From Genotypes to Ecosystems: Unraveling Microbial Interactions through Machine Learning to Engineer Stable Synthetic Communities

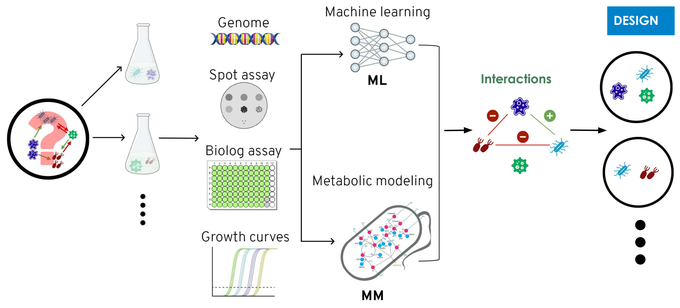
Dileep Kishore[1], Andrew Freiburger[2], Alexis L. Marsh[3], Myra B. Cohen[3], Christopher S. Henry[2], Dale A. Pelletier[1], Robert Cottingham[1], Mitchel J. Doktycz[1] and Priya Ranjan[1]

[1](Presenting author underlined) Oak Ridge National Laboratory, Oak Ridge, TN, USA; [2]Argonne National Laboratory, Lemont, IL, USA; [3]Iowa State University, Ames, IA, USA

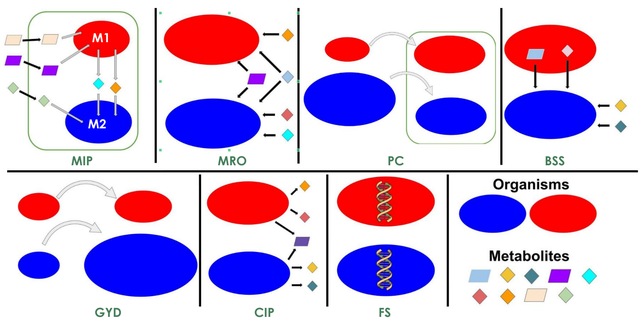
Introduction

Microbial communities are intricate ecosystems of microorganisms. Discerning each dimension of these interactions is crucial to unraveling the principles governing the range of community behavior and their dynamics. However, the complexity and scale of microbial communities lead to a combinatorial explosion of potential community interactions and pose significant challenges for computational modeling and experimental exploration.

In this project, we propose a machine-learning framework that predicts metabolic and inhibitory interactions in a microbial community using a suite of novel features derived from the members' genotypes. The microbial interaction classifier is being designed to accurately predict the interaction type and strength between community members in a given environment. This classifier uses several novel and orthogonal features: scoring metrics of metabolic cooperation and competition, secondary metabolic inhibition, and functional similarity of the annotated genomes. The classifier is being trained on a large dataset of microbial interactions, and is expected to accurately predict interactions between diverse microbes in new environments and identify stable combinations that can guide the rational design of a synthetic microbial community. This computational approach can circumvent experimental limitations, improve the efficiency of selecting communities worthy of study, and mechanistically investigate microbial community dynamics.



Prediction of metabolic interactions: CommScores

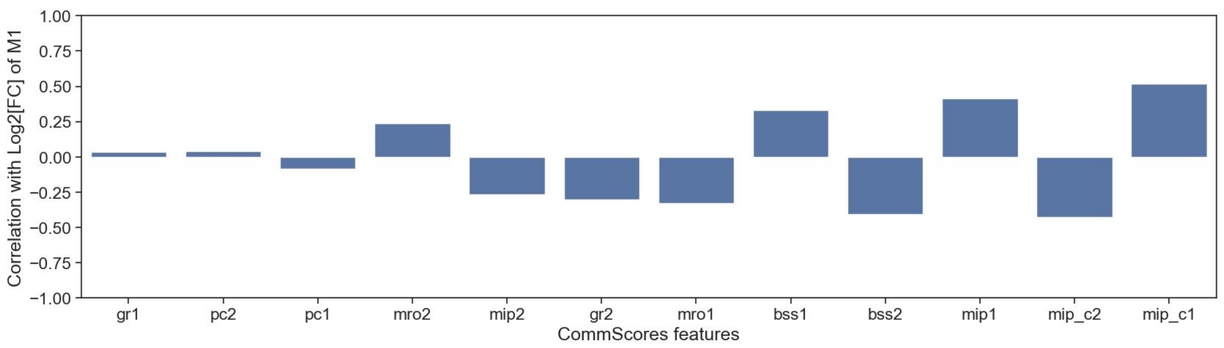


CommScores are a set of scores that encompass various dimensions of metabolic interactions in microbial communities.

Scores	Biological dimensions	
Functional Similarity (**FS**)	ecological competition	Novel
Growth Yield Difference (**GYD**)	outperformance potential	Novel
Costless Interaction Potential (**CIP**)	potential costless syntrophy	Novel
Biological Interaction Type (**BIT**)	interaction category	Novel
(forthcoming) pH Perturbation	inhibitory potential	Novel
Metabolic Interaction Potential (**MIP**)	simulated syntrophy	Curated
Metabolic Resource Overlap (**MRO**)	resource competition	Curated
Potential Cooperation (**PC**)	growth synergy	Curated
Biosynthetic Support Score (**BSS**)	parasitic potential	Curated
(forthcoming) Metabolic Support Index (**MSI**)	metabolic synergy	Curated

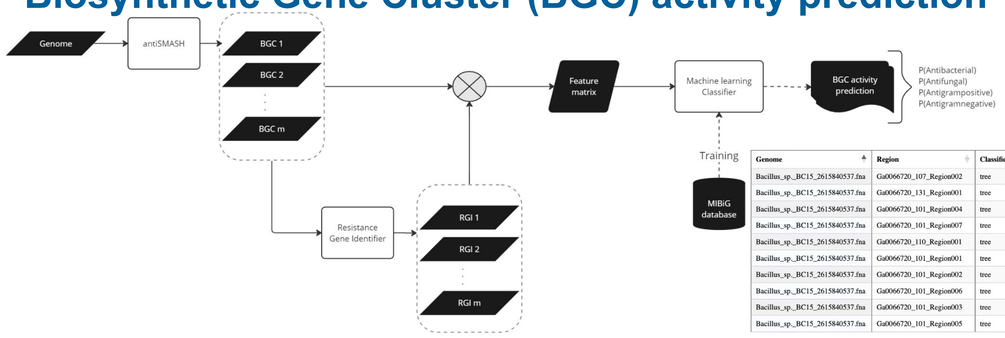
Validation with *Arabidopsis* microbiome

We validate CommScores by recapitulating an experimental study of pairwise interactions in an *Arabidopsis* microbiome (Schafer et al., 2023). Here, we compare CommScores metrics to fold changes in co-culture abundances on *Arabidopsis* leaf.



The correlation of each CommScores metric against fold changes in co-cultures. We observe that the cross-feeding scores (MIP and BSS) elicited the greatest correlations with member fold-changes.

Prediction of inhibition: Biosynthetic Gene Cluster (BGC) activity prediction



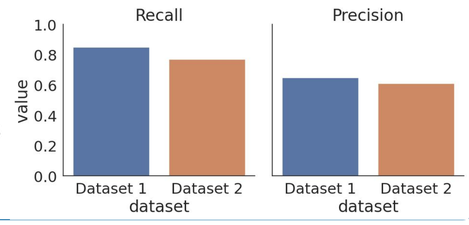
Genome	Region	Classifier	antibacterial	antifungal	antigrampos	antigramneg
Bacillus_sp._BC15_2615840537.fna	Ga0066720_107_Region002	tree	0.505	0.309	0.446	0.200
Bacillus_sp._BC15_2615840537.fna	Ga0066720_131_Region001	tree	0.924	0.034	0.605	0.040
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region004	tree	0.620	0.381	0.660	0.200
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region007	tree	0.640	0.262	0.560	0.280
Bacillus_sp._BC15_2615840537.fna	Ga0066720_110_Region001	tree	0.800	0.491	0.580	0.280
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region001	tree	0.261	0.026	0.188	0.040
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region002	tree	0.620	0.301	0.580	0.240
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region006	tree	0.656	0.048	0.425	0.120
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region003	tree	0.540	0.223	0.400	0.200
Bacillus_sp._BC15_2615840537.fna	Ga0066720_101_Region005	tree	0.746	0.028	0.604	0.400

Probabilities of antibiotic activities of BGCs in a genome

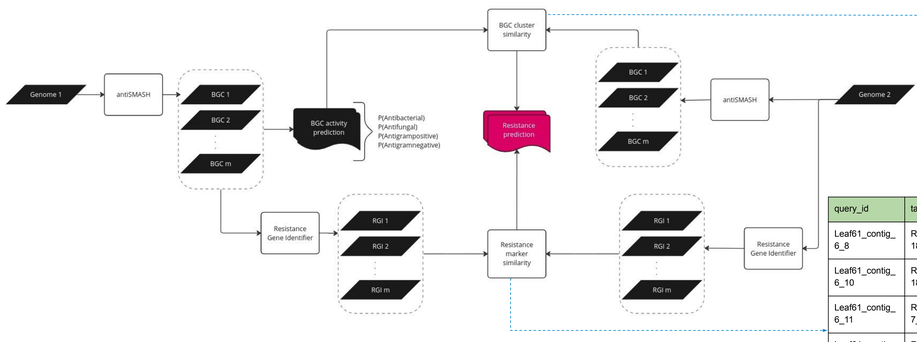
We developed a machine learning workflow that predicts the probabilities of antibiotic activity of the BGCs of a given genome.

Prediction of inhibitory microbes

We validated our BGC activity prediction pipeline using experimental inhibition screening data from two datasets: Dataset 1 (Getzke et al., 2023) containing 206 strains and Dataset 2 (Helfrich et al., 2018) containing 198 strains. Our workflow predicts >80% of the inhibitory microbes in both datasets.



Prediction of inhibition: Biosynthetic Gene Cluster (BGC) resistance prediction



source_bgc	target_bgc	cluster_similarity
Leaf61_contig_1.region001	Leaf182_contig_34.region001	0.13
Leaf61_contig_1.region002	Leaf182_contig_14.region001	0.00
Leaf61_contig_2.region001	Leaf182_contig_14.region002	0.02
Leaf61_contig_2.region002	Leaf182_contig_24.region001	0.00

BGC cluster similarity results

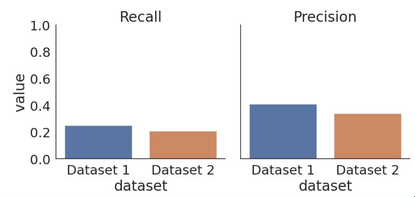
query_id	target_id	query_protein	target_protein	score
Leaf61_contig_6_8	Root401_contig_18_95	MPS…	MSTCATPP…	13.61
Leaf61_contig_6_10	Root401_contig_18_7	MPSIP…	MQLIRRITLK…	121.43
Leaf61_contig_6_11	Root401_contig_7_121	MDTITTNVTQL…	MSQELRLIRRITLK…	752.64
Leaf61_contig_6_12	Root401_contig_3_80	MSAAQ…	MAVQIA…	255.98

Resistance marker similarity results

We developed a (data-independent) pipeline that predicts the resistance of a given genome to an antibiotic BGC.

Prediction of resistant microbes

We validated our BGC activity prediction pipeline using experimental inhibition screening data from two datasets: Dataset 1 (Getzke et al., 2023) containing 206 strains and Dataset 2 (Helfrich et al., 2018) containing 198 strains. Currently, our pipeline predicts ~40% of the resistant microbes from the two datasets. Next, we plan to optimize parameters and thresholds to improve our prediction accuracy.



Prediction of pairwise interactions (combining metabolic and BGC features)

We built a random forest classifier that can predict pairwise interactions between microbes given their genomes. Currently this classifier uses the following feature for prediction:

- Metabolic network (CommScores)
- Biosynthetic gene cluster information (antiSMASH)
- Functional similarity (RAST)

In the future, we plan to add features from the inhibition prediction pipelines

RAST | CommScores | BGCs

	Functional similarity	MRO	MIP	GRD	NRPS	…
M1-M2	0.682	0.805	3	0.124	1	…
M1-M3	0.534	0.557	6	0.569	0	…
M2-M3	0.395	0.702	5	0.467	2	…
…	…					
M221-M200	0.522	0.667	0	0.001	5	…

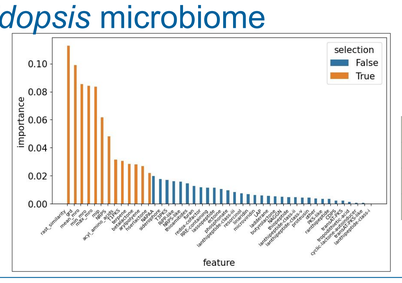
Features

Prediction

	Interaction
M1-M2	negative
M1-M3	non-negative
M2-M3	negative
…	…
M221-M200	non-negative

Predicting interactions in the *Arabidopsis* microbiome

We validated our microbial interaction classifier through experimental pairwise interaction data of microbes (224 strains) isolated from the *Arabidopsis* phyllosphere (Helfrich et al., 2018). Our classifier shows high accuracy in predicting non-negative interactions and also provides insights into the important features



	precision	recall	f1-score	support
negative	0.78	0.51	0.62	162
non-negative	0.85	0.95	0.90	462
accuracy			0.84	624

Future work

1. Optimization of the parameters used in the inhibition prediction pipelines
2. Extraction of features from the inhibition prediction pipelines and their incorporation into the pairwise interaction classifier
3. Test and engineer stable synthetic communities based on the results of the pairwise interaction classifier

References

- Getzke, Felix, M. Amine Hassani, Max Crüsemann, Milena Malisic, Pengfan Zhang, Yuji Ishigaki, Nils Böhringer, et al. "Cofunctioning of Bacterial Exometabolites Drives Root Microbiota Establishment." Proceedings of the National Academy of Sciences 120, no. 15 (April 11, 2023): e2221508120.
- Helfrich, Eric J. N., Christine M. Vogel, Reiko Ueoka, Martin Schäfer, Florian Ryffel, Daniel B. Müller, Silke Probst, Markus Kreuzer, Jörn Piel, and Julia A. Vorholt. "Bipartite Interactions, Antibiotic Production and Biosynthetic Potential of the Arabidopsis Leaf Microbiome." Nature Microbiology 3, no. 8 (August 2018): 909–19.
- Schäfer, Martin, Alan R. Pacheco, Rahel Künzler, Miriam Bortfeld-Miller, Christopher M. Field, Evangelia Vayena, Vassily Hatzimanikatis, and Julia A. Vorholt. "Metabolic Interaction Models Recapitulate Leaf Microbiota Ecology." Science 381, no. 6653 (July 7, 2023): eadf5121.
- Walker, Allison S., and Jon Clardy. "A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters." Journal of Chemical Information and Modeling 61, no. 6 (June 28, 2021): 2560–71.

Acknowledgement:
Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. This program is supported by the U. S. Department of Energy, Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research, under FWP ERKPA39.