodes are no longer connected with each other and finally become scattered nodes.

In this research, different attack strategies are developed to further understand the failure process of the network under different scenarios, specifically including random failures and targeted attacks (Wang et al., 2010). The strategies can be summarized as follows: First, there is the random attack strategy based on nodes. Second, the largest degree first attack strategy based on nodes. Third, the largest pagerank degree first attack strategy based on nodes.

7. Functional analysis of microbial genes

7.1. ARG annotation

The functional annotation of all identified proteins encoded by ARGs was based on sequence similarity searches carried out with DIAMOND BLASTP v2.0.9 (Buchfink et al., 2021) with the default settings against the HMD-ARG database (Li et al., 2021). ARGs were selected at a sequence similarity threshold of 75% and a score threshold of 60 (Yan et al., 2022). In our study, we utilized a set of equations to effectively quantify the antibiotic risk associated with the presence of ARGs within different Metagenome-MAGs in the context of the larger ecosystem. These equations, as described by equation (Zhang et al., 2022), were structured as follows:

$$Risk = \frac{num\_ARG}{total\_ARG} \times \frac{num\_ARG\_subtypes}{total\_ARG\_suntypes}$$

In this context, "num_ARG" stands for the count of ARGs detected within a specific MAG, while "total_ARG" refers to the overall count of ARGs encompassing the entire ecosystem under consideration. On the other hand, "num_ARG_subtypes" represents the tally of distinct ARG subtypes found within the same MAG, and "total_ARG_subtypes" denotes the comprehensive count of distinct ARG subtypes present within the entire ecosystem. By applying this approach, we aimed to evaluate and quantify the potential antibiotic risk posed by the presence of ARGs within each MAG, within the broader context of the ecological system. This method allowed us to gain insights into the degree of antibiotic resistance-related risk associated with specific MAGs and their ARG compositions, contributing to a more comprehensive understanding of the ecosystem's antibiotic resistance dynamics.

8. Machine learning model development

8.1. Data collection

In this experiment, 330 yak rumen microbial genomes were used as input data, along with 1992 other ruminant genomes. After comparing the protein sequences to the database using eggNOG, the proteins were grouped into different KOs (*KEGG Orthology*), with each cluster of KOs consisting of direct homologous sequences so that the function of the sequence could be inferred. The KO gene function matrix was built as an input file for machine learning.

8.2. Data preprocessing

Unbalanced data can pose challenges for machine learning models. Most machine learning models assume that the same number of samples is available for each class. Ignoring this problem can lead to errors in a few classes (and thus make the model sensitive to classification errors), causing ML models to ignore observations in a few classes. In the current work, the number of samples collected was uneven because the number of ruminant gastrointestinal samples varied from region to region. The synthetic minority oversampling technique (SMOTE) method was applied to overcome the adverse effect of learning data imbalance (Chawla et al., 2002).

8.3. Model development and tuning

Four machine learning models were developed to predict microbial plateau adaptive function in the Jupyter lab development environment (Perkel, 2018) using scikit-learn,[1] Numpy (v1.15.3), Pandas (v0.23.4), Matplotlib (v3.0.1), and Scipy (v1.1.0) for experiments. The machine learning algorithms used for classification in this work were random forest (Qi, 2012), decision tree (Somvanshi et al., 2016), light gradient boosting machine (lightGBM) (Ke et al., 2017), and XGBoost (Chen and Guestrin, 2016). These are all integrated tree-based learning methods and are rated by the machine learning community as the most popular nonlinear models today. LightGBM is a fast and efficient GBDT algorithm in the open-source promotion framework designed by Microsoft MSRA in 2016. The algorithm is used for many machine learning tasks, such as sorting, classification, and regression, and supports efficient parallel training.

8.4. Shapley additive explanation

SHAP quantifies the importance of variables by leveraging Shapley values, a concept originating from cooperative game theory introduced by Shapley in (Bouneder et al., 2020). SHAP's theoretical foundation is rooted in cooperative game theory, as highlighted by Lundberg and Lee in 2017 (Van den Broeck et al., 2022). The methodology explicates the model's predictions by embracing the idea of additive feature attribution. The fundamental principle of SHAP is to decompose the explanation of a prediction into contributions from each feature (Wang et al., 2022). It assigns each feature's contribution based on its Shapley value across different subsets of features, which is equivalent to a weighted average of feature contributions. Shapley values are a concept from cooperative game theory and denote the average contribution of a player across all possible coalition formations.

The SHAP value of feature i(ϕi) can be computed using the following equation:

1 https://scikit-learn.org

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|M|-|S|-1)!}{|M|!} \left[f_x\left(S \cup \{i\}\right) - f_x\left(S\right) \right]$$

Where: N represents the set of all features. *S is a subset of NN that does not include feature i.* $f\left(S\right)$ represents the model's prediction when considering only the feature set *S*. $f\left(S \cup \{i\}\right)$ is the model's prediction when feature ii is added to the features in subset S. The idea behind this formula is to consider all possible combinations of features, excluding feature ii, and calculate the difference in model predictions when feature ii is included in these combinations. The Shapley value concept assigns weights to these differences based on the number of ways a specific feature can contribute to different subsets of features. The summation calculates the weighted average of these differences, yielding the SHAP value for feature *i*.

In essence, the SHAP value quantifies the contribution of each feature to the model's prediction by considering how including or excluding that feature influences the model's output across various combinations of features.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YY conceived, designed the overall study, conceived, designed, and executed the bioinformatics analysis. XB and YG provided the genomic information. All authors contributed intellectually to the interpretation and presentation of the results in the manuscript, which was edited and approved by all authors.

Acknowledgments

We acknowledge Shengdong Hou from East China Jiao Tong University for helpful scientific discussion.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1147007/full#supplementary-material

References

Ban, Y., and Guan, L. L. (2021). Implication and challenges of direct-fed microbial supplementation to improve ruminant production and health. *J. Anim. Sci. Biotechnol.* 12:109. doi: 10.1186/s40104-021-00630-x

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks" in *Proceedings of the International AAAI Conference on Web and Social Media*, 361–362.

Bouneder, L., Léo, Y., and Lachapelle, A. (2020). X-SHAP: towards multiplicative explainability of machine learning. *arXiv*. doi: 10.48550/arXiv.2006.04574

Buchfink, B., Reuter, K., and Drost, H.-G. J. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, T., and Guestrin, C. (2016) Xgboost: a scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA. (2016).

Cholewinska, P., Gorniak, W., and Wojnarowski, K. (2021). Impact of selected environmental factors on microbiome of the digestive tract of ruminants. *BMC Vet. Res.* 17:25. doi: 10.1186/s12917-021-02742-y

Cortés, A., Rooney, J., Bartley, D. J., Nisbet, A. J., and Cantacessi, C. (2020). Helminths, hosts, and their microbiota: new avenues for managing gastrointestinal helminthiases in ruminants. *Expert. Rev. Anti Infect. Ther.* 18, 977–985. doi: 10.1080/14787210.2020.1782188

Csardi, G., and Nepusz, T. J. I. (2006). The igraph software package for complex network research. *Inter J Compl Syst* 1695, 1–9.

Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., et al. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc. Natl. Acad. Sci. U. S. A.* 106, 18644–18649. doi: 10.1073/pnas.0904691106

Díaz-Céspedes, M., Hernández-Guevara, J. E., and Gómez, C. (2021). Enteric methane emissions by young Brahman bulls grazing tropical pastures at different rainfall seasons in the Peruvian jungle. *Trop. Anim. Health Prod.* 53:421. doi: 10.1007/s11250-021-02871-4

Guo, W., Wang, W., Bi, S., Long, R., Ullah, F., Shafiq, M., et al. (2020). Characterization of anaerobic rumen fungal community composition in yak, Tibetan sheep and small tail Han sheep grazing on the Qinghai-Tibetan plateau. *Animals* 10:144. doi: 10.3390/ani10010144

Guo, N., et al. (2021). Seasonal dynamics of diet–gut gastrointestinal microbiota interaction in adaptation of yaks to life at high altitude. *NPJ Biofilms Microb* 7:38. doi: 10.1038/s41522-021-00207-6

Sharland, M., Pulcini, C., Harbarth, S., Zeng, M., Gandra, S., Mathur, S., et al. (2015). Expert Committee on Selection and Use of Essential Medicines. Classifying antibiotics in the WHO Essential Medicines List for optimal use-be AWaRe. *Lancet. Infect. Dis.* 18, 18–20. doi: 10.1016/S1473-3099(17)30724-7

Ihl, C., and Barboza, P. S. (2007). Nutritional value of moss for arctic ruminants: a test with muskoxen. *J Wildl Manag* 71, 752–758. doi: 10.2193/2005-745

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree" in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, USA, 3149–3157.

Li, Z., Gao, Y., Wang, S., Lu, Y., Sun, K., Jia, J., et al. (2021). Phytoplankton community response to nutrients along lake salinity and altitude gradients on the Qinghai-Tibet plateau. *Ecol. Indic.* 128:107848. doi: 10.1016/j.ecolind.2021.107848

Li, F., Li, C., Chen, Y., Liu, J., Zhang, C., Irving, B., et al. (2019). Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. *Microbiome* 7:92. doi: 10.1186/s40168-019-0699-1

Li, Y., Xu, Z., Han, W., Cao, H., Umarov, R., Yan, A., et al. (2021). HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome* 9, 1–12. doi: 10.1186/s40168-021-01002-3

Malmuthuge, N., and Guan, L. L. (2016). Gastrointestinal microbiome and omics: a new definition to ruminant production and health. *Anim. Front.* 6, 8–12. doi: 10.2527/af.2016-0017

Mennecart, B., Aiglstorfer, M., Li, Y., Li, C., and Wang, S. (2021). Ruminants reveal Eocene Asiatic palaeobiogeographical provinces as the origin of diachronous mammalian Oligocene dispersals into Europe. *Sci. Rep.* 11:17710. doi: 10.1038/s41598-021-96221-x

Pan, X., Guo, X., Li, X., Niu, X., Yang, X., Feng, M., et al. (2021). National Tibetan Plateau Data Center: promoting earth system science on the third pole. *Am Meteorol Soc* 102, E2062–E2078. doi: 10.1175/BAMS-D-21-0004.1

Pérez, V., Dorador, C., Molina, V., Yáñez, C., Hengst, M. J. A. V., and SP, L. R. (2018). Rb3, an aerobic anoxygenic phototroph which thrives in the polyextreme ecosystem of the Salar de Huasco in The Chilean Altiplano. *Antonie Van Leeuwenhoek* 111, 1449–1465. doi: 10.1007/s10482-018-1067-z

Perkel, J. M. J. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature* 563, 145–147. doi: 10.1038/d41586-018-07196-1

Pickford, M. J. (2001). Africa's smallest ruminant: a new tragulid from the Miocene of Kenya and the biostratigraphy of east African Tragulidae. *Geobios* 34, 437–447. doi: 10.1016/S0016-6995(01)80007-3

Qi, Y. (2012). "Random forest for bioinformatics" in *Ensemble machine learning*. eds. C. Zhang and Y. Q. Ma (New York: Springer), 307–323.

Sauer, C., Bertelsen, M. F., Lund, P., Weisbjerg, M. R., and Clauss, M. (2016). Quantitative macroscopic anatomy of the giraffe (*Giraffa camelopardalis*) digestive tract. *Anat. Histol. Embryol.* 45, 338–349. doi: 10.1111/ahe.12201

Seshadri, R., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., Cookson, A. L., et al. (2018). Cultivation and sequencing of rumen microbiome members from the Hungate1000 collection. *Nat. Biotechnol.* 36, 359–367. doi: 10.1038/nbt.4110

Shabat, S. K. B., Sasson, G., Doron-Faigenboim, A., Durman, T., Yaacoby, S., Miller, M. E. B., et al. (2016). Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *ISME J.* 10, 2958–2972. doi: 10.1038/ismej.2016.62.

Shen, X., Huo, B., Li, Y., Song, C., Wu, T., and He, J. (2021). Response of the critically endangered Przewalski's gazelle (*Procapra przewalskii*) to selenium deprived environment. *J. Proteome* 241:104218. doi: 10.1016/j.jprot.2021.104218

Somvanshi, M., Chavan, P., Tambade, S., and Shinde, S. (2016). "A review of machine learning techniques using decision tree and support vector machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, India, pp. 1–7.

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* 37, 953–961. doi: 10.1038/s41587-019-0202-3

Suyal, D. C., Kumar, S., Joshi, D., Soni, R., and Goel, R. (2018). Quantitative proteomics of psychotrophic diazotroph in response to nitrogen deficiency and cold stress. *J. Proteome* 187, 235–242. doi: 10.1016/j.jprot.2018.08.005

Van den Broeck, G., Lykov, A., Schleich, M., and Suciu, D. J. J. (2022). On the tractability of SHAP explanations. *J. Artif. Intell. Res.* 74, 851–886. doi: 10.1613/jair.1.13283

van Lingen, H. J., Edwards, J. E., Vaidya, J. D., van Gastelen, S., Saccenti, E., van den Bogert, B., et al. (2017). Diurnal dynamics of gaseous and dissolved metabolites and microbiota composition in the bovine rumen. *Front. Microbiol.* 8:425. doi: 10.3389/fmicb.2017.00425

Vestergaard, G., Schulz, S., Schöler, A., Schloter, M. J. B., and Soils, F. O. (2017). Making big data smart—how to use metagenomics to understand soil quality. *Biol. Fertil. Soils* 53, 479–484. doi: 10.1007/s00374-017-1191-3

Wang, Z., Scaglione, A., and Thomas, R. J. (2010). Electrical centrality measures for electric power grid vulnerability analysis. Proceedings of the 49th IEEE Conference on Decision and Control, Atlanta: IEEE, 5792–5797.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., and Tysklind, M. (2022). Towards better process management in wastewater treatment plants: process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* 301:113941. doi: 10.1016/j.jenvman.2021.113941

Wei, Y., Jin, Y., Ma, D., and Xiu, C. (2021). Impact of colored motif characteristics on the survivability of passenger airline networks in China. *Phys A* 566:125658. doi: 10.1016/j.physa.2020.125658

Xie, F., Jin, W., Si, H., Yuan, Y., Tao, Y., Liu, J., et al. (2021). An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* 9:137. doi: 10.1186/s40168-021-01078-x

Yan, Y., Li, H., Fayyaz, A., and Gai, Y. (2022). Metagenomic and network analysis revealed wide distribution of antibiotic resistance genes in monkey gut microbiota. *Microbiol. Res.* 254:126895. doi: 10.1016/j.micres.2021.126895

Yao, T., Thompson, L. G., Mosbrugger, V., Zhang, F., Ma, Y., Luo, T., et al. (2012). Third pole environment (TPE). *Environ Dev* 3, 52–64. doi: 10.1016/j.envdev.2012.04.002

Zhang, Z., Zhang, Q., Wang, T., Xu, N., Lu, T., Hong, W., et al. (2022). Assessment of global health risk of antibiotic resistance genes. *Nat. Commun.* 13:1553. doi: 10.1038/s41467-022-29283-8

Zhu, X., Guan, Y., Signore, A. V., Natarajan, C., DuBay, S. G., Cheng, Y., et al. (2018). Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet plateau. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1865–1870. doi: 10.1073/pnas.1720487115

Zhu, Z., Sun, Y., Zhu, F., Liu, Z., Pan, R., Teng, L., et al. (2020). Seasonal variation and sexual dimorphism of the microbiota in wild blue sheep (*Pseudois nayaur*). *Front. Microbiol.* 11:1260. doi: 10.3389/fmicb.2020.01260