SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-023-2569-7

RESEARCH PAPER

Predicting *Lactobacillus delbrueckii* subsp. *bulgaricus*-*Streptococcus thermophilus* interactions based on a highly accurate semi-supervised learning method

Shujuan Yang[1,2,3,4†], Mei Bai[1,2,3,4†], Weichi Liu[5,6†], Weicheng Li[1,2,3,4], Zhi Zhong[1,2,3,4], Lai-Yu Kwok[1,2,3,4], Gaifang Dong[5,6*] & Zhihong Sun[1,2,3,4*]

[1]Key Laboratory of Dairy Biotechnology and Engineering, Ministry of Education, Inner Mongolia Agricultural University, Hohhot 010018, China;
[2]Key Laboratory of Dairy Products Processing, Ministry of Agriculture and Rural Affairs, Inner Mongolia Agricultural University, Hohhot 010018, China;
[3]Inner Mongolia Key Laboratory of Dairy Biotechnology and Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China;
[4]Collaborative Innovative Center for Lactic Acid Bacteria and Fermented Dairy Products, Ministry of Education, Inner Mongolia Agricultural University, Hohhot 010018, China;
[5]College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China;
[6]Inner Mongolia Autonomous Region Key Laboratory of Big Data Research and Application of Agriculture and Animal Husbandry, Hohhot 010018, China

†Contributed equally to this work
*Corresponding authors (Gaifang Dong, email: donggf@imau.edu.cn; Zhihong Sun, email: sunzhihong78@163.com)

Received 18 December 2023; Accepted 15 March 2024; Published online 14 October 2024

***Lactobacillus delbrueckii subsp. bulgaricus* (*L. bulgaricus*) and *Streptococcus thermophilus* (*S. thermophilus*) are commonly used starters in milk fermentation. Fermentation experiments revealed that *L. bulgaricus-S. thermophilus* interactions (*LbSt*I) substantially impact dairy product quality and production. Traditional biological humidity experiments are time-consuming and labor-intensive in screening interaction combinations, an artificial intelligence-based method for screening interactive starter combinations is necessary. However, in the current research on artificial intelligence based interaction prediction in the field of bioinformatics, most successful models adopt supervised learning methods, and there is a lack of research on interaction prediction with only a small number of labeled samples. Hence, this study aimed to develop a semi-supervised learning framework for predicting *LbSt*I using genomic data from 362 isolates (181 per species). The framework consisted of a two-part model: a co-clustering prediction model (based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) dataset) and a Laplacian regularized least squares prediction model (based on K-mer analysis and gene composition of all isolates datasets). To enhance accuracy, we integrated the separate outcomes produced by each component of the two-part model to generate the ultimate *LbSt*I prediction results, which were verified through milk fermentation experiments. Validation through milk fermentation experiments confirmed a high precision rate of 85% (17/20; validated with 20 randomly selected combinations of expected interacting isolates). Our data suggest that the biosynthetic pathways of cysteine, riboflavin, teichoic acid, and exopolysaccharides, as well as the ATP-binding cassette transport systems, contribute to the mutualistic relationship between these starter bacteria during milk fermentation. However, this finding requires further experimental verification. The presented model and data are valuable resources for academics and industry professionals interested in screening dairy starter cultures and understanding their interactions.**

***Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus* | interaction prediction | semi-supervised learning | dairy starter | artificial intelligence | milk fermentation**

INTRODUCTION

Lactic acid bacteria (LAB) are important industrial microorganisms widely used in the food, pharmaceutical, and feed industries (Hatti-Kaul et al., 2018). They are Gram-positive, acid-tolerant, cocci or rod-shaped bacteria that metabolize carbohydrates as their sole or major carbon source (George et al., 2018; Wang et al., 2021b). Lactic acid bacteria are the most commonly used starter cultures in food fermentation. A starter culture is a preparation consisting of one or more bacterial species/strains that are generally incorporated in the raw material of fermented foods to accelerate and steer the fermentation process (Sharma et al., 2023). The use of this bacterial preparation is paramount to ensuring consistent and safe production of fermented foods.

There is a long history of human production and consumption of fermented dairy products. These acidic dairy products are created by starter cultures and/or specific microbes that acidify milk (Macori and Cotter, 2018; Zannini et al., 2016). Initially, natural fermentation was employed, but since the last century, the food industry has made technological advancements, commercializing fermented food production utilizing natural starter cultures and adjuvant microbes. Dairy starter exemplifies the proficient use of LAB in the dairy fermentation sector. The biological properties and activities of LAB play a crucial role in manufacturing fermented dairy products, particularly those with distinctive flavors and textures (Bintsis, 2018; Sharma et al., 2023). Fermented foods produced by functional strains can contain numerous biologically active metabolites. These include aminobutyric acid, exopolysaccharides, conjugated linoleic acid, and bacteriocins, such as reuterin, which enhance the nutraceutical properties of finished food products (Abedin et al., 2023). The metabolic capacity and diversity of LAB make them significant contributors to the functional food industry, as they improve human nutrition and health.

Lactobacillus delbrueckii subsp. *bulgaricus* (*L. bulgaricus*) and *Streptococcus thermophilus* (*S. thermophilus*) are the predominant bacteria used in the production of dairy starters. The two species typically work together in dairy fermentation, supporting each other synergistically and symbiotically for improved survival and growth in the milk environment and throughout the fermentation process. The co-culture of these two bacteria dictates the efficiency of the fermentation process and the quality of the finished products (Deshwal et al., 2021; Ge et al., 2024). Their symbiotic relationship confers desirable fermentation properties, such as rapid acid production, high viscosity, and production of diverse bioactive substances, imparting the sensory and functional quality of the fermented milk products (Yang et al., 2023). A fermented milk microbial ecosystem consists of a complex network of mutualistic and feedback interactions among various members of the microbial community, rather than a simple aggregate of independent microbes (Settachaimongkon et al., 2014; Sieuwerts, 2016). Though interactions between these two species have been studied previously, screening for optimal starter combinations has remained a challenge due to the vast diversity of LAB genomes and small genomic differences that lead to significant functional variations.

A customary approach for identifying ideal starter combinations involves traditional and experimental approaches, but this way is laborious. Alternatively, the development of computational models for predicting microbial interactions followed by verification and further data training through biological experiments is an effective avenue of research. Currently, there are four trending areas of interaction research in bioinformatics: protein-protein interactions, gene regulatory networks, drug-target interactions, and molecular interaction networks. Computational and machine learning algorithms are routinely employed to deduce the likelihood of protein interactions or probably binding patterns based on data consisting of protein structure, sequence, and function. Commonly used methods include structure-based docking simulation, sequence alignment, and machine learning (Lei et al., 2021; Lian et al., 2019; Wang et al., 2023; Xu et al., 2021). In gene regulatory networks, interactions between transcription factors and target genes are modeled as network connections, revealing the complex relationships and mechanisms of gene regulation. Methods for regulatory network analysis include correlation analysis based on expression data, topology analysis, and dynamic simulation (Jansen et al., 2022; Peng et al., 2023). Drug-target interaction focuses on the affinity and interaction patterns between drugs and target proteins by integrating bioinformatics data, drug chemistry information, and protein structure information. Commonly used methods include virtual screening based on molecular docking, drug similarity calculation, and machine learning (Deng et al., 2022; Dong et al., 2023; Gu et al., 2023; Li et al., 2022; Peng et al., 2020b; Zhang et al., 2022b; Zhou et al., 2021). Lastly, molecular interaction networking seeks to construct biomolecular interaction networks by integrating multiple experimental data and bioinformatics analysis. This process reveals physical interactions, regulatory relationships, and signaling pathways between molecules. Common methods of achieving this include graph-based network analysis, systems biology modeling, and network dynamics simulation (Li et al., 2024; Wen et al., 2017). These four types of interaction modeling have provided rich insights for predicting microbial interactions in the milk fermentation microbial ecosystem.

The objective of this study was to develop a one-on-one combination screening model for *L.* bulgaricus-*S. thermophilus* interactions (*LbSt*I) during milk fermentation. This was achieved by using a co-clustering algorithm together with the Laplacian regularized least squares (LapRLS) prediction model. This study generated a predictive model of *LbSt*I using the genomic data from a vast quantity of food-derived *L.* bulgaricus and *S. thermophilus* isolates (181 isolates per species), amounting to a total of 32,761 isolate combinations. The resulting model was substantiated by a stringent validation process through cross-validation, machine learning, and fermentation experiments (Fermentation time and viscosity of fermented milk), confirming its high accuracy in predicting starter interaction and fermentation outcome. Specifically, the combination of strains (positive and negative combinations) screened by the model shows high consistency with the fermentation experiment results. This work presents a valuable resource for academics and industry professionals for screening starter cultures and gaining insight into dairy starter interactions.

RESULTS

Co-clustering and LbStIPred_SimLapRLS results

Our experimental data comes from a wide range of sources and types, providing a solid data foundation for computer modelling (Figure 1). A total of 711 positively and 362 negatively interacting isolate combinations were predicted through co-clustering (Table S1 in Supporting Information). The *LbSt*IPred_-ed_SimLapRLS model scored 2292 combinations with isolate interaction (Table 1). These combinations were ranked according to their probability of interaction; higher rankings represent a higher likelihood of interaction, while lower rankings denote a lower probability. Based on the interaction score, an approximate interval of *LbSt*I was determined. When the interaction matrix was constructed, the labeled positive and negative combinations were set to 1 and −1, while the predicted combinations were set to 0 when constructing the interaction matrix. Thus, the top 1,314 combinations received positive scores, while the bottom 1,314 received negative scores. This parameter setting would result in a score closer to 1 if the similarity between predicted and labeled positive combinations was high, and a score closer to −1 if the similarity between predicted and labeled negative combinations was higher.

Final LbStI prediction results

Finally, the prediction results of co-clustering and *LbSt*IPred_-ed_SimLapRLS were compared and combined as the final prediction results of *LbSt*I, yielding 142 final prediction isolate combinations (Figure 2C). Positive combinations predicted by co-clustering were among the top 500 of the *LbSt*IPred_SimLapRLS rankings, while negative combinations were distributed at the back of the rankings (after 1,000), indicating that the predictions from both methods corroborated each other.

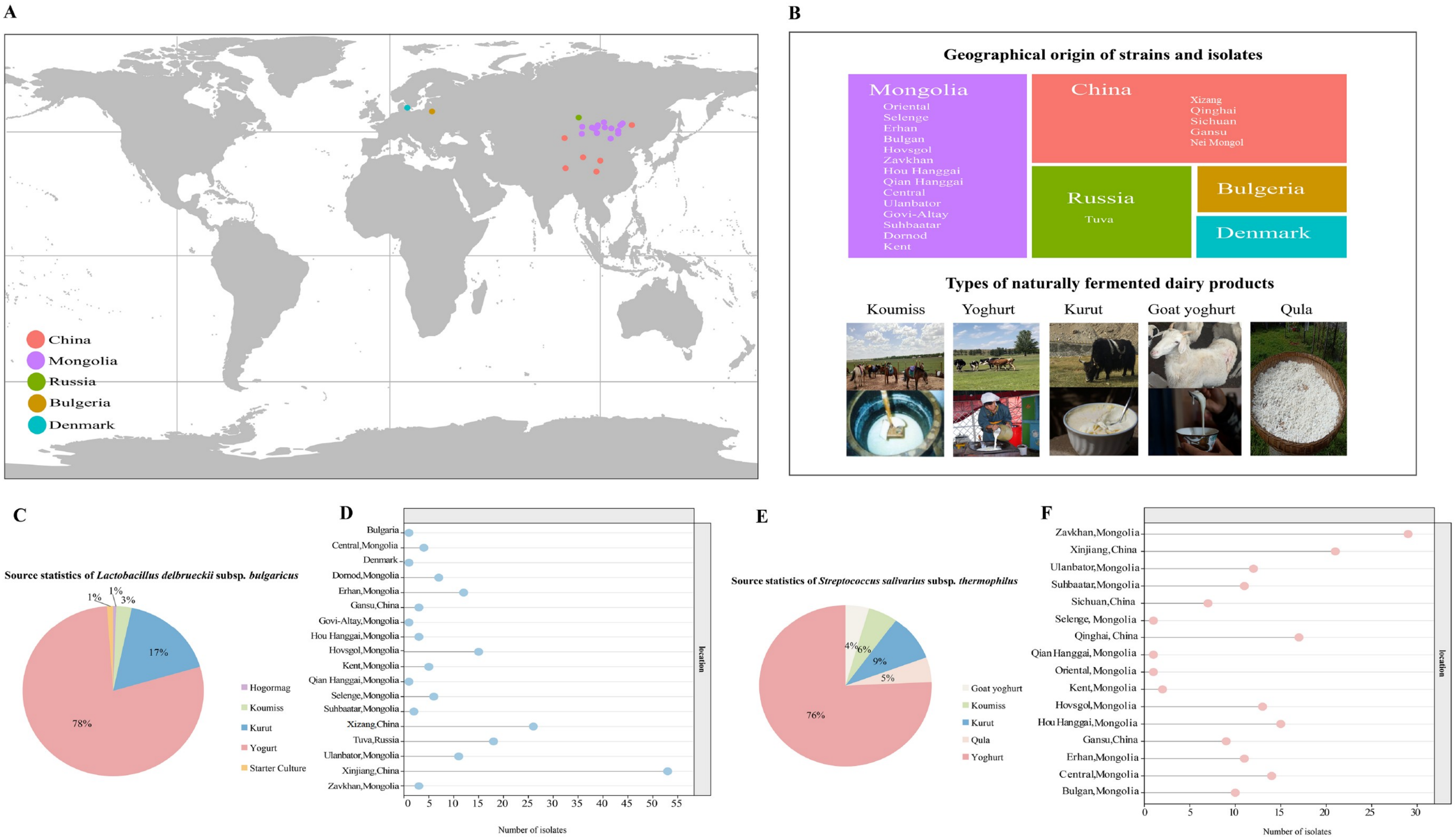

Figure 1. Geographical origins of strains and isolates. A, Global map showing the sample distribution. B, Types of naturally fermented products and geographical origin of bacterial strains and isolates. C–F, The two strains from Denmark and Bulgaria were industrial starter cultures. Source statistics of *Lactobacillus delbrueckii* subspecies *bulgaricus* (C and D) and *Streptococcus thermophilus* (E and F).

Table 1. Predicted interaction results generated by *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus* interaction prediction framework based on similarity-fusion LapRLS (*LbSt*IPred_SimLapRLS)

Interaction label	Combination (respective *Lb*, *St* isolates)	Interaction score	Rank
Positive	IMAU95110, IMAU20774	0.911868768	none
Positive	IMAU95110, IMAU80844	0.911688595	none
Positive	IMAU20450, IMAU20774	0.911535088	none
Positive	IMAU95110, IMAU20543	0.911409419	none
Positive	IMAU95110, IMAU20588	0.911395415	none
Positive	IMAU95110, IMAU20588	0.911298232	none
Positive	IMAU95110, IMAU40133	0.911261113	none
Positive	IMAU20450, IMAU20588	0.911218362	none
Positive	IMAU20450, IMAU20543	0.91121752	none
Positive	IMAU20450, IMAU40133	0.910960416	none
Positive	IMAU62091, IMAU40133	0.910062115	none
Nil	IMAU95110, IMAU80845	0.001737867	1
Nil	IMAU95110, IMAU80840	0.001730118	2
Nil	IMAU95110, IMAU80842	0.001704359	3
Nil	*	*	4–499
Nil	IMAU95087, IMAU20543	0.000614619	500
Nil	*	*	501–999
Nil	IMAU62091, IMAU32092	0.000246895	1,000
Nil	*	*	1,001–1,199
Nil	IMAU62091, IMAU205623	0.000143121	1,200
Nil	*	*	1,201–1,313
Nil	IMAU32111, IMAU80844	0.0000015572	1,314
Nil	IMAU32265, IMAU80844	−0.000000309	1,315
Nil	*	*	1,316–1,699
Nil	IMAU62081, IMAU32112	−0.000226876	1,700
Nil	*	*	1,701–2,289
Nil	IMAU32076, IMAU32476	−0.00134442	2,290
Nil	IMAU32076, IMAU80840	−0.001355872	2,291
Nil	IMAU32076, IMAU80845	−0.001363913	2,292
Negative	IMAU20450, IMAU20766	−0.909195262	none
Negative	IMAU62161, IMAU40133	−0.910488145	none
Negative	IMAU62081, IMAU40133	−0.910784237	none
Negative	IMAU32076, IMAU20543	−0.910806333	none
Negative	IMAU32076, IMAU20588	−0.910811971	none
Negative	IMAU32076, IMAU80844	−0.910828816	none
Negative	IMAU20312, IMAU40133	−0.910831868	none
Negative	IMAU32076, IMAU40133	−0.911187354	none

Notes: *Lb* and *St* represent *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus* (St), respectively. The non-labeled combinations were predicted to show starter interaction. Combinations were ranked according to how likely they were to interact; higher rankings denote a higher probability of interaction, while lower rankings denote a lower probability. Only representative results are shown in this table (refer to Table S1 in Supporting Information for complete prediction results). *Symbol markings indicate omissions.

leave-one-out cross-validation

We used LOOCV to test the predictive performance of the *LbSt*IPred_SimLapRLS model (Chen et al., 2021a). Nineteen rounds of validation were performed on the 19 labeled isolate combinations. One labeled isolate combination was chosen in each round of validation, and the corresponding value was set to 0 when constructing the interaction matrix. *LbSt*IPred_SimSimLapRLS was then used to make predictions of the picked labeled combination. Since LOOCV would modify the prediction range of *LbStI*Pred_SimLapRLS when setting the interaction matrix, the total number of predictions in each round varied. The *LbSt*IPred_SimLapRLS prediction was considered correct if the rank percentages of the positive and negative combinations were in the top and bottom 50%, respectively. A high accuracy of 89.47% (17/19) was achieved (Table 2). The validation test found that five of the 11 positive combinations (IMAU20450, IMAU20774; IMAU95110, IMAU80844; IMAU95110, IMAU20543; IMAU95110, IMAU20588; and IMAU95110, IMAU20774) ranked in the top 10% (corresponding to 7.50%, 2.18%, 5.41%, 6.59%, and 0.04%, respectively); four combinations (IMAU20450, IMAU80844; IMAU20450, IMAU20543; IMAU20450, IMAU20588; and IMAU95110, IMAU40133) had rank percentages between 10% and 20% (corresponding to

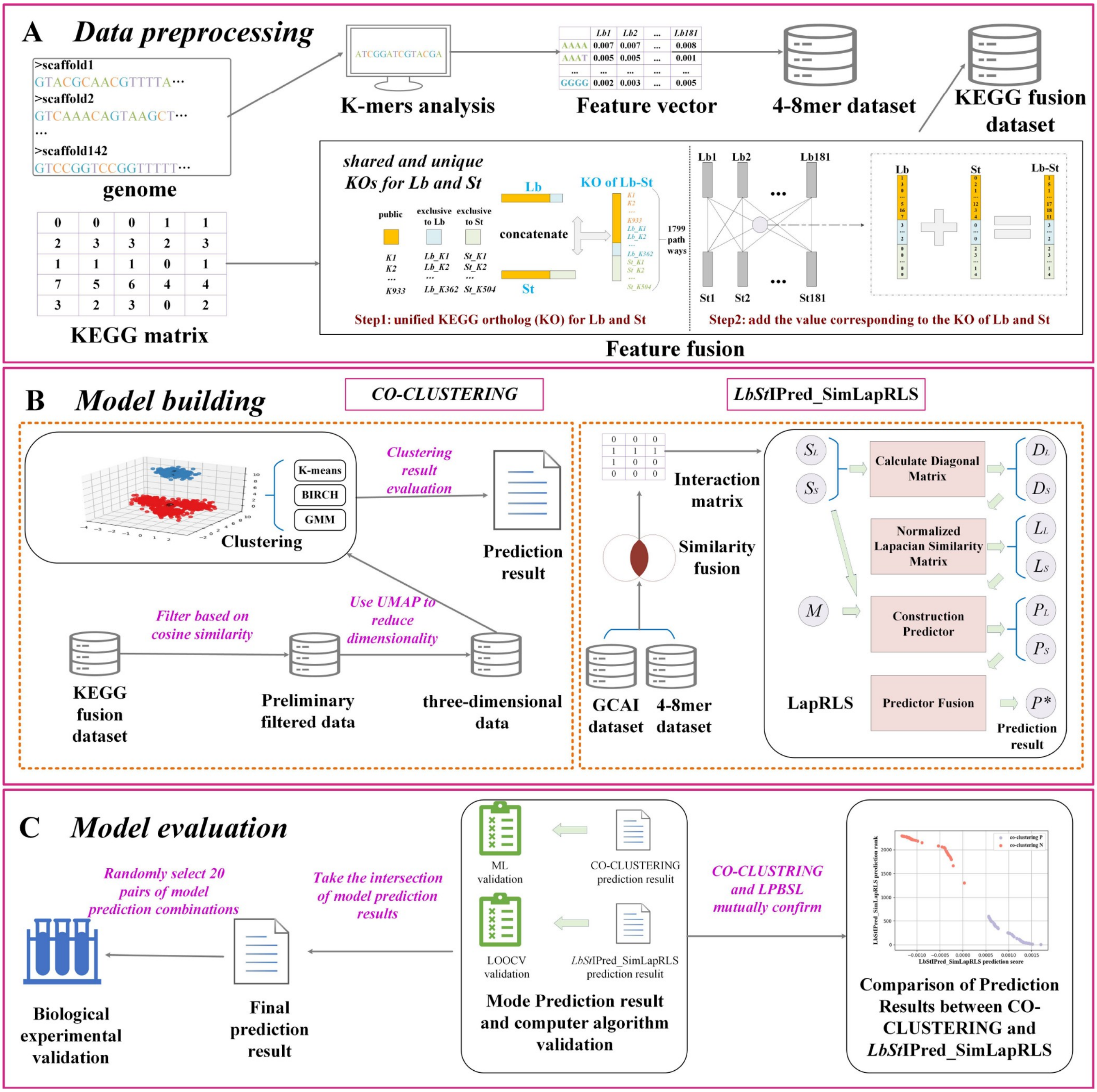
A Data preprocessing
>scaffold1
GTACGCAACGTTTTA···
>scaffold2
GTCAAACAGTAAGCT···
···
>scaffold142
GTCCGGTCCGGTTTTT···
genome
ATCGGATCGTACGA
K-mers analysis
Lb1
Lb2
...
Lb181
AAAA 0.007 0.007 ... 0.008
AAAT 0.005 0.005 ... 0.001
GGGG 0.002 0.003 ... 0.005
Feature vector
4-8mer dataset
KEGG fusion dataset
0 0 0 1 1
2 3 3 2 3
1 1 1 0 1
7 5 6 4 4
3 2 3 0 2
KEGG matrix
shared and unique KOs for Lb and St
public
exclusive to Lb
exclusive to St
K1
K2
K933
Lb_K1
Lb_K2
Lb_K362
St_K1
St_K2
St_K504
Lb
St
concatenate
KO of Lb-St
1799 path ways
Lb1
Lb2
Lb181
St1
St2
St181
Lb
St
Lb-St
Step1: unified KEGG ortholog (KO) for Lb and St
Step2: add the value corresponding to the KO of Lb and St
Feature fusion
B Model building
CO-CLUSTERING
LbStIPred_SimLapRLS
K-means
BIRCH
GMM
Clustering
Clustering result evaluation
Prediction result
Filter based on cosine similarity
Use UMAP to reduce dimensionality
KEGG fusion dataset
Preliminary filtered data
three-dimensional data
Interaction matrix
Similarity fusion
GCAI dataset
4-8mer dataset
S_L
S_S
Calculate Diagonal Matrix
D_L
D_S
Normalized Lapacian Similarity Matrix
L_L
L_S
M
Construction Predictor
P_L
P_S
LapRLS
Predictor Fusion
P*
Prediction result
C Model evaluation
Randomly select 20 pairs of model prediction combinations
Biological experimental validation
Final prediction result
Take the intersection of model prediction results
ML validation
CO-CLUSTERING prediction resulit
LOOCV validation
LbStIPred_SimLapRLS prediction resulit
Mode Prediction result and computer algorithm validation
CO-CLUSTRING and LPBSL mutually confirm
co-clustering P
co-clustering N
LbStIPred_SimLapRLS prediction rank
LbStIPred_SimLapRLS prediction score
Comparison of Prediction Results between CO-CLUSTERING and LbStIPred_SimLapRLS

Figure 2. Schematic flow diagram of this work. A, Data preprocessing—from acquiring the 4–8 mer dataset through K-mer analysis to concatenating the KEGG data. B, Model building—the process of construction of the *LbSt*I Prediction based on the Similarity-fusion LapRLS (*LbSt*IPred_SimLapRLS) and co-clustering models. C, Generation of the final results by combining the intersection prediction sets of the co-clustering and *LbSt*IPred_SimLapRLS and validation of the final results by milk fermentation experiments.

13.52%, 10.12%, 12.12%, and 11.56%, respectively). Furthermore, four of the eight negative combinations (IMAU20312, IMAU20766; IMAU32076, IMAU20543; IMAU32076, IMAU20588; and IMAU32076, IMAU40133) ranked below 70% (corresponding to 80.97%; 80.33%; 72.83%; and 91.89%, respectively). The results of LOOCV validation confirmed a good prediction performance of *LbSt*IPred_SimLapRLS.

Validation with milk fermentation experiments

We randomly selected 20 isolate combinations (13 positively and seven negatively interactive) from the 142 potential interactive combinations predicted by *LbSt*I for validation with milk fermentation experiments. A fast acid production rate is a significant criterion for identifying high-quality starter cultures, as good starter isolates can enhance the production efficiency and quality of fermented milk (Dan et al., 2017). One-day ripening improved the flavor and texture of fermented milk. Specifically, low-temperature post-ripening can control the acidity of fermented milk, produce acetone and other flavor substances, enrich the taste of yogurt, and make the state of yogurt more stable. Therefore, we chose the yogurt after one day

Table 2. Leave-one-out cross-validation (LOOCV) of predicted results

Round	Label	Combination (respective *Lb*, *St* isolates)	Rank	Total number of predictions	Rank percentage	Prediction accuracy
1	Positive	IMAU20450, IMAU80844	310	2,293	13.52%	True
2	Positive	IMAU20450, IMAU20543	232	2,293	10.12%	True
3	Positive	IMAU20450, IMAU20588	278	2,293	12.12%	True
4	Positive	IMAU20450, IMAU20774	172	2,293	7.50%	True
5	Positive	IMAU20450, IMAU40133	796	2,293	34.71%	True
6	Negative	IMAU20450, IMAU20766	61	2,119	2.88%	False
7	Positive	IMAU62091, IMAU40133	1,910	2,118	90.18%	False
8	Positive	IMAU95110, IMAU80844	50	2,293	2.18%	True
9	Positive	IMAU95110, IMAU20543	124	2,293	5.41%	True
10	Positive	IMAU95110, IMAU20588	151	2,293	6.59%	True
11	Positive	IMAU95110, IMAU20774	1	2,293	0.04%	True
12	Positive	IMAU95110, IMAU40133	265	2,293	11.56%	True
13	Negative	IMAU20312, IMAU20766	1,715	2,118	80.97%	True
14	Negative	IMAU62081, IMAU40133	1,296	2,118	61.19%	True
15	Negative	IMAU62161, IMAU40133	1,142	2,293	53.92%	True
16	Negative	IMAU32076, IMAU80844	1,523	2,293	66.42%	True
17	Negative	IMAU32076, IMAU20543	1,842	2,293	80.33%	True
18	Negative	IMAU32076, IMAU20588	1,670	2,293	72.83%	True
19	Negative	IMAU32076, IMAU40133	2,107	2,293	91.89%	True

Notes: *Lb* and *St* represent *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus* (St), respectively. *LbSt*IPred_SimLapRLS predicted results were validated by 19 rounds of LOOCV. Each labeled combination was ranked by LOOCV. The total number of predictions in each round varied due to the impact of LOOCV on the prediction range of *LbSt*IPred_SimLapRLS when generating the interaction matrix. When the rank percentage of the labeled positive and negative combination fell within the top and bottom 50%, respectively, the *LbSt*IPred_SimLapRLS prediction was deemed accurate.

of ripening to evaluate its characteristics, which are reflected in the changes in its viscosity, pH value, water retention capacity and sensory characteristics. The choice of starter culture significantly impacts these properties. Accordingly, we validated the *LbSt*I prediction results by evaluating the fermentation performance of these 20 randomly selected isolate combinations based on these attributes.

Fermentation time required to reach the fermentation endpoint

The commercial starter, Control-YF922, required the shortest time to reach the fermentation endpoint of pH 4.5 (6.07±0.15) h; Figure 3A). Nearly all (except one) of the 20 predicted positively interacting isolate combinations required less than 8 h to reach the fermentation endpoint, while all the predicted negatively interacting combinations needed over 8 h to complete the milk fermentation. These results suggest that the fermentation time and the projected starter interaction align with each other.

Fermented milk properties after 1-day ripening

(i) Viscosity. Ten out of the 13 probable positively interacting isolate combinations produced fermented milk with high viscosity (over 800 mPa·s), with the highest viscosity produced by the starter combination, IMAU95110 and IMAU10630 (mean=2,856 mPa·s; Figure 3B). The remaining three pairs produced fermented milk with low viscosity and poor texture. In contrast, all seven pairs of probable negatively interacting isolate combinations produced fermented milk with low viscosity and poor texture (Figure 3B), resembling a bean curd residue state. These data suggest that the predicted result of starter isolate interactions is largely consistent with the viscosity of fermented milk. See Table S6 in Supporting Information for the comparison details of the significant differences between the groups in Figure 3A and B.

(ii) Water-holding capacity. The water-holding capacity of fermented milk is a vital criterion for evaluating its internal structure stability and whey precipitation. The water-holding capacity of the ripened fermented milk was similar in both positively interacting (ranging from 55.6% to 63.3%) and negatively interacting combinations (ranging from 54.1% to 62.45%; Figure S1A in Supporting Information). These results suggest that there is no correlation between the water-holding capacity of fermented milk and the predicted outcome of isolate interaction.

(iii) pH. The acidity level of fermented milk significantly impacts its shelf life, flavor, and taste, affecting production time, efficiency, and product quality. A proper acidity level contributes to the pleasant taste and smooth texture of fermented milk. The pH values of 1-day ripened fermented milk produced by the 20 isolate combinations varied from 4.3 to 4.4, with no significant difference observed between the positively or negatively predicted combinations (Figure S1B in Supporting Information).

(iv) Sensory quality. Sensory evaluation is a crucial tool for assessing the popularity and quality of fermented milk, and the sensory evaluation results of fermented milk produced by the 20 isolate combinations are shown in Figure 3C. In the current sensory evaluation, the 100-point scale covered five attributes (20 points per attribute): organization state, taste, special flavor, color and luster, and liking. Our results showed that, except for three poor viscosity combinations (IMAU95110, IMAU80842; IMAU20762, IMAU20588; and IMAU20450, IMAU20465), the overall sensory scores of the predicted positively interacting combinations were higher than those of the predicted negatively interacting combinations.

Accuracy of the LbStI prediction model

The experimental validation results showed that both the co-

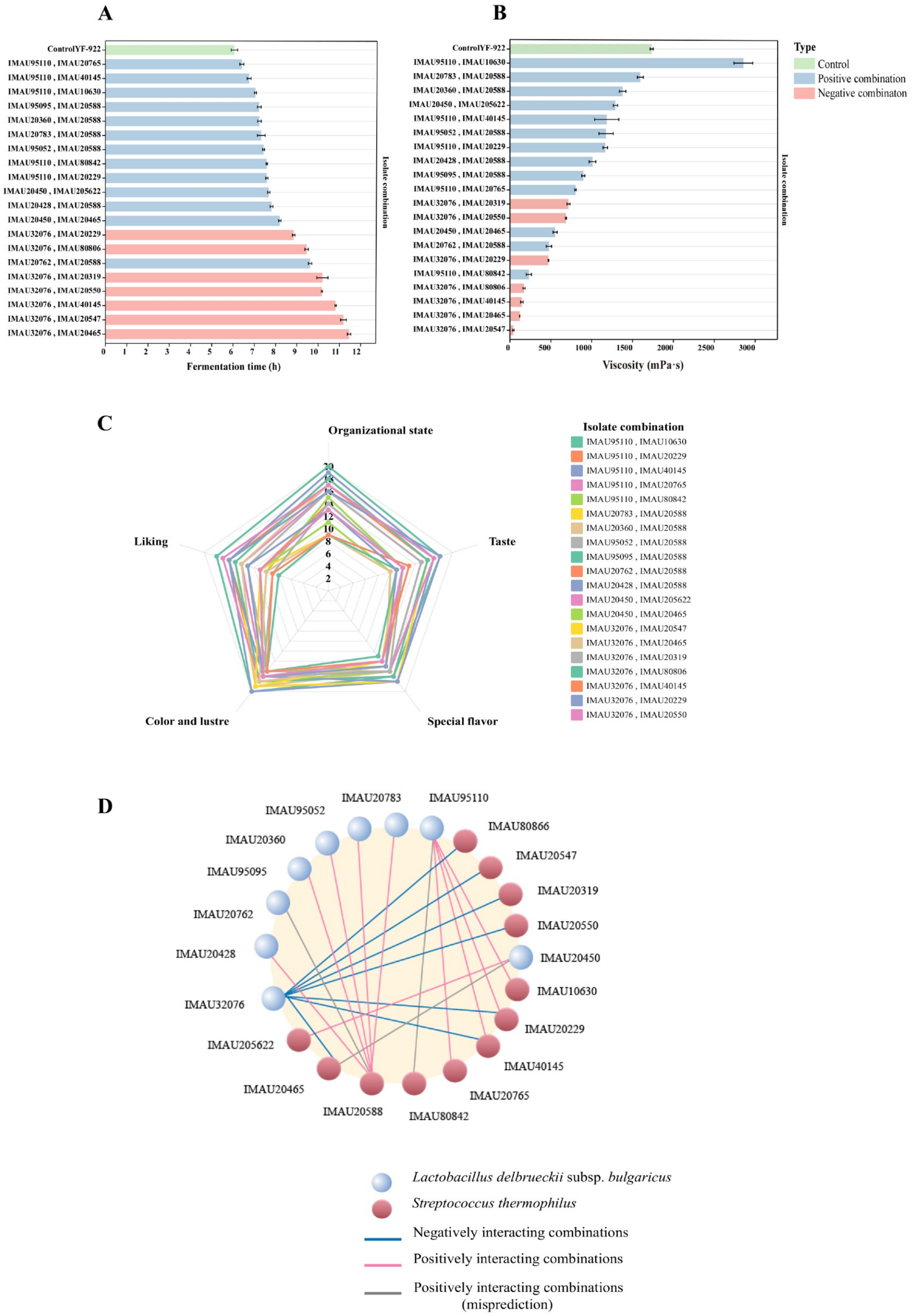

Figure 3. Fermented milk characteristics and microbial interactions of different isolate combinations. A, Fermentation time. B, viscosity of fermented milk produced by different isolate combinations. The green, blue, and peach bars represent the results of the control (a commercial starter, YF-922) and isolate combinations predicted to exhibit positive and negative interaction, respectively. Error bars represent standard deviation. C, Radar chart showing sensory attributes of 1-day ripened fermented milk produced by different isolate combinations. The sensory evaluation was based on five attributes: organizational state, taste, flavor, color and lust, and liking. Each attribute was graded on a 20-point scale: 0 (none) to 20 (extremely strong). The results for each combination are shown in a different color. D, Line chart showing the isolate combinations and their interactions. Blue and peach circles represent isolates of *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus*, respectively. The isolate interaction in milk fermentation was assessed based on the fermentation time and viscosity after 1-day ripening. Positively interacting combinations are connected by pink lines (indicating correct prediction of positive isolate combinations) and gray lines (indicating misprediction), respectively, while negatively interacting combinations are connected by blue lines.

clustering and *LbSt*IPred_SimLapRLS algorithms erroneously identified three combinations as positively interacting pairs. In fact, all three combinations exhibited a negative overall fermentation score, suggesting that they had non-mutualistic interaction as a starter culture during milk fermentation (Table 3, Figure 3D). Based on the validation results of the milk fermentation experiments, the *LbSt*I model achieved a highly accurate prediction rate of 85% (17/20).

DISCUSSION

This study used genome information from 181 *L. bulgaricus* and 181 *S. thermophilus* dairy isolates to construct a semi-supervised learning framework for predicting starter interactions in milk fermentation by co-clustering and LapRLS algorithms. The model showed a strong prediction accuracy, which was confirmed by five machine learning methods, LOOCV, and milk fermentation experiments; achieving an 85% accuracy rate (17/20) through validation with 20 randomly selected anticipated interacting isolate combinations. Meanwhile, our model construction and analysis process has disclosed intriguing biological insights on the metabolic interaction mechanisms between the starter isolates.

One strength of this study was the incorporation of putative genes and KEGG data in our model construction. The putative gene dataset represents the genomic potential of the microbe, while the KEGG database offers comprehensive biochemical pathways and information on metabolic interactions. Therefore, utilizing the complete gene and feature-fused KEGG datasets could uncover intriguing mechanistic details regarding biological interactions and synergistic effects of dairy starters during milk fermentation. To identify the factors that affect *L. bulgaricus* and *S. thermophilus* interactions, we calculated the average values with consideration of the presence/absence of genes and KOs in our datasets for 39 isolate combinations (19 labeled combinations and the 20 final predicted combinations). From these results, we identified the top 10 differential genes and KOs exhibiting the lowest *P* values in Mann-Whitney *U* tests between the positively and negatively interacting isolate combinations (Figure 4A, Table 4). Some of these factors may have important functions in initiating starter interactions in the milk fermentation process, particularly K01005, K05846, K14652 (*ribBA*), K23304, *opuCA*, *opuCB_2*, *opuCC*, and *epsF*. These genes and KOs are primarily involved in cysteine, riboflavin, teichoic acid, and exopolysaccharide biosynthesis, as well as ATP-binding cassette (ABC) transporter systems. Overall, the positively interacting isolate combinations have a higher number of these differential genes and KOs compared with those with negative interactions, suggesting that these gene functions and metabolic pathways are beneficial to milk fermentation.

K23304 is a serine O-acetyltransferase participating in the biosynthesis of cysteine. Cysteine biosynthesis mainly involves two steps: the conversion of L-serine to O-acetyl-L-serine and the subsequent formation of L-cysteine. The first step involves serine O-acetyltransferase, and the process of biochemical conversion is depicted in Figure 4B. The two substrates of the enzyme are acetyl coenzyme A and L-serine, while its two products are coenzyme A and O-acetyl-L-serine. Cysteine is an important natural sulfur-containing amino acid, which has many functions, such as promoting growth, antioxidation, and detoxification. Furthermore, it serves as the metabolic precursor of various essential biomolecules, including vitamins, cofactors, antioxidants, and many defense compounds (Alvarez et al., 2012). It plays a pivotal role in the biological system of animals. The two starter species, *S. thermophilus* and *L. bulgaricus*, require nitrogen sources in the form of small peptides and amino acids for growth, and fermentation can only occur when their amino acid requirements are met (Liu et al., 2016). Analysis of the growth media of 15 isolates of *S. thermophilus* revealed that this species relies on amino acids, such as cysteine, glutamic acid, and methionine, for growth. The absence of these amino acids impedes their growth (Letort and Juillard, 2001). *Lactobacillus delbrueckii* subsp. *bulgaricus* strain ND02 could increase the level of cysteine and other metabolites in the whey after fermentation, demonstrating its ability to degrade protein into peptides and amino acids (Peng et al., 2020a). Therefore, it can be inferred that cysteine facilitates metabolic cooperation and promotes growth interactions between *L. bulgaricus* and *S. thermophilus*.

K14652 (*ribBA*) is part of the riboflavin biosynthetic (*rib*) operon and encodes a 3,4-dihydroxy 2-butanone 4-phosphate synthase or GTP cyclohydrolase II that catalyzes the first step in riboflavin (vitamin B2) biosynthesis by hydrolyzing the initial substrate, GTP. The synthesis of riboflavin is a complex chemical reaction (Figure 4C). Riboflavin is an essential micronutrient for the human body and is acquired through exogenous food or nutritional supplements. Adequate riboflavin intake is crucial for disease prevention and treatment (Zhang et al., 2022a). Although lactic acid bacteria are typically deficient in multiple vitamins, some can synthesize certain B vitamins, such as folic acid (vitamin B9), riboflavin (vitamin B2), and cobalamin (vitamin B12) (Capozzi et al., 2012; Leblanc et al., 2011). Given its importance to human health and the ubiquity of deficiency, riboflavin has become one of the most studied vitamins produced by lactic acid bacteria (Pacheco Da Silva et al., 2016). Since riboflavin is also an essential growth factor for lactic acid bacteria (Le Boucher et al., 2013), we hypothesize that it plays a critical role in supporting the co-culture and that positively interacting isolate combinations can produce more riboflavin during metabolism than the negatively or non-interacting pairs.

K01005 encodes a polyisoprenyl-teichoic acid-peptidoglycan teichoic acid transferase, involved in teichoic acid biosynthesis. Teichoic acid encompasses a diverse family of cell surface glycopolymers containing phosphodiester-linked polyol repeat units and is a unique cell wall component of Gram-positive bacteria. We speculate that the substantial difference in K01005 function between positively and negatively interacting isolate combinations could contribute to their variation in cell growth and cell wall synthesis. However, our hypothesis awaits further experimental confirmation.

Some identified differential KO and genes encode proteins annotated as mineral and organic ion transporters under ABC transporters in KEGG categorization and are specialized in osmotic protectant uptake (opu; Figure 4D). ATP-binding cassette transporters are one of the largest known protein families and are widely distributed in bacteria. These transporters facilitate the active transport of various substrates, such as ions, sugars, lipids, sterols, peptides, proteins, and drugs, by coupling ATP hydrolysis. K05846 corresponds to the *opuBD* gene and encodes a permease protein in the osmoprotectant transport system, which is a component of the bacterial *opuB* ABC transport system. Interestingly, some components of the *opuC* ABC transport system (*opuCC*, *opuCB_2*, and *opuCA*) were also

Table 3. Validation of predicted results by milk fermentation experiments

Isolate combination (respective *Lb*, *St* isolates)	Fermentation time (h, mean±SD, in triplicate)	Viscosity (mPa·s, mean±SD, in triplicate)	Overall fermentation score	Co-clustering prediction	LPBSL rank
IMAU95110, IMAU80842	7.59±0.03 (1)	231±31.90 (−2)	−1	Positive	3
IMAU95110, IMAU10630	7.05±0.05 (1)	2856±114.47 (2)	3	Positive	13
IMAU95110, IMAU40145	6.76±0.09 (2)	1184±148.71 (2)	4	Positive	19
IMAU95110, IMAU20765	6.42±0.10 (2)	802±9.17 (0)	2	Positive	61
IMAU95110, IMAU20229	7.59±0.05 (1)	1166±27.06 (2)	3	Positive	99
IMAU20450, IMAU205622	7.68±0.06 (1)	1290±26.15 (2)	3	Positive	172
IMAU20450, IMAU20465	8.20±0.06 (0)	552±26.15 (−2)	−2	Positive	237
IMAU20783, IMAU20588	7.33±0.19 (1)	1596±36.50 (2)	3	Positive	379
IMAU95095, IMAU20588	7.23±0.09 (1)	905±18.33 (1)	2	Positive	405
IMAU95052, IMAU20588	7.43±0.05 (1)	1176±87.16 (2)	3	Positive	409
IMAU20428, IMAU20588	8.22±0.07 (0)	1010±40.84 (1)	1	Positive	507
IMAU20360, IMAU20588	7.27±0.08 (1)	1378±39.95 (2)	3	Positive	510
IMAU20762, IMAU20588	9.62±0.08 (−2)	476±33.05 (−2)	−4	Positive	549
IMAU32076, IMAU20465	11.46±0.08 (−2)	120±0 (−2)	−4	Negative	2200
IMAU32076, IMAU20319	10.20±0.26 (−2)	750±18.9 (−1)	−3	Negative	2217
IMAU32076, IMAU20550	10.18±0.03 (−2)	686±6.93 (−2)	−4	Negative	2226
IMAU32076, IMAU20229	8.85±0.06 (0)	470±3.46 (−2)	−2	Negative	2227
IMAU32076, IMAU80806	9.46±0.08 (−1)	172±12.49 (−2)	−3	Negative	2243
IMAU32076, IMAU20547	11.19±0.14 (−2)	40±10.44 (−2)	−4	Negative	2255
IMAU32076, IMAU40145	10.82±0.03 (−2)	146±15.10 (−2)	−4	Negative	2285

Notes: The outcome of milk fermentation of a specific combination of *Lactobacillus delbrueckii* subsp. *bulgaricus* (*Lb*) and *Streptococcus thermophilus* (*St*) isolates was assessed by two indicators, namely the fermentation time required to reach the fermentation endpoint (4.5<pH<4.6) and the fermented milk viscosity after 1-day day ripening; the indicator scores are written in brackets. The fermentation time score ranged from 2 (high fermentation rate) to −2 (low fermentation rate); 2 (<7 h), 1 (⩾7 h but <8 h), 0 (⩾8 h but <9 h), −1 (⩾9 h but <10 h), and −2 (⩾10 h), respectively. The viscosity score ranged from 2 (high viscosity) to −2 (low viscosity); 2 (>1,000 mPa·s), 1 (⩾900 but <1,000 mPa·s), 0 (⩾800 but <900 mPa·s), −1 (⩾700 mPa·s but <800 mPa·s), and −2 (⩾700 mPa·s). The overall fermentation score was calculated by adding the scores of fermentation time and viscosity. The three inaccurately predicted LPBSL results are shown in bold font.

identified as significantly differential genes between positively and negatively interacting isolates. In *Bacillus subtilis*, the cells import various osmostress protectants in the cell cytoplasm under hyperosmotic conditions via five osmotically controlled transport systems (*opuA* to *opuE*). *OpuB* specializes in importing choline, which is required for the biosynthesis of glycine betaine, while o*puC* imports a broad spectrum of compatible solutes including choline and glycine betaine (Rath et al., 2020). Osmotic protectants are essential in maintaining cell expansion pressure and inhibiting cell osmotic imbalance, especially in environments of low water potential and high ion pressure, to prevent hypertonic shock (Kiousi et al., 2022). Therefore, the presence of such genes in the positively interacting isolates likely helps in maintaining cell stability even amid dynamic metabolic conditions during fermentation.

Another differential gene is *epsF*, which is part of the exopolysaccharide gene cluster responsible for the production of extracellular polymeric substances (Figure 4E) (Lamothe et al., 2002; Wu et al., 2014). The complete exopolysaccharide gene cluster encompasses components responsible for regulation, chain-length determination, biosynthesis (by the glycosyltransferase) and polymerization of repeating units, and export of exopolysaccharides (Wu et al., 2023). The *epsF* encodes the glycosyltransferase and is conserved across all major phylogenetic groups (Stingele et al., 1999). The study of Folkenberg et al. (2006) observed that the interactions between different exopolysaccharide-producing *S. thermophilus* isolates and non-exopolysaccharide-producing *L. bulgaricus* could substantially change the texture of fermented milk. Moreover, capsular polysaccharide-producing isolates contribute to the high viscosity and desired mouth thickness and creaminess of fermented milk (Folkenberg et al., 2006). Thus, *epsF* is likely the reason for the higher viscosity and better texture of fermented milk produced by the positively interacting isolates over the negatively interacting ones.

This work utilized genomic data from numerous food-derived *S. thermophilus* and *L. bulgaricus* isolates to construct a rigorously validated model for predicting dairy starter interactions. This model aids in the selection for starter isolate combinations, improving milk fermentation efficiency. In the process of model building, we innovatively applied three algorithms of distinct clustering methods, with consideration of the biological information obtained from the isolate gene composition and KEGG features, maximizing the prediction accuracy. While the model exhibited high accuracy, it is not without limitations. For example, the current model consists of two low-coupling sub-models: co-clustering and Laplacian regularized least squares. These sub-models were combined to generate the final results by accepting their intersecting prediction sets. However, the two sub-models were not fully integrated at a deeper level of construction. This lack of comprehensive integration may hinder the capability of the model to capture complex interactions and potential synergies between different components. Moreover, it is noteworthy that we did not compare our method with other existing approaches for two reasons. Firstly, we had a limited amount of labeled data available, consisting of only 19 pairs of isolates. This small dataset poses challenges when attempting meaningful comparisons with methods that rely on supervised

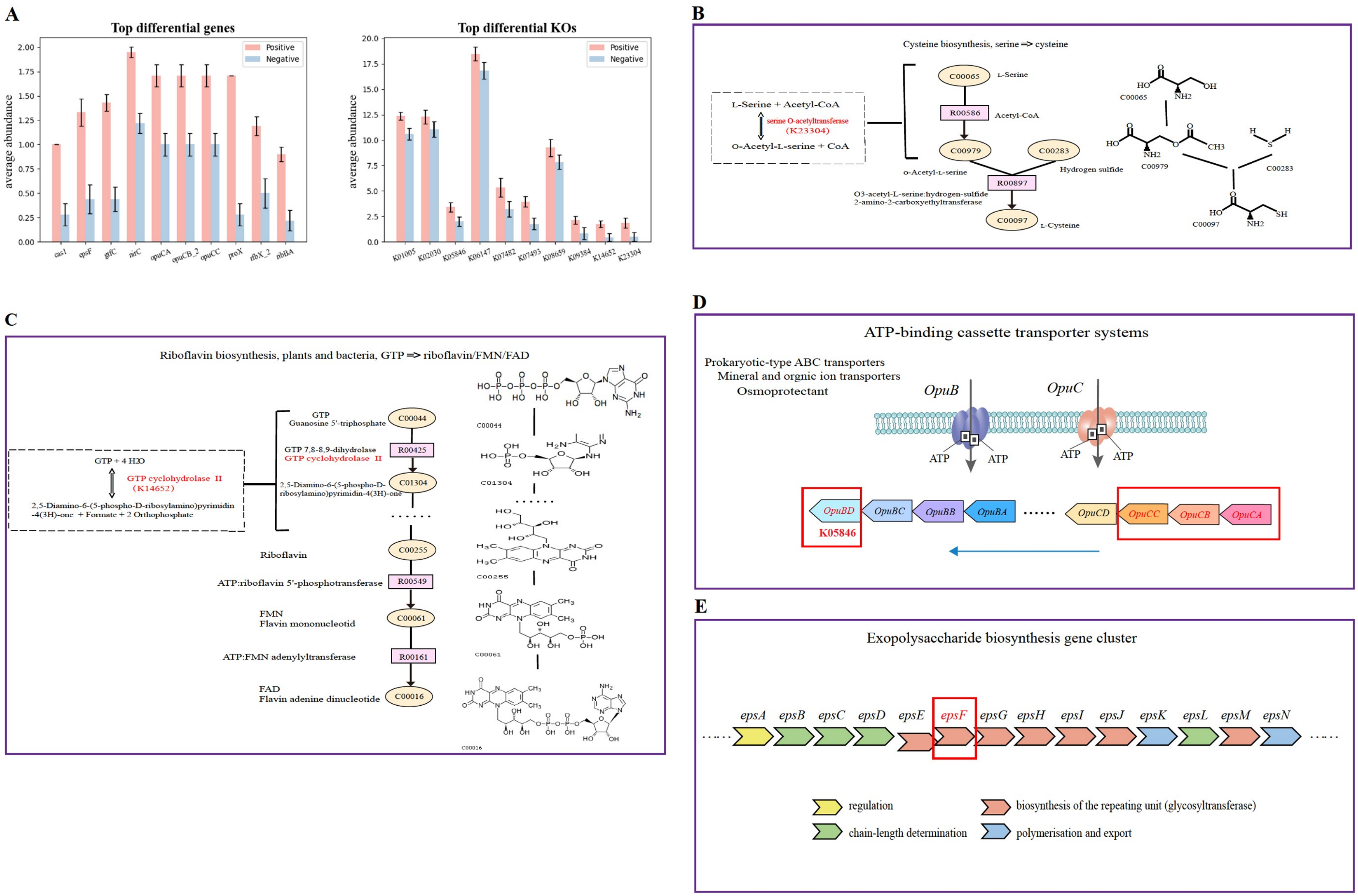
A
Top differential genes
Positive
Negative
average abundance
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
cas1
epsF
glfC
nirC
opuCA
opuCB_2
opuCC
proX
rfbX_2
ribBA
Top differential KOs
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
K01005
K02030
K05846
K06147
K07482
K07493
K08659
K09384
K14652
K23304
B
Cysteine biosynthesis, serine ⇒ cysteine
L-Serine + Acetyl-CoA
serine O-acetyltransferase (K23304)
O-Acetyl-L-serine + CoA
C00065
L-Serine
R00586
Acetyl-CoA
C00979
O-Acetyl-L-serine
C00283
Hydrogen sulfide
R00897
O3-acetyl-L-serine:hydrogen-sulfide 2-amino-2-carboxyethyltransferase
C00097
L-Cysteine
C
Riboflavin biosynthesis, plants and bacteria, GTP ⇒ riboflavin/FMN/FAD
GTP + 4 H2O
GTP cyclohydrolase II (K14652)
2,5-Diamino-6-(5-phospho-D-ribosylamino)pyrimidin-4(3H)-one + Formate + 2 Orthophosphate
GTP
Guanosine 5'-triphosphate
C00044
GTP 7,8-8,9-dihydrolase
GTP cyclohydrolase II
R00425
2,5-Diamino-6-(5-phospho-D-ribosylamino)pyrimidin-4(3H)-one
C01304
Riboflavin
C00255
ATP:riboflavin 5'-phosphotransferase
R00549
FMN
Flavin mononucleotid
C00061
ATP:FMN adenylyltransferase
R00161
FAD
Flavin adenine dinucleotide
C00016
D
ATP-binding cassette transporter systems
Prokaryotic-type ABC transporters
Mineral and orgnic ion transporters
Osmoprotectant
OpuB
OpuC
ATP
ATP
ATP
ATP
OpuBD
K05846
OpuBC
OpuBB
OpuBA
OpuCD
OpuCC
OpuCB
OpuCA
E
Exopolysaccharide biosynthesis gene cluster
epsA
epsB
epsC
epsD
epsE
epsF
epsG
epsH
epsI
epsJ
epsK
epsL
epsM
epsN
regulation
chain-length determination
biosynthesis of the repeating unit (glycosyltransferase)
polymerisation and export

Figure 4. Key functional differences between positively and negatively interacting isolate combinations. A, Bar charts showing the top differential genes (identified from the gene composition of all isolates dataset, GCAI dataset) and Kyoto Encyclopedia of Genes and Genomes orthologs (KOs), showing the top 10 differential genes (KOs) between positively and negatively interacting isolate combinations (see Table 4 for the *P* value of each gene and KO comparison). Error bars represent Standard Deviation. B, The cysteine biosynthesis pathway functions to transform L-serine to L-cysteine. In the process, serine O-acetyltransferase transfers acetyl-CoA to L-serine to form O-acetyl-L-serine, which further reacts with hydrogen sulfide to generate L-cysteine. C, Riboflavin biosynthesis pathway involves the transformation of GTP to riboflavin. In the first step, GTP cyclohydrolase II catalyzes GTP to 2,5-diamino-6-(5-phospho-d-robosylamino) pyrimidin-4(3h)-one. D, Examples of bacterial ABC transporter systems are illustrated. Osmotic protectant uptake (Opu) proteins import solutes into the cell cytoplasm to maintain cell expansion pressure and inhibit cell osmotic imbalance. E, Schematic representation of a typical exopolysaccharide gene cluster of lactic acid bacteria.

Table 4. Top 20 differential genes and KOs between positively and negatively interacting isolate combinations

Top differential genes and KOs	*P* value, Mann-Whitney *U* test
cas1	0.00000272
epsF	0.00016648
gtfC	0.00007490
nirC	0.00000431
opuCA	0.00014164
opuCB_2	0.00014164
opuCC	0.00014164
proX	0.00000272
rfbX_2	0.00034107
ribBA	0.00002160
K01005	0.00005365
K02030	0.00627884
K05846	0.00014164
K06147	0.00182316
K07482	0.00102930
K07493	0.00001106
K08659	0.00400906
K09384	0.00034157
K14652	0.00001945
K23304	0.00004671

Notes: The top 20 significantly different genes (determined from the gene composition of all isolates dataset (GCAI dataset) by considering the presence/absence of specific genes) and KEGG orthologs (KOs). These differential genes and KOs exhibited the lowest *P* value (Mann-Whitney *U* test) between the positively and negatively interacting isolate combinations across the complete dataset. Only the 19 labeled and 20 predicted isolate combinations were included in this analysis.

learning techniques and larger datasets. Secondly, in the field of bioinformatic interactions, most approaches employ supervised learning where each data point has corresponding labels, enabling more robust training and evaluation processes (Lei et al., 2021; Li et al., 2023a). Alternatively, semi-supervised learning with a larger amount of labeled data is employed in other cases (Dalkıran et al., 2023). Our modeling approach leans towards unsupervised learning, in which the model learns patterns and relationships from the unlabeled data itself. Direct comparison of model performance is impractical and inconclusive when the research methods and underlying principles are fundamentally different.

These limitations highlight the need for future research efforts to explore ways to strengthen model coupling and amass a larger amount of labeled data for more comprehensive evaluations and comparisons with existing methods. Nonetheless, this study effectively constructed a predictive model and garnered an understanding of the interactions between dairy starters during milk fermentation (Figure 5).

This study developed a semi-supervised learning framework (*LbSt*I prediction model) for predicting the interactions between the two starter bacteria, *L. bulgaricus* and *S. thermophilus*, in milk fermentation, with a high accuracy of 85%. Our prediction model also revealed that the mutualistic starter isolates interactions in optimal fermentation rely on cysteine, riboflavin, teichoic acid, and exopolysaccharide biosynthesis and the function of ABC transport. The model and data generated in this study will aid in screening starter cultures and guide the development of fermented products.

MATERIALS AND METHODS

Background information of bacterial isolates

The genomic information of 362 *L.* bulgaricus and *S. thermophilus* isolates (181 isolates per species) was used in this study. Most isolates were obtained from naturally fermented milk

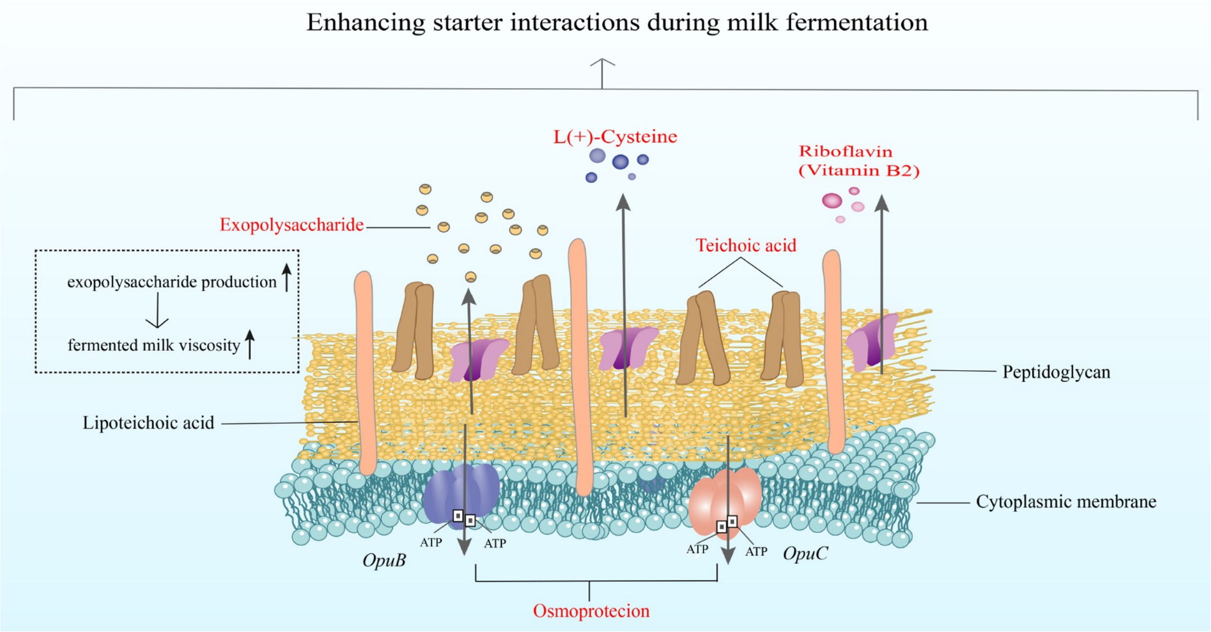

Figure 5. Schematic diagram showing proposed mechanisms of enhanced interactions between starter isolates of *Lactobacillus delbrueckii* subspecies *bulgaricus* and *Streptococcus thermophilus* during milk fermentation. The differential metabolic pathways of cysteine, riboflavin, exopolysaccharides, teichoic acid, and osmotic protectant transporter function are crucial determinants of starter combination interactions. Our findings suggest that these factors influence the outcomes of starter interactions.

samples collected from three different countries (Figure 1; Tables S2 and S3 in Supporting Information), including different geological locations in China (Gansu, Nei Monggol, Qinghai, Sichuan, Xizang, and Xinjiang), Mongolia (Bulgan, Central Mongolia, Dornod, Erhan, Govi-Altay, Hou Hanggai, Hovsgol, Kent, Oriental, Qian Hanggai, Selenge, Suhbaatar, Ulanbator, and Zavkhan), and Russia (Tuva). Two *L.* bulgaricus isolates were industrial starter cultures originating from Denmark and Bulgaria, respectively. .All strains were provided by the National Collection of Microbial Resource for Feed (Nei Monggol).

The genomes of all isolates were previously sequenced by Illumina HiSeq high-throughput sequencing, and high-quality data were selected for genome assembly and subsequent analyses (https://figshare.com/s/4471858b0681af3ea270; Tables S2 and S3 in Supporting Information) (Song et al., 2021; Zhao et al., 2021). The genome sequencing data are available in the National Center for Biotechnology Information database (BioProject IDs: PRJNA594100 for *S. thermophilus*; PRJNA594100 for *L. bulgaricus*).

Positive and negative training samples

In milk fermentation, different combinations of *L. bulgaricus* and *S. thermophilus* are used as starter cultures. The 181 *L. bulgaricus* and 181 *S. thermophilus* in our dataset could generate a total of 32,761 (181×181) unique isolate combinations. The study aimed to establish a one-on-one screening method to detect isolated interactions from these 32,761 combinations. To accomplish this, 19 isolate combinations were selected randomly and trained as positive and negative samples.

To determine their milk fermentation efficiency, we evaluated the fermentation time required to reach the endpoint of pH 4.5 to 4.6 and viscosity after 1-day ripening. The fermentation time score ranged from 2 (high fermentation rate) to −2 (low fermentation rate); where 2 indicated a time less than seven hours, 1 was for time between seven and eight hours, 0 between eight and nine hours, −1 between nine and ten hours, and −2 indicated time greater than or equal to ten hours. The viscosity score ranged from 2 (high viscosity) to −2 (low viscosity), with 2 indicating a high viscosity of greater than 1,000 mPa·s, 1 representing a viscosity of 900 to less than 1,000 mPa·s, 0 representing a viscosity of 800 to less than 900 mPa·s, −1 representing a viscosity of 700 to less than 800 mPa·s, and −2 referring to a low viscosity of less than 700 mPa·s.

The overall fermentation score was calculated by adding the scores of fermentation time and viscosity. Combinations with a positive sum score were considered as having potential for a positive starter culture interaction, while those with a negative sum score were deemed to have a negative starter culture interaction. Subsequent analyses labeled them as positive and negative interactions, respectively (Table 5).

Datasets

Our study analyzed the *LbSt*I utilizing three genomic sequence-based datasets, which were the KEGG dataset, the K-mer analysis dataset, and the gene composition of all isolates (GCAI) dataset (Tables S4 and S5 in Supporting Information). The KEGG dataset is an extensive database of biological information that covers genes, chemicals, metabolic pathways, cell signaling, and human diseases in life sciences. We utilized metabolic pathway data for our study. The term K-mer refers to a substring of length k, typically employed to describe a continuous sequence of DNA or RNA. Our study implemented a 4–8mer matrix. The GCAI dataset comprised the entire set of unique genes from all isolates, totaling 7,026 genes. Each gene in the GCAI matrix was assigned a value of 1 or 0, depending on its presence or absence in a specific isolate.

Model overview

The model was created through three stages: data preprocessing, model building, and evaluation (Figure 2). In the preprocessing phase, K-mers analysis and KEGG feature fusion were performed on the data (Figure 2A). The *LbSt*I prediction was based on a semi-supervised learning framework that integrated two models (Figure 2B). In the co-clustering model, fused KEGG data was subjected to data dimension reduction by the Uniform Manifold Approximation and Projection (UMAP) method before the actual co-clustering process to obtain the *LbSt*I prediction results. The predicted results were validated by machine-learning methods. Another model, *LbSt*I Prediction based on Similarity-fusion LapRLS (*LbSt*IPred_SimLapRLS), implemented the LapRLS method for *LbSt*I prediction using the similarity matrix, K-mer matrix, GCAI, and the interaction matrix. The predicted results were validated by leave-one-out cross-validation (LOOCV). A final set of results was generated by combining the predicted outcomes of the two models, which were further verified by milk fermentation experiments (Figure 2C).

Data preprocessing

In the data preprocessing stage, we acquired a 4-8mer dataset through K-mer analysis and a feature-fused KEGG dataset by concatenating the KEGG data of the two species (Figure 2A). The feature fusion step was necessary because of the occurrence of inter- and intraspecific genomic variations, which are represented by the presence of both unique and overlapping genes in and between species or isolates. Thus, it is necessary to perform a first step to align the KEGG ortholog (KO) profile of *L. bulgaricus* and *S. thermophilus* genomes to enable direct KO-based numerical addition when performing the one-on-one KO profile comparison between the two species. Specifically, the genomes of the *L. bulgaricus* and *S. thermophilus* species contained 1,295 and 504 non-overlapping features, respectively, so the total number of features of each species would add up to 1,799, and the numerical value of 0 would be assigned to the increased features in the respective isolate. The second step of the generation of the feature-fused KEGG dataset was the actual KO addition operation performed between each starter isolate combination, resulting in a total of 32,761 fused KEGG data from the two sets of 181 isolates of the two species. The KEGG database contains a vast amount of information regarding biochemical pathways and metabolic interactions. If the values of identical or related KEGG pathways from two isolates in a starter combination result in a non-zero-sum, this potentially indicates a connection between these pathways. This connection further implies a growth interaction when the two isolates are co-cultured. By utilizing the feature-fused-KEGG dataset, it is possible to analyze and understand the interdependent interactions and synergistic effects of various combinations of dairy starters during milk fermentation.

https://doi.org/10.1007/s11427-023-2569-7

Table 5. Combination of starter cultures and fermentation outcome

Isolate combination (respective *Lb*, *St* isolates)	Interaction label	Code	Fermentation time (h, mean±SD, in triplicate)	Viscosity (mPa·s, mean±SD, in triplicate)	Overall fermentation score
IMAU20450, IMAU80844	Positive	P1	7.42±0.36 (1)	1,720±78.50 (2)	3
IMAU20450, IMAU20543	Positive	P2	7.72±0.03 (1)	1,698±63.38 (2)	3
IMAU20450, IMAU20588	Positive	P3	7.47±0.17 (1)	1,706±78.05 (2)	3
IMAU20450, IMAU20774	Positive	P4	7.23±0.03 (1)	1,688±54.51 (2)	3
IMAU20450, IMAU40133	Positive	P5	6.53±0.03 (2)	1,765±4.58 (2)	4
IMAU62091, IMAU40133	Positive	P6	8.17±0.03 (0)	1,805±39.51 (2)	2
IMAU95110, IMAU80844	Positive	P7	7.88±0.01 (1)	1,731±52.00 (2)	3
IMAU95110, IMAU20543	Positive	P8	7.83±0.03 (1)	1,693±58.20 (2)	3
IMAU95110, IMAU20588	Positive	P9	8.38±0.02 (0)	1,707±12.13 (2)	2
IMAU95110, IMAU20774	Positive	P10	6.75±0.12 (2)	1,714±75.19 (2)	4
IMAU95110, IMAU40133	Positive	P11	6.63±0.03 (2)	1,754±33.86 (2)	4
IMAU20312, IMAU40133	Negative	N1	10.2±0.02 (−2)	685±70.50 (−2)	−4
IMAU20450, IMAU20766	Negative	N2	9.75±0.02 (−1)	657±68.00 (−2)	−3
IMAU62081, IMAU40133	Negative	N3	9.53±0.05 (−1)	721±16.04 (−1)	−2
IMAU62161, IMAU40133	Negative	N4	9.87±0.01 (−1)	735±18.03 (−1)	−2
IMAU32076, IMAU80844	Negative	N5	10.22±0.03 (−2)	662±60.00 (−2)	−4
IMAU32076, IMAU20543	Negative	N6	11.55±0.04 (−2)	697±78.35 (−2)	−4
IMAU32076, IMAU20588	Negative	N7	10.41±0.03 (−2)	674±8.14 (−2)	−4
IMAU32076, IMAU40133	Negative	N8	7.43±0.04 (1)	638±7.51 (−2)	−1

Notes: The outcome of milk fermentation of a specific combination of *Lactobacillus delbrueckii* subsp. *bulgaricus* (*Lb*) and *Streptococcus thermophilus* (*St*) isolates was assessed by two indicators, namely the fermentation endpoint (4.5 < pH < 4.6) and the fermented milk viscosity after 1-day ripening; the indicator scores are written in brackets. The fermentation time score ranged from 2 (high fermentation rate) to −2 (low fermentation rate); 2 (<7 h), 1 (⩾7 h but <8 h), 0 (⩾8 h but <9 h), −1 (⩾9 h but <10 h), and −2 (⩾10 h), respectively. The viscosity score ranged from 2 (high viscosity) to −2 (low viscosity); 2 (>1,000 mPa·s), 1 (⩾900 but < 1,000 mPa·s), 0 (⩾800 but <900 mPa·s), −1 (⩾700 mPa·s but <800 mPa·s), and −2 (⩾700 mPa·s). The overall fermentation score was calculated by adding the scores of fermentation time and viscosity. Combinations having a positive and a negative sum score were considered to have a potentially positive and negative starter culture interaction, respectively. Each combination was given a different code.

Co-clustering prediction

The co-clustering model is a statistical technique that clusters data from various perspectives and standards to predict *LbSt*I in this case. This is achieved by classifying unlabeled data through labeled distribution. This model could enhance prediction accuracy by establishing complementary relationships between predicted data and perspectives. The process involves initial data screening, UMAP reduction of data dimension, and multiple clustering using K-means, Gaussian mixture model (GMM), and balanced iterative reducing and clustering using hierarchies (BIRCH).

The computational load of clustering was reduced by conducting a preliminary screening of 32,761 combinations based on cosine similarity between unlabeled and labeled combinations (Chen et al., 2021c) by using formula (1):

$$\cos\ similarity = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \qquad (1)$$

Where A_i and B_i represented each component of the vectors, *A* and *B*, respectively.

The preliminary data screening involved analyzing the difference in similarity among labeled combinations using similarity. We analyzed the average similarity between all candidate combinations and 11 positive combinations (P1–P11), selecting the top 3,000 with the highest similarities. From these 3,000 combinations, the 1,000 combinations that showed the least average similarity with the 8 labeled negative combinations (N1-N8) were chosen as the primary positive combinations. Similarly, we calculated the average similarities between all the candidate combinations and N1-N8, selecting the top 3,000 with the highest similarities. From these 3,000 combinations, the 1,000 combinations that showed the least average similarity with the 11 labeled positive combinations (P1–P11) were chosen as the prior negative combinations.

After the initial data screening, we conducted a cluster prediction using the UMAP dimension reduction method, which decreased the 1,799 dimensions (primary positive combinations, prior negative combinations, and 19 labeled combinations) to three dimensions. Three clustering algorithms (K-means, Gaussian Mixture Model, and Balanced Iterative Reducing and Clustering using Hierarchies) were then used to cluster the data. The results of these three clustering algorithms divided all combinations into two groups. Ideally, all labeled positive combinations were assigned to one group, while all labeled negative combinations were assigned to the other group. However, empirical experiments revealed that achieving this ideal result was challenging, and only a relatively optimal outcome was attainable (Table 6). For example, K-means clustering achieved an accuracy rate of 16 out of 19, but three positively labeled combinations were mistakenly clustered into group 1. The unlabeled combinations in groups 0 and 1 were considered positive and negative isolate combinations, respectively.

If all three clustering techniques yielded somewhat optimal results, the final prediction of interacting and non-interacting combinations was determined by intersecting the data from the positive and negative clusters of each technique. If the clustering algorithms did not reach the relative optimum, the UMAP

Table 6. Results of clustering analyses

Isolate combination code	K-means	GMM	BIRCH
P1	0	1	0
P2	0	1	0
P3	0	1	0
P4	0	1	0
P5	1	0	1
P6	1	0	1
P7	0	1	0
P8	0	1	0
P9	0	1	0
P10	0	1	0
P11	1	0	0
N1	1	0	1
N2	1	0	0
N3	1	0	1
N4	1	0	1
N5	1	0	0
N6	1	0	1
N7	1	0	1
N8	1	0	1
Clustering precision rate	16/19	16/19	15/19

Notes: three clustering analyses were performed, namely K-means, GMM, and BIRCH, to assign each isolate combination to cluster 0 or 1 in each case. Presumably, isolate combinations should form clusters based on a positive or negative overall fermentation score; isolate combinations that did not cluster correctly are written in bold font.

operation would be repeated with a re-adjustment of the random_state parameter until all three clustering algorithms satisfied the requirements. Congruent results from the three clustering algorithms were regarded as the predicted results of the co-clustering model. This process ensured the best possible outcome.

LbStIPred_SimLapRLS prediction

To predict potential combinations of *LbSt*I, we developed a semi-supervised learning framework, namely *LbSt*I Prediction Framework based on Similarity-fusion LapRLS (*LbSt*IPred-SimLap) for interaction prediction (Chen et al., 2016; Chen et al., 2012). The development process involved three steps: constructing an interaction matrix from known *LbSt*I combinations (Chen et al., 2021b), calculating interspecies KEGG-GCAI fusion similarity of individual *L. bulgaricus* and *S. thermophilus*, and developing the LapRLS prediction module, which combined the *L. bulgaricus* and *S. thermophilus* classifiers.

The interaction matrix M was constructed using data from 181 *L. bulgaricus* and 181 *S. thermophilus*, which were 181×181 in size. The initial values of all combinations were set to 0, awaiting the prediction of the probability of interactions. Then, the values of labeled combinations were modified, with positive and negative values set to 1s and –1 s, respectively. These labeled combinations can be used as prior knowledge in semi-supervised learning to predict the interaction probability for unlabeled combinations. The final construction of interaction matrix M is shown in Table 7.

We calculated the cosine similarity of the 181 *L. bulgaricus* isolates and constructed the similarity matrix (S_L), with K-mer (4–8 mer) and GCAI using formulae (2) and (3).

$$S_{LB\_kmer} = \alpha \times \sum_{k=4}^{8} S_{LB\_\text{K-mer}} \tag{2}$$

$$S_L = \alpha S_{LB\_\text{K-mer}} + \left(1 - \alpha\right) S_{LB\_GCAI} \tag{3}$$

We used the weighted averaging method for the *L. bulgaricus* similarity integration, which assigned equal weight to each of the similar measures. α was the weighting factor, and it was set to 0.2 (formula 3) and 0.5 (formula 4) in this work. The cosine similarity matrix (S_S) for the 181 *S. thermophilus* isolates was calculated using the same method. After constructing the interaction matrix M and similarity fusion matrices ($S_L$and S_S), LapRLS was used to construct the *LbSt*I predictor (Figure 2B).

To implement the LapRLS Predictor, the diagonal matrices for the *L. bulgaricus* and *S. thermophilus* isolates (D_L and D_S) must first be defined. D_L and D_S were defined such that $D_L(i, i)$ and $D_S(i, i)$ were the sum of rows of $S_L$and S_S, respectively (Chen et al., 2021b). Then, the normalized Laplacian similarity matrices were calculated by using formulae (4) and (5):

$$L_L = (D_L)^{-1/2}(D_L - S_L)(D_L)^{-1/2} \tag{4}$$

$$L_S = (D_S)^{-1/2}(D_S - S_S)(D_S)^{-1/2} \tag{5}$$

The predictor, P^*, was obtained based on the theoretical assumption of previous works (Chen et al., 2016; Chen et al., 2012; Xia et al., 2010): if *Lb*1 and *St*1 could interact, then *Lb*2, which was very similar to *Lb*1, was assumed to be able to interact with *St*1. Based on the above assumption, formula (6) was used to define the optimal prediction function of *L. bulgaricus* space (P_L^*), which was essentially a cost function,

$$P_L^* = \arg\min_{P_L}[\,||\;M - P_L\;||_F^2\; + \theta_L\;||\;P_L^T L_L P_L\;||_F^2\,] \tag{6}$$

where $||\;.\;||_F$ was Frobenius norm, and θ_L was the weight parameter in the *L. bulgaricus* space. After that, we could solve formula (6) to get P_L^* using equation (7) (Xia et al., 2010). P_s^*could be calculated in a similar way using formula (8). It should be noted that when calculating P_s^*, M here must be transposed.

$$P_L^* = S_L(S_L + \theta_L L_L S_L)^{-1} M \tag{7}$$

$$P_S^* = S_S(S_S + \theta_S L_S S_S)^{-1} M^T \tag{8}$$

Finally, the ultimate predictor, P^*, was obtained by combining P_L^* and P_s^*, as shown in formula (9):

$$P^* = \frac{P_L^* + (P_S^*)^T}{2} \tag{9}$$

The final output of the predictor, P^*, was a score table arranged in descending order. The closer the top of the table, the higher the score, indicating a higher likelihood of *LbSt*I.

Verification with milk fermentation experiments

Preparation of freeze-dried powder of bacterial isolates

The experimental isolates were activated and subcultured being expanded in a 5 L reagent bottle with a high boron silicon thread mouth (Sichuan Shubo Co., Ltd., Chongzhou, China). They were then inoculated into de Man-Rogosa-Sharpe broth and M17

https://doi.org/10.1007/s11427-023-2569-7

broth for static fermentation in a 42°C constant temperature incubator (LBH-250 biochemical incubator, Shanghai Yiheng Scientific Instrument Co., Ltd., Shanghai, China). Bacterial cells were harvested by centrifugation (DL-6M centrifuge, Hunan Xiangyi Laboratory Instrument Development Co., Ltd, Changsha, China), mixed with a cryoprotective agent, and frozen in a freezer at −80°C for 24 h. Afterward, the frozen cells were subjected to vacuum freeze-drying for 48 h in a freeze dryer (EYALA DRC-1000/EYERA FDU-1100, Tokyo Physical and Chemical Instrument Co., Ltd., Japan), and the freeze-dried bacterial preparations were ground into powder and stored in aluminum foil bags after viable cell count. These steps ensured that the correct isolates were used in the fermentation experiments and that the stored frozen cells were still viable after the freeze-drying procedure.

Preparation of fermented milk

Stirred yogurt was prepared per the protocol described in Kearney et al. (2011). Briefly, 93.5% pure milk (pure milk containing 3% protein, Inner Mongolia Mengniu Dairy Co., Ltd., Hohhot, China) was used as the base material. The milk was preheated to 62°C to 65°C before 6.5% sucrose was added. The milk mixture was left for 10 min to allow the sucrose to dissolve, followed by homogenization under high pressure of 20 mPa·s (SHR 60-70, Shanghai Shenlu Homogenizer Co., Ltd., Shanghai, China), pasteurization at 95°C for 5 min, and cooling to 42°C in a water bath. The cooled milk was aseptically inoculated with the appropriate isolate combinations (*S. thermophilus* of 2×10^6 CFU mL^{-1} and *L. bulgaricus* of 2×10^4 CFU mL^{-1}, corresponding to a species inoculation ratio of 100:1) and fermented in a 42°C incubator until reaching the fermentation endpoint of pH 4.5 to pH 4.6. The fermented milk was cooled in an ice water bath for 30 min. Then, the fermented milk was ripened in a refrigerator at 4°C for 24 h.

Evaluation of characteristics of fermented milk after one day of ripening

(i) Determination of pH. The pH level of fermented milk was measured thrice by a FE28 pH meter (Mettler Toledo, USA).

(ii) Determination of titratable acidity. The titratable acidity of fermented milk was determined by mixing 5 g of it with 20 mL of boiled and cooled distilled water. The mixture was then titrated with 0.1 N NaOH in the presence of 0.5% phenolphthalein indicator (Bai et al., 2020). The measurement was performed in triplicate.

(iii) Determination of viscosity. The viscosity of fermented milk was measured thrice with a Brookfield DV-1 viscometer (Brookfield Co., Middleboro, MA, USA) using a No. 4 rotor, torque range of 10%–100%, rotating speed of 100 r min^{-1}, and measuring time of 30 s (Dan et al., 2018).

(iv) Determination of water-holding capacity. Twenty grams of

Table 7. Interaction matrix M

	St_1	St_2	St_3	St_4	St_5	St_6	St_7	…	St_181
Lb_1	1	1	1	1	1	−1	0	…	0
Lb_2	0	0	0	0	1	0	0	…	0
Lb_3	1	1	1	1	1	0	0	…	0
Lb_4	0	0	0	0	−1	0	0	…	0
Lb_5	0	0	0	0	−1	0	0	…	0
Lb_6	0	0	0	0	−1	0	0	…	0
Lb_7	−1	−1	−1	0	−1	0	0	…	0
…	…	…	…	…	…	…	…	…	0
Lb_181	0	0	0	0	0	0	0	0	0

Notes: The data of 181 each of *Lactobacillus delbrueckii* subsp. *bulgaricus* (*Lb*) and *Streptococcus thermophilus* (*St*) were analyzed in this work. The subfix number code represents the isolate number. "…": The table shows a matrix with 181 rows and 181 columns.

Table 8. Sensory evaluation scoring standards

Sensory attribute	Description of scoring standard	Score (20 per attribute)
Organizational state	The state is even and delicate, without stratification, bubbles, and whey precipitation.	15–20
	The state is relatively uniform, without stratification, but with a small amount of whey precipitation.	10–15
	The state is uneven, with curd lumps or particles, obvious stratification, and a large amount of whey precipitation.	0–10
Taste	Moderate acidity, delicate and smooth taste.	15–20
	Excessive or insufficient acidity imparts a less delicate and smooth taste.	10–15
	No sour or rough taste, granular or sandy.	0–10
Special flavor	There is a mellow yogurt smell; no peculiar smell.	15–20
	Suitable sweet and sour balance; a dull aroma, but no peculiar smell.	10–15
	Sweet and sour imbalance; no yogurt smell or with peculiar smell.	0–10
Color and luster	The color is very uniform, milky white or yellowish.	15–20
	The color is relatively uniform, light yellow or light gray.	10–15
	The color is dark, uneven, and abnormal.	0–10
Liking	Like the product very much.	15–20
	Generally, like the product.	10–15
	Do not like the product.	0–10

fermented milk was filtered through a funnel with a piece of qualitative filter paper at room temperature for 2 h. The filtrate was collected and weighed. The water-holding capacity (%) was calculated as (1−(filtrate mass/sample mass))×100.

(v) Sensory evaluation of fermented milk. A sensory evaluation was conducted on the ripened fermented milk by a team of ten food researchers and dairy processing professionals on the first day of ripening. The sensory quality of fermented milk was evaluated objectively from five aspects: organizational state, taste, flavor, color, and liking (Table 8). Each attribute was scored out of 20 points, with a maximum score of 100 points (Szajnar et al., 2020).

Compliance and ethics

The author(s) declare that they have no conflict of interest.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (2022YFD2100700), the National Natural Science Foundation of China (32325040), Basic Scientific Research Business Fee Project of Universities Directly (BR22-14-01), the National Dairy Science and Technology Innovation Center (2022-Open Subject -6), Inner Mongolia Natural Science Foundation Project (2021MS06023), Inner Mongolia Science & Technology planning project (2022YFSJ0017), and the earmarked fund for China Agricultural Research System (CARS36).

Supporting information

The supporting information is available online at https://doi.org/10.1007/s11427-023-2569-7. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Abedin, M.M., Chourasia, R., Phukon, L.C., Sarkar, P., Ray, R.C., Singh, S.P., and Rai, A.K. (2023). Lactic acid bacteria in the functional food industry: biotechnological properties and potential applications. Crit Rev Food Sci Nutr 5, 1–19.

Alvarez, C., Ángeles Bermúdez, M., Romero, L.C., Gotor, C., and García, I. (2012). Cysteine homeostasis plays an essential role in plant immunity. New Phytol 193, 165–177.

Bai, M., Huang, T., Guo, S., Wang, Y., Wang, J., Kwok, L.Y., Dan, T., Zhang, H., and Bilige, M. (2020). Probiotic *Lactobacillus casei* Zhang improved the properties of stirred yogurt. Food Biosci 37, 100718.

Bintsis, T. (2018). Lactic acid bacteria as starter cultures: an update in their metabolism and genetics. AIMS Microbiol 4, 665–684.

Capozzi, V., Russo, P., Dueñas, M.T., López, P., and Spano, G. (2012). Lactic acid bacteria producing B-group vitamins: a great potential for functional cereals products. Appl Microbiol Biotechnol 96, 1383–1394.

Chen, X., Li, T.H., Zhao, Y., Wang, C.C., and Zhu, C.C. (2021a). Deep-belief network for predicting potential miRNA-disease associations. Brief BioInf 22, bbaa186.

Chen, X., Liu, M.X., Cui, Q.H., and Yan, G.Y. (2012). Prediction of disease-related interactions between MicroRNAs and environmental factors based on a semi-supervised classifier. PLoS ONE 7, e43425.

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. PLoS Comput Biol 12, e1004975.

Chen, X., Sun, L.G., and Zhao, Y. (2021b). NCMCMDA: miRNA-disease association prediction through neighborhood constraint matrix completion. Brief BioInf 22, 485–496.

Chen, X., Zhou, C., Wang, C.C., and Zhao, Y. (2021c). Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization. Brief BioInf 22, bbab328.

Dalkıran, A., Atakan, A., Rifaioğlu, A.S., Martin, M.J., Atalay, R.C., Acar, A.C., Doğan, T., and Atalay, V. (2023). Transfer learning for drug-target interaction prediction. Bioinformatics 39, i103–i110.

Dan, T., Jin, R., Ren, W., Li, T., Chen, H., and Sun, T. (2018). Characteristics of milk fermented by *Streptococcus thermophilus* MGA45-4 and the profiles of associated volatile compounds during fermentation and storage. Molecules 23, 878.

Dan, T., Wang, D., Wu, S., Jin, R., Ren, W., and Sun, T. (2017). Profiles of volatile flavor compounds in milk fermented with different proportional combinations of *Lactobacillus delbrueckii* subsp. bulgaricus and *Streptococcus thermophilus*. Molecules 22, 1633.

Deng, Y., Qiu, Y., Xu, X., Liu, S., Zhang, Z., Zhu, S., and Zhang, W. (2022). META-DDIE: predicting drug-drug interaction events with few-shot learning. Brief BioInf 23, bbab514.

Deshwal, G.K., Tiwari, S., Kumar, A., Raman, R.K., and Kadyan, S. (2021). Review on factors affecting and control of post-acidification in yoghurt and related products. Trends Food Sci Tech 109, 499–512.

Dong, W., Yang, Q., Wang, J., Xu, L., Li, X., Luo, G., and Gao, X. (2023). Multi-modality attribute learning-based method for drug-protein interaction prediction based on deep neural network. Brief BioInf 24, bbad161.

Folkenberg, D.M., Dejmek, P., Skriver, A., and Ipsen, R. (2006). Interactions between EPS-producing *Streptococcus thermophilus* strains in mixed yoghurt cultures. J Dairy Res 73, 385–393.

Ge, Y., Yu, X., Zhao, X., Liu, C., Li, T., Mu, S., Zhang, L., Chen, Z., Zhang, Z., Song, Z., et al., (2024). Fermentation characteristics and postacidification of yogurt by Streptococcus thermophilus CICC 6038 and *Lactobacillus delbrueckii* ssp. bulgaricus CICC 6047 at optimal inoculum ratio. J Dairy Sci 107, 123–140.

George, F., Daniel, C., Thomas, M., Singer, E., Guilbaud, A., Tessier, F.J., Revol-Junelles, A.M., Borges, F., and Foligné, B. (2018). Occurrence and dynamism of lactic acid bacteria in distinct ecological niches: a multifaceted functional health perspective. Front Microbiol 9, 2899.

Gu, J., Bang, D., Yi, J., Lee, S., Kim, D.K., and Kim, S. (2023). A model-agnostic framework to enhance knowledge graph-based drug combination prediction with drug-drug interaction data and supervised contrastive learning. Brief BioInf 24, bbad285.

Hatti-Kaul, R., Chen, L., Dishisha, T., and Enshasy, H.E. (2018). Lactic acid bacteria: from starter cultures to producers of chemicals. FEMS Microbiol Lett 365.

Jansen, J.E., Aschenbrenner, D., Uhlig, H.H., Coles, M.C., and Gaffney, E.A. (2022). A method for the inference of cytokine interaction networks. PLoS Comput Biol 18, e1010112.

Kearney, N., Stack, H.M., Tobin, J.T., Chaurin, V., Fenelon, M.A., Fitzgerald, G.F., Ross, R.P., and Stanton, C. (2011). *Lactobacillus paracasei* NFBC 338 producing recombinant beta-glucan positively influences the functional properties of yoghurt. Int Dairy J 21, 561–567.

Kiousi, D.E., Efstathiou, C., Tegopoulos, K., Mantzourani, I., Alexopoulos, A., Plessas, S., Kolovos, P., Koffa, M., and Galanis, A. (2022). Genomic insight into *Lacticaseibacillus paracasei* SP5, reveals genes and gene clusters of probiotic interest and biotechnological potential. Front Microbiol 13, 922689.

Lamothe, G., Jolly, L., Mollet, B., and Stingele, F. (2002). Genetic and biochemical characterization of exopolysaccharide biosynthesis by *Lactobacillus delbrueckii* subsp. bulgaricus. Arch Microbiol 178, 218–228.

LeBlanc, J.G., Laiño, J.E., del Valle, M.J., Vannini, V., van Sinderen, D., Taranto, M.P., de Valdez, G.F., de Giori, G.S., and Sesma, F. (2011). B-Group vitamin production by lactic acid bacteria-current knowledge and potential applications. J Appl Microbiol 111, 1297–1309.

Le Boucher, C., Courant, F., Jeanson, S., Chereau, S., Maillard, M.B., Royer, A.L., Thierry, A., Dervilly-Pinel, G., Le Bizec, B., and Lortal, S. (2013). First mass spectrometry metabolic fingerprinting of bacterial metabolism in a model cheese. Food Chem 141, 1032–1040.

Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. (2021). A deep-learning framework for multi-level peptide-protein interaction prediction. Nat Commun 12, 5465.

Letort, C., and Juillard, V. (2001). Development of a minimal chemically-defined medium for the exponential growth of *Streptococcus thermophilus*. J Appl Microbiol 91, 1023–1029.

Li, K., Quan, L., Jiang, Y., Wu, H., Wu, J., Li, Y., Zhou, Y., Wu, T., and Lyu, Q. (2023). Simultaneous prediction of interaction sites on the protein and peptide sides of complexes through multilayer graph convolutional networks. J Chem Inf Model 63, 2251–2262.

Li, Y., Qiao, G., Gao, X., and Wang, G. (2022). Supervised graph co-contrastive learning for drug-target interaction prediction 38, 2847–2854.

Li, Y.C., You, Z.H., Yu, C.Q., Wang, L., Hu, L., Hu, P.W., Qiao, Y., Wang, X.F., and Huang, Y.A. (2024). DeepCMI: a graph-based model for accurate prediction of circRNA-miRNA interactions with multiple information. Brief Funct Genomics 23, 276–285.

Lian, X., Yang, S., Li, H., Fu, C., and Zhang, Z. (2019). Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. J Proteome Res 18, 2195–2205.

Liu, E., Zheng, H., Shi, T., Ye, L., Konno, T., Oda, M., Shen, H., and Ji, Z.S. (2016). Relationship between *Lactobacillus bulgaricus* and *Streptococcus thermophilus* under whey conditions: focus on amino acid formation. Int Dairy J 56, 141–150.

Macori, G., and Cotter, P.D. (2018). Novel insights into the microbiology of fermented dairy foods. Curr Opin Biotechnol 49, 172–178.

Pacheco Da Silva, F.F., Biscola, V., LeBlanc, J.G., and Gombossy de Melo Franco, B.D. (2016). Effect of indigenous lactic acid bacteria isolated from goat milk and cheeses on folate and riboflavin content of fermented goat milk. LWT-Food Sci Tech 71, 155–161.

Peng, J., Li, D., Liu, Y., Zhang, W., and Sun, T. (2020a). Metabolic characteristics of *L.*

bulgaricus ND02 during whey fermentation (in Chinese). J Food Sci Biotechnol 39, 25–33.

Peng, J., Li, J., and Shang, X. (2020b). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. BMC BioInf 21, 394.

Peng, X., Lei, Y., Feng, P., Jia, L., Ma, J., Zhao, D., and Zeng, J. (2023). Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. Nat Mach Intell 5, 395–407.

Rath, H., Reder, A., Hoffmann, T., Hammer, E., Seubert, A., Bremer, E., Völker, U., and Mäder, U. (2020). Management of osmoprotectant uptake hierarchy in bacillus subtilis via a SigB-dependent antisense RNA. Front Microbiol 11, 622.

Settachaimongkon, S., Nout, M.J.R., Fernandes, E.C.A., Hettinga, K.A., Vervoort, J.J. M., Hooijdonk, A.C.M. van, et al., (2014). Influence of different proteolytic strains of *Streptococcus thermophilus* in co-culture with *Lactobacillus delbrueckii* subsp. *bulgaricus* on the metabolite profile of set-yoghurt, 177, 29–36.

Sharma, H., Ozogul, F., Bartkiene, E., and Rocha, J.M. (2023). Impact of lactic acid bacteria and their metabolites on the techno-functional properties and health benefits of fermented dairy products. Crit Rev Food Sci Nutr 63, 4819–4841.

Sieuwerts, S. (2016). Microbial interactions in the yoghurt consortium: current status and product implications. SOJMID 4, 01–05.

Song, Y., Zhao, J., Liu, W., Li, W., Sun, Z., Cui, Y., and Zhang, H. (2021). Exploring the industrial potential of *Lactobacillus delbrueckii* ssp. *bulgaricus* by population genomics and genome-wide association study analysis. J Dairy Sci 104, 4044–4055.

Stingele, F., Newell, J.W., and Neeser, J.R. (1999). Unraveling the function of glycosyltransferases in *Streptococcus thermophilus* Sfi6. J Bacteriol 181, 6354–6360.

Szajnar, K., Znamirowska, A., and Kuźniar, P. (2020). Sensory and textural properties of fermented milk with viability of *Lactobacillus rhamnosus* and *Bifidobacterium animalis* ssp. *lactis* Bb-12 and increased calcium concentration. Int J Food Properties 23, 582–598.

Wang, G., Liu, X., Wang, K., Gao, Y., Li, G., Baptista-Hon, D.T., Yang, X.H., Xue, K., Tai, W.H., Jiang, Z., et al., (2023). Deep-learning-enabled protein-protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. Nat Med 29, 2007–2018.

Wang, Y., Wu, J., Lv, M., Shao, Z., Hungwe, M., Wang, J., Bai, X., Xie, J., Wang, Y., and Geng, W. (2021b). Metabolism characteristics of lactic acid bacteria and the expanding applications in food industry. Front Bioeng Biotechnol 9, 612285.

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., and Lu, H. (2017). Deep-learning-based drug-target interaction prediction. J Proteome Res 16, 1401–1409.

Wu, J., Han, X., Ye, M., Li, Y., Wang, X., and Zhong, Q. (2023). Exopolysaccharides synthesized by lactic acid bacteria: biosynthesis pathway, structure-function relationship, structural modification and applicability. Crit Rev Food Sci Nutr 63, 7043–7064.

Wu, Q., Tun, H.M., Leung, F.C.C., and Shah, N.P. (2014). Genomic insights into high exopolysaccharide-producing dairy starter bacterium *Streptococcus thermophilus* ASCC 1275. Sci Rep 4, 4974.

Xia, Z., Wu, L.Y., Zhou, X., and Wong, S.T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol 4, S6.

Xu, H., Xu, D., Zhang, N., Zhang, Y., and Gao, R. (2021). Protein-protein interaction prediction based on spectral radius and general regression neural network. J Proteome Res 20, 1657–1665.

Yang, S., Bai, M., Kwok, L.Y., Zhong, Z., and Sun, Z. (2023). The intricate symbiotic relationship between lactic acid bacterial starters in the milk fermentation ecosystem. Crit Rev Food Sci Nutr doi: 10.1080/10408398.2023.2280706, 1–18.

Zannini, E., Waters, D.M., Coffey, A., and Arendt, E.K. (2016). Production, properties, and industrial food application of lactic acid bacteria-derived exopolysaccharides. Appl Microbiol Biotechnol 100, 1121–1135.

Zhang, J.R., Ge, Y.Y., Liu, P.H., Wu, D.T., Liu, H.Y., Li, H.B., Corke, H., and Gan, R.Y. (2022a). Biotechnological strategies of riboflavin biosynthesis in microbes. Engineering 12, 115–127.

Zhang, L., Wang, C.C., and Chen, X. (2022b). Predicting drug-target binding affinity through molecule representation block based on multi-head attention and skip connection. Brief BioInf 23, bbac468.

Zhao, J., Wu, L., Li, W., Wang, Y., Zheng, H., Sun, T., Zhang, H., Xi, R., Liu, W., and Sun, Z. (2021). Genomics landscape of 185 Streptococcus thermophilus and identification of fermentation biomarkers. Food Res Int 150, 110711.

Zhou, D., Xu, Z., Li, W., Xie, X., Peng, S. (2021). MultiDTI: drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. Bioinformatics 37, 4485–4492.