Predicting Host-Interaction Traits in Probiotic Bacteria using Machine Learning and Functional Data

Gabriela de Quadros da Luz , Kristofer Stift Kappel , Rafaella Sinnott Dias , Fábio Pereira Leivas Leite , Frederico Schmitt Kremer *

Posted Date: 29 May 2026

doi: 10.20944/preprints202605.2061.v1

Keywords: probiotics; microbial genomics; functional genomics; annotation; screening; supervised machine learning

 doi:10.20944/preprints202605.2061.v1

Article

Predicting Host-Interaction Traits in Probiotic Bacteria Using Machine Learning and Functional Data

Gabriela de Quadros da Luz [1], Kristofer Stift Kappel [2], Rafaella Sinnott Dias [1], Fábio Pereira Leivas Leite and Frederico Schmitt Kremer [1,*]

[1] Omixlab Bioinformatics Laboratory, Center for Technological Development, Federal University of Pelotas, Pelotas, Rio Grande do Sul, Brazil

[2] Center for Computational Sciences (C3), Federal University of Rio Grande - FURG, Rio Grande do Sul, Brazil

* Correspondence: fred.s.kremer@gmail.com; Tel.: +55-53-99156-2706

Abstract

The identification of probiotic microorganisms is a complex process that requires integrating genomic data with evidence of functional health benefits. This study presents a machine learning framework, applied to a dataset of 1,185 probiotic and non-probiotic genomes, to predict seven key probiotic phenotypes (acid resistance, bile resistance, adhesion, antimicrobial activity, immunomodulation, antioxidant, and antiproliferative potential) using functional (COG) and metabolic (AntiSMASH) genomic annotations. We evaluated several algorithms, including CatBoost, Random Forest, and XGBoost, across three validation scenarios: cross-validation, leave-one-out, and train-test split. Models for antioxidant and antiproliferative activities achieved the highest predictive performance. Shapley Additive explanations (SHAP) analysis revealed specific metabolic signatures for each trait, such as the critical role of Type III polyketide synthases (T3PKS) in antioxidant activity and lipid metabolism, which contribute to antiproliferative effects. These results demonstrate that combining machine learning with functional analysis is a robust strategy for the in silico screening and selection of novel probiotic strains.

Keywords: probiotics; microbial genomics; functional genomics; annotation; screening; supervised machine learning

1. Introduction

Probiotics are defined as "Live microorganisms that, when administered in adequate amounts, confer health benefits to the host" [1]. These microorganisms can be present in the gastrointestinal and vaginal microbiota, as well as in foods and dietary supplements [2]. They can play a role in the production of vitamins, amino acids, and short-chain fatty acids, in ion absorption, in prophylaxis against various associated pathologies, in the development of both local and systemic immunity, and in reducing inflammation by acting on various signaling pathways [3].

However, for a microorganism to exert its beneficial effects as a probiotic, it needs to be resistant to bile acid and gastric acid, and to adhere to remain in place and colonize the mucosa [4]. Moreover, by remaining in the mucosa, it can exert its immunomodulatory effect, which is a major ally of the intestinal microbiota, as stimulating immune cells promotes modulation of the intestinal barrier, thereby maintaining its homeostasis [5]. The interaction between probiotics and the host is fundamental for maintaining the integrity of the intestinal barrier. This homeostatic process involves strengthening intercellular junctions and stimulating mucosal immune cells, thereby reducing intestinal permeability and preventing inflammatory processes [6]. The failure of these mechanisms

 doi:10.20944/preprints202605.2061.v1

is associated with pathologies such as cancer, allergies, intolerances, inflammatory diseases, diarrhea, and irritable bowel syndrome [7].

Interest in understanding the functions of probiotic microbial genes and their potential associations with human health has increased. Comparative genomic analysis is frequently used for this purpose [8]. However, genomic analysis based solely on sequence alignments and homology fails to capture the synergy between different genomic components [9]. The transition to advanced computational models is justified by the need to map multidimensional interactions between functional categories and biosynthetic genes (BGCs), which are the true determinants of biological efficacy, thereby enabling more robust prediction of the probiotic phenotype before experimental validation. In this context, Machine Learning emerges as a promising tool for large-scale genomic data analysis [10]. Machine learning algorithms can identify hidden patterns in high-dimensional datasets, enabling the classification of strains based on functional and metabolic profiles with high predictive accuracy [11,12].

In light of this, a better understanding of the mechanisms of action of probiotics is important. Consequently, machine learning tools stand out in in silico research for their ability to rapidly generate predictions from analyses of millions of experimental data points [12,13]. However, a gap remains in integrating functional genomic data to predict multiple biological properties simultaneously. The present study is justified by the need for computational models that not only classify microorganisms but also help to elucidate which genomic features are the primary determinants of probiotic activities.

To address this need, the present study proposes a combination of machine learning models to predict seven probiotic characteristics. To this end, functional and metabolic genomic annotation was performed on 782 probiotic and 403 non-probiotic genomes. Based on these results, various machine learning algorithms were applied to predict the acid- and bile-resistance, adhesion, antimicrobial activity, immunomodulation, and antioxidant and antiproliferative activities of the probiotics. The evaluation was conducted under three different validation scenarios: Stratified Cross-Validation, Leave-One-Out (LOO), and the classic Train-Test split (Holdout).

2. Materials and Methods

2.1. Data Acquisition

The genomes of the probiotic and non-probiotic microorganisms used as references were obtained directly from the National Center for Biotechnology Information (NCBI) Datasets platform (https://www.ncbi.nlm.nih.gov/datasets/genome/) in GenBank format (.GBK). The total number of probiotic genomes was 782, and the number of non-probiotic genomes was 403. Dataset selection was based on reference databases previously described in the literature. Positive examples (probiotic microorganisms) were obtained from the Probio-Ichnos database, which catalogs microorganisms with in vitro-proven probiotic properties [14]. From this set, the seven phenotypic characteristics of interest (acid resistance, bile resistance, adhesion, antimicrobial activity, immunomodulation, and antioxidant and antiproliferative potential) were extracted and used as labels in the predictive models. For the negative examples (non-probiotics), the genomes were selected from the dataset of inactive instances of the iProbiotics platform, ensuring the inclusion of strains with no reported probiotic activity [15].

2.2. Functional and Metabolic Annotation

The functional annotation of the protein sequences of the microorganisms used in this study was performed using COGClassifier 1.0.5 (https://pypi.org/project/cogclassifier/). The sequences were classified into functional categories using the Clusters of Orthologous Genes (COG) database via similarity alignments. The completeness of the genomes used for training the models was evaluated using the BUSCO and Check2 programs.

 doi:10.20944/preprints202605.2061.v1

For the analysis of biosynthetic potential, AntiSMASH 7.0.0 (Antibiotics and Secondary Metabolite Analysis Shell) was used to identify and annotate secondary metabolite biosynthetic gene clusters (BGCs). AntiSMASH identifies genomic regions encoding enzymes involved in the production of bioactive compounds by detecting conserved protein domains and comparing them with known clusters [16].

Based on the results from both tools, a structured dataset was constructed. The data derived from COGClassifier and AntiSMASH were treated as binary presence/absence variables per genome. This dataset included the functional categories and the secondary metabolite profile, along with labels for the seven probiotic activities of interest (acid resistance, bile resistance, adhesion, antimicrobial, immunomodulation, antioxidant, and antiproliferative). Table 1 presents the seven characteristics of interest as well as the class distribution for each characteristic.

Table 1. Target labels and class distribution within the genomic dataset. This table lists the seven probiotic phenotypes selected as prediction targets: acid resistance, bile resistance, adhesion, antimicrobial activity, immunomodulation, antioxidant, and antiproliferative potential.

Labels	classes	Sample	percentage (%)
Resistant to acid	True	321	41.1
	False	460	58.9
Resistant to bile	True	358	45.8
	False	423	54.2
Adhesion	True	272	34.8
	False	509	65.2
Antimicrobial	True	428	54.8
	False	352	45.2
Immunomodulation	True	240	30.7
	False	541	69.3
Antioxidant	True	99	12.6
	False	682	87.4
Antiproliferative	True	35	4.4
	False	746	95.6

Data preprocessing and the development of predictive models were implemented in Python, using the Scikit-Learn (https://scikit-learn.org/stable/), Pandas (https://pandas.pydata.org/), and NumPy (https://numpy.org/) libraries. For better data analysis and evaluation, each of the 7 biological labels was treated as an independent binary outcome, generating 7 distinct datasets.

Each dataset was independently divided into training (80%) and testing (20%) subsets using the train_test_split function. This split was stratified by outcome class, ensuring that the original proportions of positive and negative samples were maintained in both sets. To ensure the reproducibility of the experiment, the random state was fixed (random_state=42).

To mitigate the bias caused by class imbalance, five resampling strategies were applied exclusively to the training data: (1) Undersampling, (2) Oversampling with SMOTE, (3) Tomek Links associated with Undersampling, (4) SMOTE combined with Tomek Links, and (5) the combination of Undersampling, SMOTE, and Tomek Links. The test set remained untouched and in its original imbalanced state to reflect a real-world prediction scenario.

Seven machine learning algorithms, covering different predictive paradigms, were evaluated: Logistic Regression (linear), Decision Tree, and tree ensemble-based approaches (Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost). All models employed maintained their default hyperparameters as loaded from Scikit-Learn, XGBoost, LightGBM, and CatBoost to establish a baseline performance and evaluate the algorithms' native predictive capability.

The evaluation of the models was conducted on the training set using two complementary strategies: 5-fold Stratified Cross-Validation, with prior data shuffling (shuffle=True,

 doi:10.20944/preprints202605.2061.v1

random_state=42), and Leave-One-Out. The performance of the models was evaluated using accuracy, recall, precision, and F1-score.

To assess the impact and contribution of each biological feature on the predictions, the SHAP (Shapley Additive exPlanations) approach was applied, based on the calculation of Shapley values. The F1-score results from the validation step were used as the primary criterion to select the best predictive model for each outcome, and the resulting models were subjected to explanatory analysis using SHAP.

3. Results

The performance metrics for the optimal models across each label are presented in Tables 2, 3, and 4. Three distinct validation scenarios were employed for all models: cross-validation (Table 2), leave-one-out (Table 3), and train-test split (Table 4). A general observation was that both model performance and the efficacy of the applied balancing methods varied across labels.

The cross-validation assessment yielded results characterized by balanced performance across all metrics. The best models, identified through SHAP analysis, were CatBoost, Random Forest, and XGBoost, which were utilized in conjunction with specific balancing strategies, notably SMOTE and Tomek Links, and SMOTE combined with Oversampling. This suggests that the use of class-balancing techniques played a significant role in model evaluation. Among the superior results, the Antioxidant label, classified by XGBoost with SMOTE and Tomek Links, demonstrated the highest overall performance (Accuracy = 0.92 ± 0.02; F1-score = 0.92 ± 0.02). Subsequently, the Antiproliferative label, using Random Forest combined with SMOTE and Oversampling, achieved meaningful metrics (Accuracy = 0.97 ± 0.01). Meanwhile, the Acid Resistant, Bile Resistant, Adhesion, and Antimicrobial labels showed intermediate metric values (Accuracy between 0.74 and 0.77) and good mutual consistency. It is worth noting that, despite high accuracy, in vitro phenotypic validation remains necessary, as none of the characteristics predicted from genomic information can be considered fully accurate.

Regarding leave-one-out validation, the results showed lower metric values than those obtained with cross-validation. It was noted that the best results for the Antioxidant and Antiproliferative labels were maintained even with the leave-one-out approach, with accuracies exceeding 0.94 and F1 scores ranging from 0.47 to 0.49. Similar to the other labels, Acid Resistant, Bile Resistant, and Antimicrobial demonstrated reduced performance, with F1-score ranging from 0.38 to 0.41. This indicates increased difficulty in prediction for these specific cases. This difficulty and the discrepancy relative to the cross-validation results are attributed to the greater complexity of the leave-one-out approach, which, given the high variability of results across labels, suggests that the models are sensitive to minor data perturbations.

In the train-test split scenario, the results were comparable to those obtained with the other validations assessed. Among the top-performing models in this group of results were LightGBM, CatBoost, and Random Forest, with their superiority varying by label. The LightGBM model yielded the best results for the Antiproliferative (Accuracy = 0.96; F1 = 0.72, without balancing) and Antioxidant (Accuracy = 0.89; F1 = 0.70, using the SMOTE and Tomek Links combination) labels, respectively. Unlike the other validation scenarios analyzed, certain labels performed better when no balancing was applied, suggesting that the test set maintained more favorable proportions under this split.

Furthermore, the natural class distribution presented in Table 1 helps contextualize the observed differences among the three evaluation scenarios. Labels with more pronounced imbalance, such as Antioxidant (12.6% positive) and Antiproliferative (4.4% positive), typically require more aggressive balancing techniques, which helps explain the favorable performance achieved with SMOTE, Tomek Links, and Oversampling. This is because these balancing methods generate sufficient synthetic examples to represent the minority class during training adequately. This accounts for the prominence of these two labels across all three evaluation scenarios.

 doi:10.20944/preprints202605.2061.v1

To characterize the functional drivers of the predicted probiotic traits, an explainability analysis was conducted using SHAP (Shapley Additive exPlanations) on the models validated via the train-test split. This methodology facilitated the systematic identification of the most relevant COG (Clusters of Orthologous Genes) and metabolic categories associated with each phenotype. The results are visualized through two complementary perspectives: the SHAP summary bar plot, which ranks features by their average absolute contribution to the model's decision-making (Figure 1), and the SHAP beeswarm plot, which illustrates the distribution and directionality of these impacts across individual genomic samples (Figure 2). Collectively, these visualizations reveal the specific metabolic and functional signatures (ranging from acid-resistance signaling to lipid metabolism in antiproliferative activity) that define the probiotic genotype-to-phenotype relationship.

For the Acid-Resistant label, categories associated with signal transduction mechanisms, carbohydrate metabolism, cell motility, defense mechanisms, and gene expression were prominent, exhibiting the highest average SHAP values (Figure 1a). An examination of the beeswarm plot confirms that, among the categories highlighted by the average SHAP values, signal transduction mechanisms are crucial for identifying acid-resistance activity in a probiotic microorganism (Figure 2a). Even when present at low concentrations in samples, it is strongly correlated with a positive impact on the prediction of this label. This is followed by characteristics related to carbohydrate metabolism, present in a significant number of samples. Regarding the characteristic associated with cell motility, SHAP's impact analysis indicates that a higher count of motility-related genes in the microorganism correlates with lower acid resistance (Figure 2a). Furthermore, factors associated with defense and transcription mechanisms show a direct relationship between the number of samples possessing these characteristics and acid resistance. However, the "defense mechanisms" category exhibits a slightly greater impact on that prediction, consistent with the order of importance presented in Figures 1 and 2.

A similar behavior was observed in the SHAP results for the bile resistance probiotic characteristic, where many of the categories highlighted by the SHAP values were recurrent, but with specific biological behavior for this type of stress (Figure 1b). In this context, defense mechanisms stand out as the most relevant characteristic for predicting this label, indicating that, for the model, a low abundance of genes related to cellular defense has a strong positive effect on the prediction (Figure 2b). The second highlighted characteristic was nucleotide transport and metabolism, with a significant concentration of samples on the right side of the beeswarm plot and red dots on the left, suggesting that the presence of a few genes associated with nucleotide transport and metabolism favors classifying the microorganism as Bile Resistant. Conversely, a high concentration of these genes contributes to the microorganism's designation as not bile-resistant. In contrast, for the cell wall, membrane, envelope biogenesis category, the presence of genes linked to cell structure is crucial for the microorganism to exhibit this characteristic.

Regarding the results obtained from the analysis of the adhesion characteristic in probiotics, mobile genetic elements (mobilome), defense mechanisms, cellular biogenesis, replication and repair processes, and cellular structure are among the most relevant categories for predicting this characteristic (Figure 1c). Among these, mobile genetic elements are present in the majority of samples. However, an analysis of the impact of this characteristic on the model's prediction suggests that, although a significant concentration of samples indicates a high contribution to this prediction, some red dots indicate that a low abundance of these genes is sufficient to impair cell adhesion (Figure 2c). Conversely, the plot demonstrates a direct relationship between the presence of genes related to cellular defense and adhesion ability, such that a small number of these genes are sufficient to predict this label. Furthermore, regulatory and metabolic processes contribute to the adhesion characteristic, with categories related to structure, cell cycle control, and energy production among those positively influencing this prediction (Figure 2c).

While adhesion necessitates an investment in surface elements and genomic plasticity, the analysis of antimicrobial activity reveals a shift in focus toward basal metabolism and specialized biosynthesis (Figure 1d). This is supported by a strong positive correlation between nucleotide

 doi:10.20944/preprints202605.2061.v1

transport and metabolism and antimicrobial activity, with the model predicting that microorganisms with a greater number of genes in this category will exhibit greater antimicrobial activity (Figure 2d). Similarly, the presence of the secondary metabolites biosynthesis, transport, and catabolism category, in which the existence of a few genes related to the ability to synthesize and transport secondary metabolites out of the cell positively aids the prediction of this characteristic. Conversely, categories related to lipid transport and metabolism, and genes related to cell replication, recombination, and repair, demonstrated a different influence: the absence or low abundance of genes associated with these categories may be linked to lower antimicrobial activity.

Unlike the strong metabolic influence observed in predicting antimicrobial function, immunomodulation shows greater dependence on ribosomal biogenesis, cellular structure, carbohydrate metabolism, and defense mechanisms (Figure 1e). Among the categories that reinforced the classification and characterization of the immunomodulation label, translation, ribosomal structure, and biogenesis stood out as the most influential factors in this prediction, with a lower count of genes in this category associated with a higher probability that the microorganism is immunomodulatory (Figure 2e). Conversely, the presence of numerous genes involved in carbohydrate transport and metabolism is a positive predictor of this probiotic characteristic.

For the label related to antioxidant activity, the T3PKS category presented the highest average SHAP contribution in the evaluated scenario, followed by categories such as cell wall, membrane and envelope biogenesis, carbohydrate transport and metabolism, and transcription (Figure 1f). Consequently, the assessment of the impact of these categories on the model's predictions indicates that the presence of genes associated with type III polyketide synthases (T3PKS) is strongly correlated with the model's ability to predict antioxidant activity, with the effect being negative when present in high quantities and positive in low quantities. Thus, cell lines classified by the model as antioxidant tend to have a low abundance of these specific enzymes. Conversely, in cell structure, the low abundance of genes associated with this category favors antioxidant activity, whereas a high abundance of these genes is detrimental. The positive prediction of this label appears to be supported by a directed investment in nucleotide transport and metabolism and by intense transcriptional activity, whose high values directly contribute to the classification of samples as antioxidant (Figure 2f).

Finally, antiproliferative activity is strongly characterized by a metabolic signature centered on the economy of lipid and structural pathways (Figure 1g). According to the SHAP impact results, the lipid transport and metabolism category shows the highest average relevance for prediction, with an inverse relationship between the abundance of these genes and the model's classification of the microorganism as antiproliferative. As with lipid metabolism, the model associates the presence of a small number of cell-structure-related genes with a positive impact on prediction. Furthermore, the presence of mobile genetic elements activity acts as a secondary positive predictor, demonstrating that the ability to inhibit cell proliferation is a specialized trait in probiotic microorganisms (Figure 2g)

 doi:10.20944/preprints202605.2061.v1

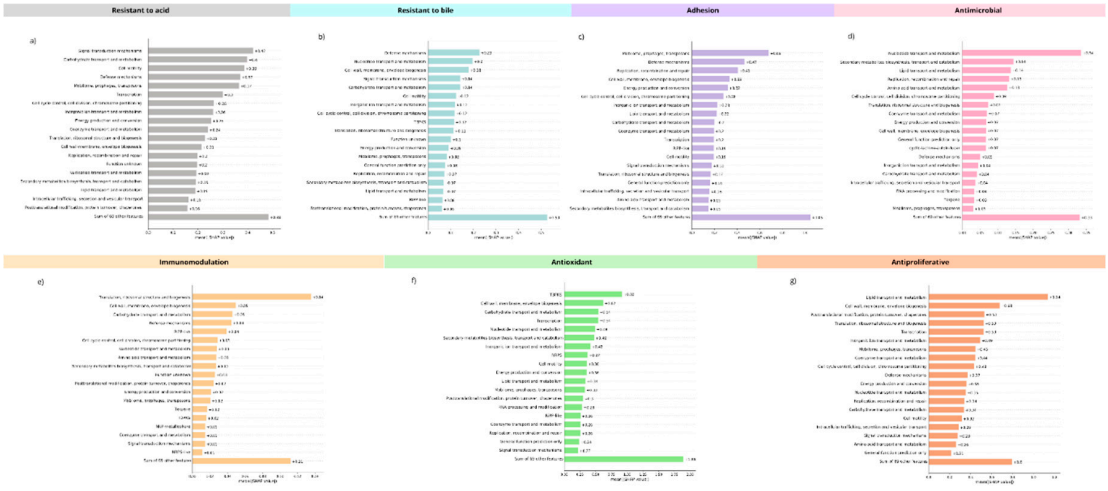

Figure 1. SHAP Interpretability Analysis for Resistance to acid, Resistance to bile, Adhesion, Antimicrobial, Immunomodulation, Antioxidant, and Antiproliferative. Bar plots show the average impact (mean absolute SHAP value) of the most relevant COG functional categories on the model's prediction of acid resistance.

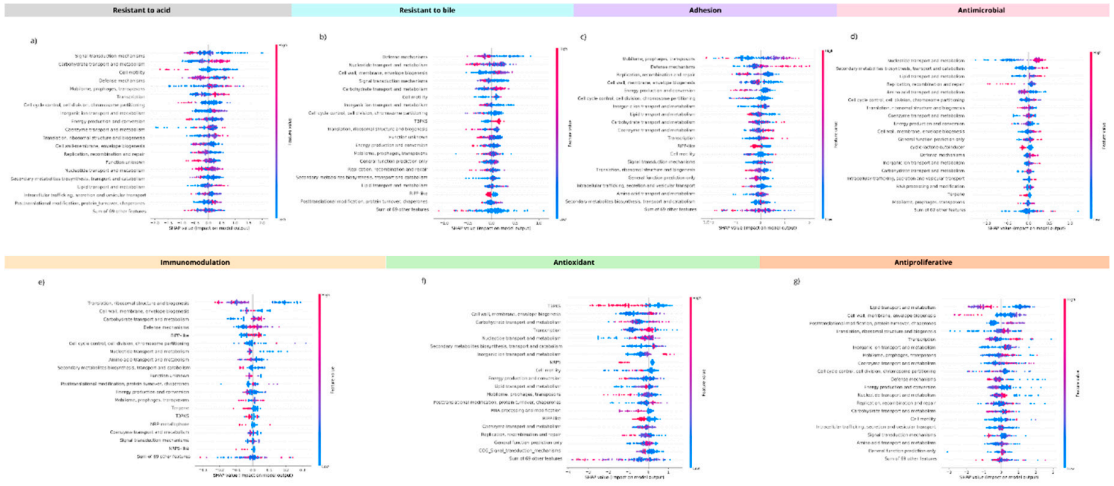

Figure 2. SHAP Interpretability Analysis for Resistance to acid, Resistance to bile, Adhesion, Antimicrobial, Immunomodulation, Antioxidant, and Antiproliferative using the Beeswarm Plot. Detailed view showing how feature values influence the model's outputs. Each point represents an individual genome, with its position on the x-axis indicating the SHAP value (impact on prediction): points to the right increase the probability of the trait, while points to the left decrease it. The color scale denotes relative feature values, with red indicating high gene abundance and blue indicating low abundance. The vertical thickness of the "swarm" illustrates the density of samples at a specific impact level.

Table 2. Optimal model performance metrics achieved through cross-validation (CV). Performance results are presented for the best-performing machine learning algorithm and class-balancing strategy for each probiotic label during k-fold cross-validation. Metrics include Accuracy, Precision, Recall, and F1-score, reflecting the models' ability to generalize across partitioned training data. CB = Catboost; RF = Random Forest; XGB = XGBoost.

Label	Best Model	Balancing	Accuracy	Precision	Recall	F1-score
Resistant to acid	CB	SMOTE + Tomeklinks	0.77 ± 0.02	0.75 ± 0.03	0.77 ± 0.04	0.76 ± 0.02

 doi:10.20944/preprints202605.2061.v1

Resistant to bile	CB	SMOTE + Tomeklinks	0.75 ± 0.04	0.73 ± 0.06	0.73 ± 0.02	0.73 ± 0.04
Adhesion	RF	SMOTE + Tomeklinks	0.77 ± 0.03	0.76 ± 0.03	0.78 ± 0.02	0.77 ± 0.02
Antimicrobial	CB	SMOTE + Tomeklinks	0.74 ± 0.02	0.72 ± 0.03	0.70 ± 0.02	0.71 ± 0.02
Immunomodulation	RF	SMOTE + Tomeklinks	0.82 ± 0.01	0.81 ± 0.01	0.84 ± 0.02	0.82 ± 0.01
Antioxidant	XGB	SMOTE + Tomeklinks	0.92 ± 0.02	0.89 ± 0.01	0.94 ± 0.03	0.92 ± 0.02
Antiproliferative	RF	SMOTE + Oversampling	0.97 ± 0.01	0.96 ± 0.02	0.97 ± 0.01	0.97 ± 0.01

Table 3. Model performance evaluation using Leave-One-Out (LOO) validation. This table presents the metrics obtained when each sample is used as the test set, with the remaining samples serving as the training set. Metrics include Accuracy, Precision, Recall, and F1-score, which reflect the models' ability to generalize to new observations. CB = Catboost; LR = Logistic Regression; RF = Random Forest; XGB = XGBoost.

Label	Best Model	Balancing	Accuracy	Precision	Recall	F1-score
Resistant to acid	CB	SMOTE + Oversampling	0.72	0.38	0.38	0.38
Resistant to bile	LR	Tomeklinks + Undersampling	0.68	0.39	0.39	0.39
Adhesion	CB	SMOTE + Oversampling	0.76	0.40	0.40	0.40
Antimicrobial	GB	Without Balancing	0.67	0.41	0.41	0.41
Immunomodulation	CB	SMOTE + Oversampling	0.81	0.42	0.42	42
Antioxidant	XGB	SMOTE + Tomeklinks	0.94	0.47	0.47	47
Antiproliferative	XGB	SMOTE + Oversampling	0.97	0.49	0.49	0.49

Table 3. Performance metrics for the best models in the Train-Test Split scenario. The results summarized here correspond to a traditional random split of the data into training and testing sets. Metrics include Accuracy, Precision, Recall, and F1-score, reflecting the models' ability to generalize beyond the training dataset, for a test (holdout) set. LGBM = LightGBM; CB = Catboost.

Label	Best Model	Balancing	Accuracy	Precision	Recall	F1-score
Resistant to acid	LGBM	Without Balancing	0.71	0.70	0.70	0.70
Resistant to bile	CB	Without Balancing	0.76	0.76	0.76	0.76
Adhesion	CB	SMOTE + Tomeklinks	0.73	0.71	0.69	0.69
Antimicrobial	CB	Without Balancing	0.77	0.77	0.76	0.76
Immunomodulation	CB	Tomeklinks + Undersampling	0.68	0.65	0.67	0.65
Antioxidant	LGBM	SMOTE + Tomeklinks	0.89	0.77	0.66	0.70
Antiproliferative	LGBM	Without Balancing	0.96	0.74	0.70	0.72

4. Discussion

The classification of microorganisms as probiotics is widely recognized as a complex process that goes beyond the simple presence of specific genomic sequences, requiring that these microorganisms be alive and capable of conferring health benefits to the host when administered in appropriate doses [17]. The inherent complexity of these effects, which result from the integrated action of multiple biological capabilities, significantly challenges predictive models based solely on genomic data. Therefore, the adoption of supervised machine learning models, as proposed in this study, is a promising strategy for investigating the genomic patterns associated with these properties [18].

The SHAP interpretability analysis contributes to this reflection by highlighting that, although k-mer-based models have yielded satisfactory results, their biological interpretation remains limited in the absence of functional information. This result is particularly relevant to the discussion, as it reinforces that identifying sequence patterns in isolation does not necessarily translate into a mechanical understanding of probiotic properties [13]. Given that host health benefits are associated with abilities such as resistance to stomach acid and bile salts, adhesion to the intestinal epithelium,

 doi:10.20944/preprints202605.2061.v1

and the production of antimicrobial compounds, it is evident that integrating genomic and functional data is essential for a biologically consistent interpretation [19,20].

The ability to withstand adverse conditions, such as those in the upper gastrointestinal tract, is a prerequisite for the microorganism to reach the intestine and exert its probiotic activity. At the same time, proper adhesion promotes competitive exclusion of pathogens and enhances the beneficial effects on the host [21]. Furthermore, mechanisms such as the production of bacteriocins, stimulation of the host immune system, and maintenance of intestinal barrier function represent complementary strategies for combating pathogen infection [22]. Further evidence suggests that these microorganisms may also help maintain host health under specific pathological conditions, including by controlling the proliferation of cancerous cells and increasing resistance to oxidative stress [23,24].

Given the relevance of these properties, the results presented in this study are particularly important, as they allow the identification of the characteristics that determine each of these abilities. However, the findings also highlight limitations inherent in approaches based exclusively on k-mers, reinforcing that the prediction of probiotic potential must be interpreted critically and in an integrated manner, considering both computational performance and the biological significance of the identified characteristics, as explored in subsequent sections.

4.1. Acid Resistance

The model's prediction of acid resistance characteristics highlighted signal transduction mechanisms, carbohydrate metabolism, cell motility, and defense mechanisms as the most relevant categories (Figure 1a). While these findings are biologically plausible, reflecting processes of stress adaptation and cellular homeostasis, they also suggest potential informational redundancy among predictors, as the importance of variables is distributed across broad and functionally overlapping categories. This moderate impact of multiple clustered features reinforces the multifactorial nature of the phenotype. It emphasizes the need for further validation to confirm the robustness and independence of these variables across different datasets. This perspective is corroborated by studies like Ma et al., who applied SHAP-interpretable algorithms to 16S rRNA data to investigate associations between the gut microbiome and atopic dermatitis, using a train-test split for validation.[25] The authors showed that, although *Bifidobacterium* sp. emerged as a key predictor, the model's contribution was distributed across various taxa. This illustrates how explainable approaches can detect coherent biological signals even in complex, multifactorial scenarios.

The ability of probiotics to withstand acidic conditions is crucial to their function in the gastrointestinal tract. To successfully colonize and exert their beneficial effects, these microorganisms must survive the extremely acidic conditions of the stomach, where hydrochloric acid and a very low pH severely threaten their viability [26]. The SHAP results demonstrated that pathways related to signal transduction, carbohydrate transport and metabolism, cell motility, and defense mechanisms contributed significantly to the classification of the acid-resistant phenotype, suggesting that these predictors may reflect biological mechanisms involved in bacterial adaptation to acid stress (Figure 2a). However, the specific functional participation of these pathways still requires experimental validation. Furthermore, for a strain to be recognized as probiotic, it must meet fundamental criteria, such as safety for human consumption and the ability to reach the intestine in sufficient quantity to multiply and establish adequate colonization. In addition, adverse physiological conditions, such as pepsin in the stomach and continuous intestinal flow, which hinder mucoadhesion, can further compromise the viability and persistence of these microorganisms [27]. A multifactorial assessment is crucial for selecting and validating potential probiotics, integrating computational findings with classical functional criteria. A central requirement for these candidates is survival during gastrointestinal transit, which necessitates tolerance to low gastric pH and bile acid effects in the intestine. [28]

 doi:10.20944/preprints202605.2061.v1

4.2. Bile Resistance

SHAP values indicate a multifactorial pattern for the bile resistance trait, highlighting the contributions of defense mechanisms, nucleotide metabolism, cell wall/membrane biogenesis, and signal transduction to the model's output trained via a train-test split (Figure 1b). From a biological perspective, these findings are consistent with the detergent effect of bile salts, which promote membrane damage, protein destabilization, and metabolic disturbances, requiring integrated responses of structural repair and cellular regulation. [29] The observed predominance of broad functional categories raises a crucial methodological consideration: the apparent significance might stem from gene co-occurrence or structural correlations among functionally related pathways, rather than reflecting genuine causal independence among the predictors. Furthermore, the moderate and dispersed nature of the feature impacts suggests that the model effectively captures a significant yet diffuse biological signal characteristic of complex adaptive phenotypes. However, this characteristic also heightens the risk of informational redundancy due to the multitude of contributing features. [30] Understanding the mechanisms of bile resistance in the intestinal environment is relevant because it directly reveals the environment's complexity. Concerning probiotic microorganisms, the need for combined functional criteria, and not just isolated characteristics, to predict their resistance and performance in the gastrointestinal environment is highlighted. [31] The conversion of primary bile acids into secondary metabolites by the gut microbiota highlights a complex functional axis between microorganisms and the host, in which alterations in microbial composition can directly affect the profile and activity of these metabolites. This process is not passive, as it depends on specific enzymatic repertoires, suggesting that structural changes in the microbiota lead to meaningful functional variations in intestinal metabolism. [32]

In parallel, bile acids themselves exert strong selective pressure on the microbiota due to their detergent and antimicrobial properties, acting as a limiting factor for bacterial survival. [33] Bile resistance is a crucial attribute for microorganisms with probiotic potential, as it dictates their survival and functional efficacy within the intestine. This resistance, however, should be viewed not just as an individual adaptation, but also as a factor that can influence the overall balance (or pool) of bile acids. This highlights the bidirectional nature of this interaction and emphasizes the necessity for more stringent functional criteria when selecting probiotic microorganisms. [34]

4.3. Adhesion

Following similar logic, the prediction of adhesion characteristics also proved to depend on multiple factors. However, in this model, functions related to the mobilome (prophages and transposons), defense mechanisms, replication/recombination/repair, and cell wall/membrane biogenesis showed the greatest average contribution (Figure 1c). The beeswarm plot demonstrates that higher values in these categories tend to increase the predicted probability of adhesion, suggesting that genomic plasticity, structural integrity, and stress-response capacity are important determinants of cell-surface interaction (Figure 1c). Additionally, metabolic functions, such as energy metabolism, lipid and carbohydrate transport, and inorganic metabolism, also presented a consistent contribution. Taken together, these results indicate that adhesive competence does not depend exclusively on structural components of the cell surface, but also on an integrated physiological state associated with cellular adaptation, membrane maintenance, and metabolic regulation [29,35]. The capacity to adhere to the intestinal mucosa is one of the fundamental characteristics for a microorganism to be classified as a probiotic. This property enables colonization of the gastrointestinal tract. It promotes interactions between probiotic strains and the host, playing an important role in modulating antagonism against pathogenic microorganisms and acting on the immune system. Furthermore, the presence of adhesins facilitates the attachment of probiotic cells to the mucosal layer, thereby promoting their persistence and activity within the intestinal environment [36]. Recent studies demonstrate that the adhesion capacity of probiotic strains, especially those of the genus Lactobacillus, is one of the main functional criteria for colonization potential and the competitive exclusion of pathogens in the gastrointestinal tract. In this context, Xing et al. observed

 doi:10.20944/preprints202605.2061.v1

that *Lactobacillus acidophilus* AD125 exhibited high adhesion to Caco-2 cells, which was associated with surface hydrophobicity and autoaggregation. These findings suggest that the physicochemical properties of the bacterial surface play a determining role in the interaction with the intestinal epithelium [37]. In a complementary way, Zawistowska-Rojek et al. suggested that adhesion varies significantly among species and strains of Lactobacillaceae, including *Lactobacillus acidophilus*, *Lactiplantibacillus plantarum*, and *Lacticaseibacillus rhamnosus*, indicating that the adhesive phenotype is strongly dependent on the strain's lineage and origin. [38] Furthermore, a review study conducted by Zheng et al. highlights that specific molecular mechanisms, such as surface layer proteins (SLPs), mucin-binding proteins, and other cell wall-associated adhesins, are determinants of the bacteria-host interaction, modulating not only physical adhesion to the epithelium but also transient persistence and local immunological communication. [39] Taken together, these findings support the idea that probiotic adherence is a multifactorial phenomenon, dependent on both the strain's specific structural characteristics and the conditions of the gut microbiota, thereby reinforcing its importance in the rational selection of microorganisms with probiotic potential. It is also known that a highly important characteristic for adhesion is the cell wall of Lactobacillus sp. Research by Kong et al. suggested that the S-layer protein (SLP) of *Lactobacillus acidophilus* CICC 6074 possesses distinct functional domains directly involved in cell wall structure and adhesion [40]. The authors found that the C-terminal domain of SLP anchors to the cell wall via interactions with teichoic acids, ensuring proper exposure on the bacterial surface. At the same time, the N-terminal regions play a significant role in adhesion to human HT-29 epithelial cells. These findings demonstrate that structural components of the cell wall, especially surface layer proteins, not only confer stability on the bacterium but also actively mediate adhesion to the host.

4.4. Antimicrobial

The results of the SHAP analysis for predicting antimicrobial activity reveal a hierarchy in which basic cellular metabolism takes center stage. In this scenario, the model highlights characteristics related to nucleotide transport and metabolism, as well as pathways for secondary metabolites, lipids, and replication and repair processes (Figure 1d). Although the relevance of secondary metabolites is consistent with antimicrobial activity, the emphasis on broad metabolic functions suggests that the model may be identifying a general profile of greater cellular activity or adaptation, rather than specific mechanisms for the production of antimicrobial compounds. Furthermore, the cumulative contribution of several other variables indicates that the antimicrobial phenotype is likely the result of multiple combined factors, making it difficult to pinpoint a single determinant mechanism. [41] Antimicrobial agents, whether synthetic or natural, are essential for inhibiting or eliminating pathogenic microorganisms that cause acute or chronic infections. [42] In this context, the role of probiotics and their metabolites is notable, as they exhibit antimicrobial activity by inhibiting the growth of pathogenic microorganisms. This effect may be related to the production of organic acids and bacteriocins, as well as to competition for nutrients and adhesion sites. [43] *Lactobacillus*, one of the most widely studied probiotic genera, produces antimicrobial substances, such as hydrogen peroxide, lactic acid, and bacteriocins, that play a fundamental role in promoting microbial homeostasis. [44] Additionally, their ability to produce organic acids lowers the pH of the environment to which they are attached and, consequently, creates unfavorable conditions for the growth of pathogenic microorganisms. Furthermore, this genus is recognized for its ability to stimulate and modulate the immune system, inducing the production of cytokines and chemokines and contributing to the maintenance of the host's immune balance. [45] These local and systemic effects do not occur in isolation, but are part of a broader set of antimicrobial mechanisms described in recent literature. Published studies reinforce that the antimicrobial activity of probiotics involves multiple complementary strategies. A review by Costa et al. indicates that probiotics from the genera *Lactobacillus*, *Bifidobacterium*, *Streptococcus*, and *Saccharomyces* consistently demonstrate antimicrobial activity in in vitro assays. This activity is primarily attributed to the production of organic acids, bacteriocins, and other bioactive metabolites.[46] More specifically, Al-Shamiri et al.'s experimental

 doi:10.20944/preprints202605.2061.v1

study indicated *that Lactobacillus rhamnosus* strains produced biosurfactants. These biosurfactants were effective at disrupting *Acinetobacter baumannii* biofilms and directly damaging the pathogen's cell membrane. [47] Aghamohammad et al. also demonstrated that postbiotics derived from *Lactiplantibacillus plantarum* and *Ligilactobacillus salivarius* acted synergistically with antibiotics against multidrug-resistant *Klebsiella pneumoniae*. [48] Thus, it is observed that the antimicrobial effects of probiotics range from acidification of the medium and immunomodulation to the production of specific metabolites and structural interference in biofilms, indicating a multifactorial mechanism that depends on the genus, species, and strain evaluated.

4.5. Immunomodulation

The SHAP analysis of immunomodulation highlighted the role of factors associated with ribosomal biogenesis, cellular structure, carbohydrate metabolism and transport, and defense mechanisms. This finding is consistent with the scientific literature, as probiotic-induced immunomodulation relies on the synthesis of structural components and the secretion of bioactive compounds [49]. These act as effectors in communication with the host, triggering signaling cascades that culminate in the modulation of cytokine synthesis and release by immune cells, thus establishing the profile of gastrointestinal immunomodulation. These effectors communicate with the host, initiating signaling cascades that modulate cytokine synthesis and release by immune cells. This process consequently determines the profile of gastrointestinal immunomodulation [50]. In addition, cell movement and the presence of specialized organelles are crucial for host-microorganism interaction, and this biological basis is consistent with the trait importance analysis conducted in this study. Specifically, the analysis in the split train-test scenario reveals a marked dominance (average SHAP value +0.14) of the category "Translation, ribosomal structure and biogenesis". This suggests that this characteristic contributed almost exclusively to the model's prediction of this class (Figure 1e). However, this concentration of importance demonstrates a specific sensitivity of the model to this scenario and data distribution, which affects its performance. This is evidenced by the fact that, compared with the other evaluations studied, the best model selected to predict this label, using a train-test split, yielded low values across the evaluated metrics (Table 4). However, when analyzing the metric values achieved by the best models, it is evident that applying cross-validation to the model proved more effective, reaching an F1-score of 0.82 (Table 2). With the use of leave-one-out, the value dropped to 0.42 (Table 3).

4.6. Antioxidant

The analysis of antioxidant potential, specifically through SHAP results, highlights the critical role of Type III Polyketide Synthases (T3PKS). T3PKS was identified as the most impactful predictor for this label, underscoring the biological significance of polyketides in counteracting oxidative stress (Figure 1f). These polyketides are instrumental in the biosynthesis of complex, branching secondary metabolites, which in turn yield diverse bioactive substances with proven antioxidant, antimicrobial, and immunomodulatory properties. [20] The presence of these enzymes in probiotics is associated with a high affinity between the microorganism and the host organism, thereby characterizing the probiotic as potent. [51] In addition, features such as cell wall, membrane, envelope biogenesis, carbohydrate transport, and metabolism. They appear to be of similar importance, demonstrating that cellular structure and biochemistry are also involved in this process.

The stability of the category hierarchy reflected in the SHAP values for this label suggests that the model identified a robust phenotype less susceptible to sampling variation. This consistency is highlighted when observing the F1-score values achieved by the best models in each scenario, where, in the train-test split and cross-evaluation settings, the values were 0.70 and 0.92, respectively. These values validate the effectiveness of the machine learning approach in identifying probiotic microorganisms with antioxidant capacity based on their genetic content.

 doi:10.20944/preprints202605.2061.v1

4.7. Antiproliferative

For the antiproliferative label, the feature importance analysis revealed one of the most stable profiles among all the labels evaluated. In the category Lipid transport and metabolism, it played an absolute leading role, suggesting a link between lipid composition and/or the ability to secrete derived metabolites and the inhibition of cell proliferation (Figure 1g). This indication aligns with the literature, which demonstrates that the conversion of dietary fatty acids into conjugated linoleic acid (CLA) by probiotic bacteria is a primary mechanism for inducing apoptosis and cell-cycle arrest in cancerous cells. [52] This effect is also achieved through modulation of lipid transport, in which probiotics and their derivatives regulate lipid accumulation in the membranes of tumor cells and pathogenic microorganisms. [53] Furthermore, the antiproliferative response can be achieved using structural or intracellular components, such as cytoplasmic and cell wall extracts (Figure 2g) [54]. Reinforcing the compatibility of the most important categories presented by the models with the literature.

In general, SHAP results showed stable importance distributions across categories. However, despite maintaining the logic across the selected categories, there was a discrepancy in the importance value for the lipid transport and metabolism category, indicating potential sensitivity of the model in this scenario (Figure 1g). It is worth noting that among the labels studied in this article, antiproliferative activity was the most imbalanced, with only 4.4% of the data as true (Table 1), and that the best model for the train-test split was selected without balancing (Table 4). In other words, since there was no balance between the classes, the hypothesis is that the model identified this characteristic as a robust predictor capable of recognizing the minority class without the need to expand the data synthetically.

5. Conclusions

This study demonstrates that integrating machine learning algorithms with functional genomic analysis provides a capable and scalable strategy for predicting complex probiotic phenotypes directly from sequence data. The developed models exhibited high predictive performance, particularly for antioxidant and antiproliferative activities, achieving superior accuracy and F1-score. Using SHAP explainability analysis, specific metabolic signatures and shared functional categories were identified for each probiotic trait, offering deeper insights into the underlying biological mechanisms.

The present study contributes to the understanding of genomic data, suggesting that these data can be used efficiently for the screening and selection of new probiotic strains, thereby reducing exclusive reliance on experimental assays and accelerating the prospection of new microorganisms. Despite the promising results, the predictions should be experimentally validated to confirm the association between genomic data and probiotic phenotypes. From a future perspective, the data and models developed in this study serve as a basis for building software that predicts probiotic characteristics from genomic data. Furthermore, integrating machine learning tools and functional analyses into software is expected to assist in the screening, selection, and prospecting of new microbial strains with probiotic and biotechnological potential.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualization, G.Q.L, R.S.D. and F.S.K.; methodology, F.S.K.; software, G.Q.L, K.S.K; validation, G.Q.L, K.S.K., and F.S.K.; resources, F.S.K.; data curation, X.X.; writing—original draft preparation, G.Q.L and R.S.D.; writing—review and editing, G.Q.L, K.S.K, R.S.D., F.P.L.L and F.S.K.; visualization, X.X.; supervision, F.P.L.L and F.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil).

 doi:10.20944/preprints202605.2061.v1

References

1. FAO/WHO Joint FAO/WHO Consultation on Evaluation of Health and Nutritional Properties of Probiotics in Food, Including Powder Milk with Live Lactic Acid Bacteria. *Relatório Téc.* **2001**.
2. Latif, A.; Shehzad, A.; Niazi, S.; Zahid, A.; Ashraf, W.; Iqbal, M.W.; Rehman, A.; Riaz, T.; Aadil, R.M.; Khan, I.M.; et al. Probiotics: Mechanism of Action, Health Benefits, and Their Application in Food Industries. *Front. Microbiol.* **2023**, *14*, 1216674, doi:10.3389/fmicb.2023.1216674.
3. Sepehr, A.; Miri, S.T.; Aghamohammad, S.; Rahimirad, N.; Milani, M.; Pourshafie, M.-R.; Rohani, M. Health Benefits, Antimicrobial Activities, and Potential Applications of Probiotics: A Review. *Medicine (Baltimore)* **2024**, *103*, e32412, doi:10.1097/MD.0000000000032412.
4. Ayyash, M.M.; Abdalla, A.K.; AlKalbani, N.S.; Baig, M.A.; Turner, M.S.; Liu, S.-Q.; Shah, N.P. Invited Review: Characterization of New Probiotics from Dairy and Nondairy Products—Insights into Acid Tolerance, Bile Metabolism and Tolerance, and Adhesion Capability. *J. Dairy Sci.* **2021**, *104*, 8363–8379, doi:10.3168/jds.2021-20398.
5. Mazziotta, C.; Tognon, M.; Martini, F.; Torreggiani, E.; Rotondo, J.C. Probiotics Mechanism of Action on Immune Cells and Beneficial Effects on Human Health. *Cells* **2023**, *12*, 184, doi:10.3390/cells12010184.
6. Mercado-Monroy, J.; Falfán-Cortés, R.N.; Muñóz-Pérez, V.M.; Gómez-Aldapa, C.A.; Castro-Rosas, J. Probiotics as Modulators of Intestinal Barrier Integrity and Immune Homeostasis: A Comprehensive Review. *J. Sci. Food Agric.* **2026**, *106*, 2578–2590, doi:10.1002/jsfa.70168.
7. Grom, L.C.; Coutinho, N.M.; Guimarães, J.T.; Balthazar, C.F.; Silva, R.; Rocha, R.S.; Freitas, M.Q.; Duarte, M.C.K.H.; Pimentel, T.C.; Esmerino, E.A.; et al. Probiotic Dairy Foods and Postprandial Glycemia: A Mini-Review. *Trends Food Sci. Technol.* **2020**, *101*, 165–171, doi:10.1016/j.tifs.2020.05.012.
8. Petraro, S.; Tarracchini, C.; Lugli, G.A.; Mancabelli, L.; Fontana, F.; Turroni, F.; Ventura, M.; Milani, C. Comparative Genome Analysis of Microbial Strains Marketed for Probiotic Interventions: An Extension of the Integrated Probiotic Database. *Microbiome Res. Rep.* **2024**, *3*, doi:10.20517/mrr.2024.11.
9. Churin, A.A.; Sokolyanskaya, L.O.; Lukina, A.P.; Karnachuk, O.V. Current Concepts in Probiotic Safety and Efficacy. *Nutrients* **2026**, *18*, 696, doi:10.3390/nu18040696.
10. Abavisani, M.; Khoshrou, A.; Foroushan, S.K.; Ebadpour, N.; Sahebkar, A. Deciphering the Gut Microbiome: The Revolution of Artificial Intelligence in Microbiota Analysis and Intervention. *Curr. Res. Biotechnol.* **2024**, *7*, 100211, doi:10.1016/j.crbiot.2024.100211.
11. Galal, A.; Talal, M.; Moustafa, A. Applications of Machine Learning in Metabolomics: Disease Modeling and Classification. *Front. Genet.* **2022**, *13*, 1017340, doi:10.3389/fgene.2022.1017340.
12. Sun, Y.; Li, H.; Zheng, L.; Li, J.; Hong, Y.; Liang, P.; Kwok, L.-Y.; Zuo, Y.; Zhang, W.; Zhang, H. iProbiotics: A Machine Learning Platform for Rapid Identification of Probiotic Properties from Whole-Genome Primary Sequences. *Brief. Bioinform.* **2022**, *23*, bbab477, doi:10.1093/bib/bbab477.
13. Dias, R.S.; Martinez, D.P.; Leite, F.P.L.; Avila, L.F.D.C.D.; Kremer, F.S. Pato: Prediction of Probiotic Bacteria Using Metabolic Features. *Braz. J. Microbiol.* **2025**, *56*, 1169–1178, doi:10.1007/s42770-025-01678-x.
14. Tsifintaris, M.; Kiousi, D.E.; Repanas, P.; Kamarinou, C.S.; Kavakiotis, I.; Galanis, A. Probio-Ichnos: A Database of Microorganisms with In Vitro Probiotic Properties. *Microorganisms* **2024**, *12*, 1955, doi:10.3390/microorganisms12101955.
15. Sun, Y.; Li, H.; Zheng, L.; Li, J.; Hong, Y.; Liang, P.; Kwok, L.-Y.; Zuo, Y.; Zhang, W.; Zhang, H. iProbiotics: A Machine Learning Platform for Rapid Identification of Probiotic Properties from Whole-Genome Primary Sequences. *Brief. Bioinform.* **2022**, *23*, bbab477, doi:10.1093/bib/bbab477.
16. Blin, K.; Shaw, S.; Augustijn, H.E.; Reitz, Z.L.; Biermann, F.; Alanjary, M.; Fetter, A.; Terlouw, B.R.; Metcalf, W.W.; Helfrich, E.J.N.; et al. antiSMASH 7.0: New and Improved Predictions for Detection, Regulation, Chemical Structures and Visualisation. *Nucleic Acids Res.* **2023**, *51*, W46–W50, doi:10.1093/nar/gkad344.
17. Orkkatteri Krishnan, A.; Mudgal, L.N.; Soni, V.; Prakash, T. ProbML: A Machine Learning-Based Genome Classifier for Identifying Probiotic Organisms. *Mol. Nutr. Food Res.* **2025**, *69*, e70025, doi:10.1002/mnfr.70025.
18. Jo, D.-M.; Ko, S.-C.; Kim, K.W.; Yang, D.; Kim, J.-Y.; Oh, G.-W.; Choi, G.; Lee, D.-S.; Tabassum, N.; Kim, Y.-M.; et al. Artificial Intelligence-Driven Strategies to Enhance the Application of Lactic Acid Bacteria as Functional Probiotics: Health Promotion and Optimization for Industrial Applications. *Trends Food Sci. Technol.* **2025**, *165*, 105309, doi:10.1016/j.tifs.2025.105309.

 doi:10.20944/preprints202605.2061.v1

19. Tenea, G.N.; Hidalgo, J.; Pircalabioru, G.G.; Cifuentes, V. Genomic and Functional Insights into Lactiplantibacillus Plantarum UTNGt3 from Ecuadorian Amazon *Chrysophyllum Oliviforme*: A Safe and Promising Probiotic. *Front. Microbiol.* **2025**, *16*, 1634475, doi:10.3389/fmicb.2025.1634475.
20. Vorapreeda, T.; Rampai, T.; Chamkhuy, W.; Nopgasorn, R.; Wannawilai, S.; Laoteng, K. Genomic Insights into the Probiotic Functionality and Safety of *Lactiplantibacillus Pentosus* Strain TBRC 20328 for Future Food Innovation. *Foods* **2025**, *14*, 2973, doi:10.3390/foods14172973.
21. Khoo, S.C.; Chin, K.W.; Ting, T.Z.; Luang-In, V.; Chi-Wei Lan, J.; Ma, N.L. Stress Tolerance and Metabolism Profiling of Selected Functional Probiotic Strains. *Food Biosci.* **2025**, *64*, 105919, doi:10.1016/j.fbio.2025.105919.
22. Fijan, S. Probiotics and Their Antimicrobial Effect. *Microorganisms* **2023**, *11*, 528, doi:10.3390/microorganisms11020528.
23. Dolati, M.; Department of Biology, Parand Branch, Islamic Azad University, Parand, Iran; Tafvizi, F.; Department of Biology, Parand Branch, Islamic Azad University, Parand, Iran; Salehipour, M.; Department of Biology, Parand Branch, Islamic Azad University, Parand, Iran; Komeili Movahed, T.; Cellular and Molecular Research Center, Qom University of Medical Sciences, Qom, Iran; Jafari, P.; Microbiology Department, Faculty of Science, Arak Branch, Islamic Azad University, Arak, Iran Antiproliferative Potential of *Bacillus Coagulans* Supernatant on SKBR3 Breast Cancer Cell Line. *Iran. J. Med. Microbiol.* **2022**, *16*, 383–391, doi:10.30699/ijmm.16.5.383.
24. Caliman-Sturdza, O.A.; Vaz, J.A.; Lupaescu, A.V.; Lobiuc, A.; Bran, C.; Gheorghita, R.E. Antioxidant and Anti-Inflammatory Activities of Probiotic Strains. *Int. J. Mol. Sci.* **2026**, *27*, 1079, doi:10.3390/ijms27021079.
25. Ma, J.; Fang, Y.; Li, S.; Zeng, L.; Chen, S.; Li, Z.; Ji, G.; Yang, X.; Wu, W. Interpretable Machine Learning Algorithms Reveal Gut Microbiome Features Associated with Atopic Dermatitis. *Front. Immunol.* **2025**, *16*, 1528046, doi:10.3389/fimmu.2025.1528046.
26. Bernatek, M.; Żukiewicz-Sobczak, W.; Lachowicz-Wiśniewska, S.; Piątek, J. Factors Determining Effective Probiotic Activity: Evaluation of Survival and Antibacterial Activity of Selected Probiotic Products Using an "In Vitro" Study. *Nutrients* **2022**, *14*, 3323, doi:10.3390/nu14163323.
27. Guo, P.; Wang, W.; Xiang, Q.; Pan, C.; Qiu, Y.; Li, T.; Wang, D.; Ouyang, J.; Jia, R.; Shi, M.; et al. Engineered Probiotic Ameliorates Ulcerative Colitis by Restoring Gut Microbiota and Redox Homeostasis. *Cell Host Microbe* **2024**, *32*, 1502-1518.e9, doi:10.1016/j.chom.2024.07.028.
28. Fuochi, V.; Petronio, G.P.; Lissandrello, E.; Furneri, P.M. Evaluation of Resistance to Low pH and Bile Salts of Human *Lactobacillus* Spp. Isolates. *Int. J. Immunopathol. Pharmacol.* **2015**, *28*, 426–433, doi:10.1177/0394632015590948.
29. Wu, H.; Li, Y.; Jiang, Y.; Li, X.; Wang, S.; Zhao, C.; Yang, X.; Chang, B.; Yang, J.; Qiao, J. Machine Learning Prediction of Obesity-Associated Gut Microbiota: Identifying *Bifidobacterium Pseudocatenulatum* as a Potential Therapeutic Target. *Front. Microbiol.* **2025**, *15*, 1488656, doi:10.3389/fmicb.2024.1488656.
30. Cao, Y.; Hu, X.; Guo, J.; Fang, T. Machine Learning-Based Prediction of Recurrent Extrahepatic Bile Duct Stones after Common Bile Duct Exploration: A Comparative Study of Models and SHAP-Driven Interpretability Analysis. *Front. Med.* **2025**, *12*, 1691519, doi:10.3389/fmed.2025.1691519.
31. Kiriyama, Y.; Nochi, H. Physiological Role of Bile Acids Modified by the Gut Microbiome. *Microorganisms* **2021**, *10*, 68, doi:10.3390/microorganisms10010068.
32. Green, M.; Arora, K.; Prakash, S. Microbial Medicine: Prebiotic and Probiotic Functional Foods to Target Obesity and Metabolic Syndrome. *Int. J. Mol. Sci.* **2020**, *21*, 2890, doi:10.3390/ijms21082890.
33. Liu, W.; Li, Z.; Ze, X.; Deng, C.; Xu, S.; Ye, F. Multispecies Probiotics Complex Improves Bile Acids and Gut Microbiota Metabolism Status in an in Vitro Fermentation Model. *Front. Microbiol.* **2024**, *15*, 1314528, doi:10.3389/fmicb.2024.1314528.
34. Ruiz, L.; Margolles, A.; Sánchez, B. Bile Resistance Mechanisms in *Lactobacillus* and *Bifidobacterium*. *Front. Microbiol.* **2013**, *4*, doi:10.3389/fmicb.2013.00396.
35. Contreas, L.; Hook, A.L.; Winkler, D.A.; Figueredo, G.; Williams, P.; Laughton, C.A.; Alexander, M.R.; Williams, P.M. Linear Binary Classifier to Predict Bacterial Biofilm Formation on Polyacrylates. *ACS Appl. Mater. Interfaces* **2023**, acsami.2c23182, doi:10.1021/acsami.2c23182.

 doi:10.20944/preprints202605.2061.v1

36. Plaza-Diaz J, Ruiz-Ojeda FJ, Gil-Campos M, Gil A Mechanisms of Action of Probiotics. *Adv. Nutr.* **2020**, *11*, 1054, doi:10.1093/advances/nmaa042.
37. Xing, T.; Bian, X.; Ma, C.; Yang, Y.; Liu, X.; Wang, Y.; Fan, J.; Zhang, N. In Vitro Evaluation of Probiotic Properties of *Lactobacillus Acidophilus* AD125 and Antagonism against *Escherichia Coli* O157:H7 Adhesion to Caco-2 Cell. *Food Funct.* **2023**, *14*, 2472–2480, doi:10.1039/D2FO03200G.
38. Zawistowska-Rojek, A.; Kośmider, A.; Stępień, K.; Tyski, S. Adhesion and Aggregation Properties of Lactobacillaceae Strains as Protection Ways against Enteropathogenic Bacteria. *Arch. Microbiol.* **2022**, *204*, 285, doi:10.1007/s00203-022-02889-8.
39. Zheng, Y.; Zhao, J.; Nie, X.; Chitrakar, B.; Gao, J.; Sang, Y. Mutual Adhesion of Lactobacillus Spp. to Intestinal Cells: A Review of Perspectives on Surface Layer Proteins and Cell Surface Receptors. *Int. J. Biol. Macromol.* **2024**, *282*, 137031, doi:10.1016/j.ijbiomac.2024.137031.
40. Kong, W.; Gan, J.; Su, M.; Xiong, B.; Jiang, X.; Zhang, T.; Zeng, X.; Wu, Z.; Sun, Y.; Pan, D.; et al. Identification and Characterization of Domains Responsible for Cell Wall Binding, Self-Assembly, and Adhesion of S-Layer Protein from *Lactobacillus Acidophilus* CICC 6074. *J. Agric. Food Chem.* **2022**, *70*, 12982–12989, doi:10.1021/acs.jafc.2c03907.
41. Jain, R.; Voss, A.L.; Del Rosario, J.; Hale, J.D.F. Efficacy of a Topical Live Probiotic in Improving Skin Health. *Int. J. Cosmet. Sci.* **2025**, *47*, 488–496, doi:10.1111/ics.13043.
42. Aghamohammad, S.; Rohani, M. Antibiotic Resistance and the Alternatives to Conventional Antibiotics: The Role of Probiotics and Microbiota in Combating Antimicrobial Resistance. *Microbiol. Res.* **2023**, *267*, 127275, doi:10.1016/j.micres.2022.127275.
43. Chornchoem, P.; Tandhavanant, S.; Saiprom, N.; Preechanukul, A.; Thongchompoo, N.; Sensorn, I.; Chantratita, W.; Chantratita, N. Metagenomic Evaluation, Antimicrobial Activities, and Immune Stimulation of Probiotics from Dietary Supplements and Dairy Products. *Sci. Rep.* **2025**, *15*, 11537, doi:10.1038/s41598-025-95664-w.
44. Scillato, M.; Spitale, A.; Mongelli, G.; Privitera, G.F.; Mangano, K.; Cianci, A.; Stefani, S.; Santagati, M. Antimicrobial Properties of *Lactobacillus* Cell-free Supernatants against Multidrug-resistant Urogenital Pathogens. *MicrobiologyOpen* **2021**, *10*, e1173, doi:10.1002/mbo3.1173.
45. He, G.; Long, H.; He, J.; Zhu, C. The Immunomodulatory Effects and Applications of Probiotic *Lactiplantibacillus Plantarum* in Vaccine Development. *Probiotics Antimicrob. Proteins* **2024**, *16*, 2229–2250, doi:10.1007/s12602-024-10338-9.
46. Costa, F.F.; Dias, T.G.; Mendes, P.M.; Viana, J.P.M.; Madeira, E.B.; Pereira, A.L.F.; Ferreira, A.G.N.; Neto, M.S.; Dutra, R.P.; Reis, A.S.; et al. Antioxidant and Antimicrobial Properties of Probiotics: Insights from In Vitro Assays. *Probiotics Antimicrob. Proteins* **2025**, *17*, 2721–2731, doi:10.1007/s12602-024-10426-w.
47. Al-Shamiri, M.M.; Wang, J.; Zhang, S.; Li, P.; Odhiambo, W.O.; Chen, Y.; Han, B.; Yang, E.; Xun, M.; Han, L.; et al. Probiotic *Lactobacillus* Species and Their Biosurfactants Eliminate *Acinetobacter Baumannii* Biofilm in Various Manners. *Microbiol. Spectr.* **2023**, *11*, e04614-22, doi:10.1128/spectrum.04614-22.
48. Aghamohammad, S.; Sakaki, F.; Haghighatshenas, Z.; Khiavi, E.H.A.G.; Badmasti, F.; Rohani, M. Novel Antimicrobial Strategy: Native Postbiotics Synergize with Antibiotics to Overcome Resistance in *Klebsiella Pneumoniae*. *BMC Microbiol.* **2025**, *25*, 723, doi:10.1186/s12866-025-04474-7.
49. Thoda, C.; Touraki, M. Immunomodulatory Properties of Probiotics and Their Derived Bioactive Compounds. *Appl. Sci.* **2023**, *13*, 4726, doi:10.3390/app13084726.
50. Adejumo, S.A.; Oli, A.N.; Rowaiye, A.B.; Igbokwe, N.H.; Ezejiegu, C.K.; Yahaya, Z.S. Immunomodulatory Benefits of Probiotic Bacteria: A Review of Evidence. *OBM Genet.* **2023**, *07*, 1–73, doi:10.21926/obm.genet.2304206.
51. Abriouel, H.; Caballero Gómez, N.; Manetsberger, J.; Benomar, N. Dual Effects of a Bacteriocin-Producing Lactiplantibacillus Pentosus CF-6HA, Isolated from Fermented Aloreña Table Olives, as Potential Probiotic and Antimicrobial Agent. *Heliyon* **2024**, *10*, e28408, doi:10.1016/j.heliyon.2024.e28408.
52. González, A.; Fullaondo, A.; Rodríguez, J.; Tirnauca, C.; Odriozola, I.; Odriozola, A. Conjugated Linoleic Acid Metabolite Impact in Colorectal Cancer: A Potential Microbiome-Based Precision Nutrition Approach. *Nutr. Rev.* **2025**, *83*, e602–e614, doi:10.1093/nutrit/nuae046.

 doi:10.20944/preprints202605.2061.v1

53. Pattapulavar, V.; Ramanujam, S.; Kini, B.; Christopher, J.G. Probiotic-Derived Postbiotics: A Perspective on next-Generation Therapeutics. *Front. Nutr.* **2025**, *12*, 1624539, doi:10.3389/fnut.2025.1624539.
54. Chuah, L.-O.; Foo, H.L.; Loh, T.C.; Mohammed Alitheen, N.B.; Yeap, S.K.; Abdul Mutalib, N.E.; Abdul Rahim, R.; Yusoff, K. Postbiotic Metabolites Produced by *Lactobacillus Plantarum* Strains Exert Selective Cytotoxicity Effects on Cancer Cells. *BMC Complement. Altern. Med.* **2019**, *19*, 114, doi:10.1186/s12906-019-2528-2.