MACHINE LEARNING IDENTIFICATION OF HONEY BEE BIO-MARKERS USING MICROBIOME AND PROTEOME DATA

Sidki Bouslama[1], Pierre-Luc Mercier[1], Pierre Giovenazzo[2], Leonard Foster[3], Amro Zayed[4], Nicolas Derome[1]

[1]Institute for Integrative and Systems Biology(IBIS), Université Laval, Québec, QC, Canada
[2]Departement of Biology, Université Laval, Québec, QC, Canada
[3]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC, Canada
[4]Departement of Biology, York University, Toronto, ON, Canada

contact: sidki.bouslama.1@ulaval.ca

MOTIVATION

Recent studies have highlighted the importance of the microbiome on host health on a variety of organisms, including *Apis mellifera*, the european honey bee, where the gut microbiome was shown to be involved in carbohydrate processing, metabolite synthesis, immunity, etc..

The economic value of honey bees and other pollinators in Canada and around the world cannot be stressed enough, and current yearly colony losses in Canada motivated an effort to better understand the adversities and stressors at play.

In that regard, our work seeks to employ novel machine-learning methods in order to explore the possibility of using the honey bee microbiome as a bio-marker for host health.

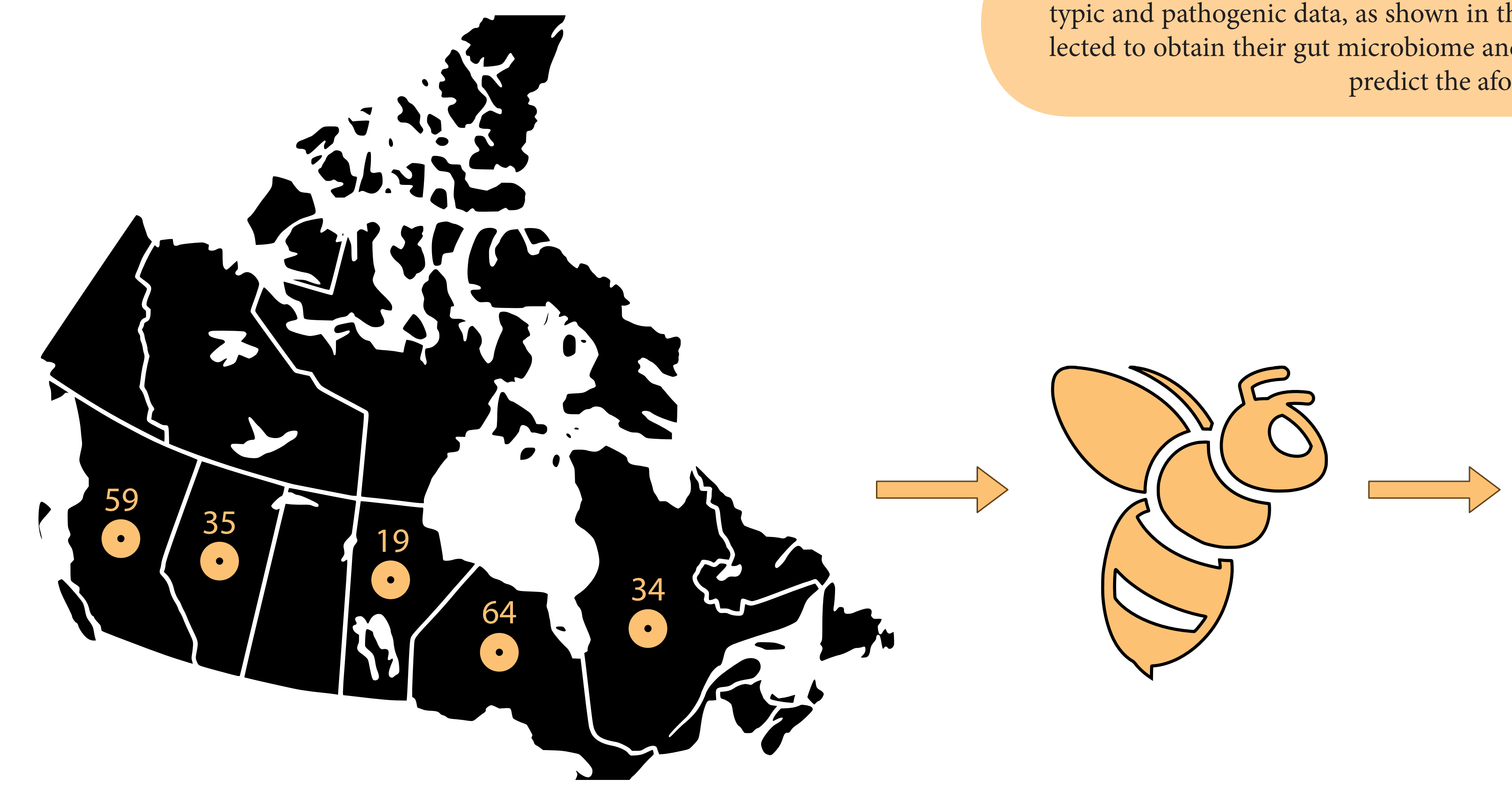
APPROACH

16S metagenomic data and host proteomic data were used to train a machine learning model in order to predict values of metrics pertaining to phenotypic or pathogenic caracteristics. An implementation of the Random forest algorithm was used over multiple iterations and multiple randomized subsamples in order to minimize the risk of overfitting. Machine learning models with an acceptable predictive power were analyzed in order to extract elements in the metagenomic/proteomic data that contributed most to the predictions.

These elements were then used to produce a shorter and simpler dataset that was used to produce a machine learning model with similar or better predictive capacity, the end result expected being a short list of bio-markers with a significant predictive power on the biological factors/stressors at play.

DESIGN

In total, 211 colonies were sampled from across Canada. Each colony was thouroughly screened for various phenotypic and pathogenic data, as shown in the diagram below. For each colony, random individual honey bees were selected to obtain their gut microbiome and proteome. This data was fed to our machine learning pipeline in order to predict the aforementionned phenotypic and pathogenic traits.



16S GUT MICROBIOME DATA

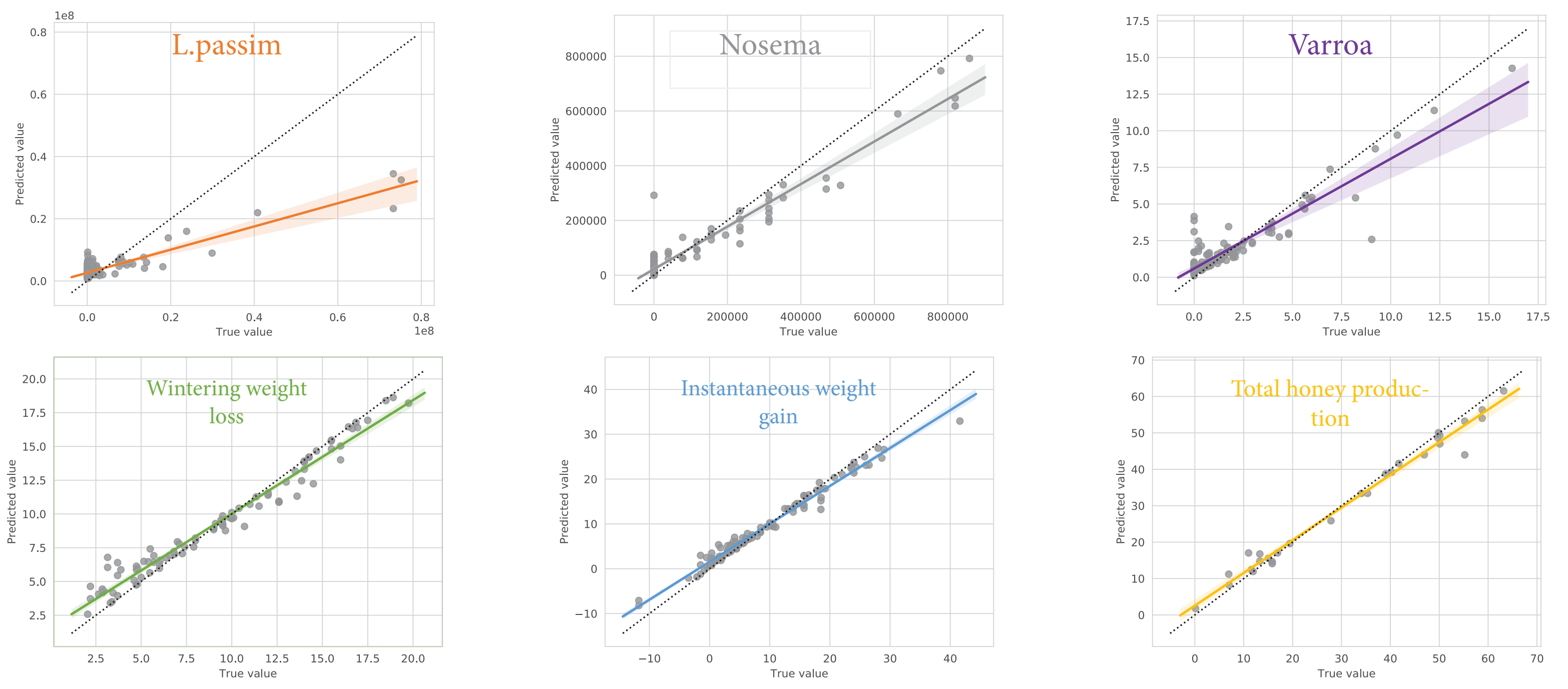
HOST PROTEOME DATA

PREDICTS

- Phenotypic data (wintering weight loss, supervized weight gain, total honey production)
- Parasitic load (nosema, lotmaria passim)
- Colony location
- Viral load (BQCV, DWV, SBV, IAPV)

RESULTS

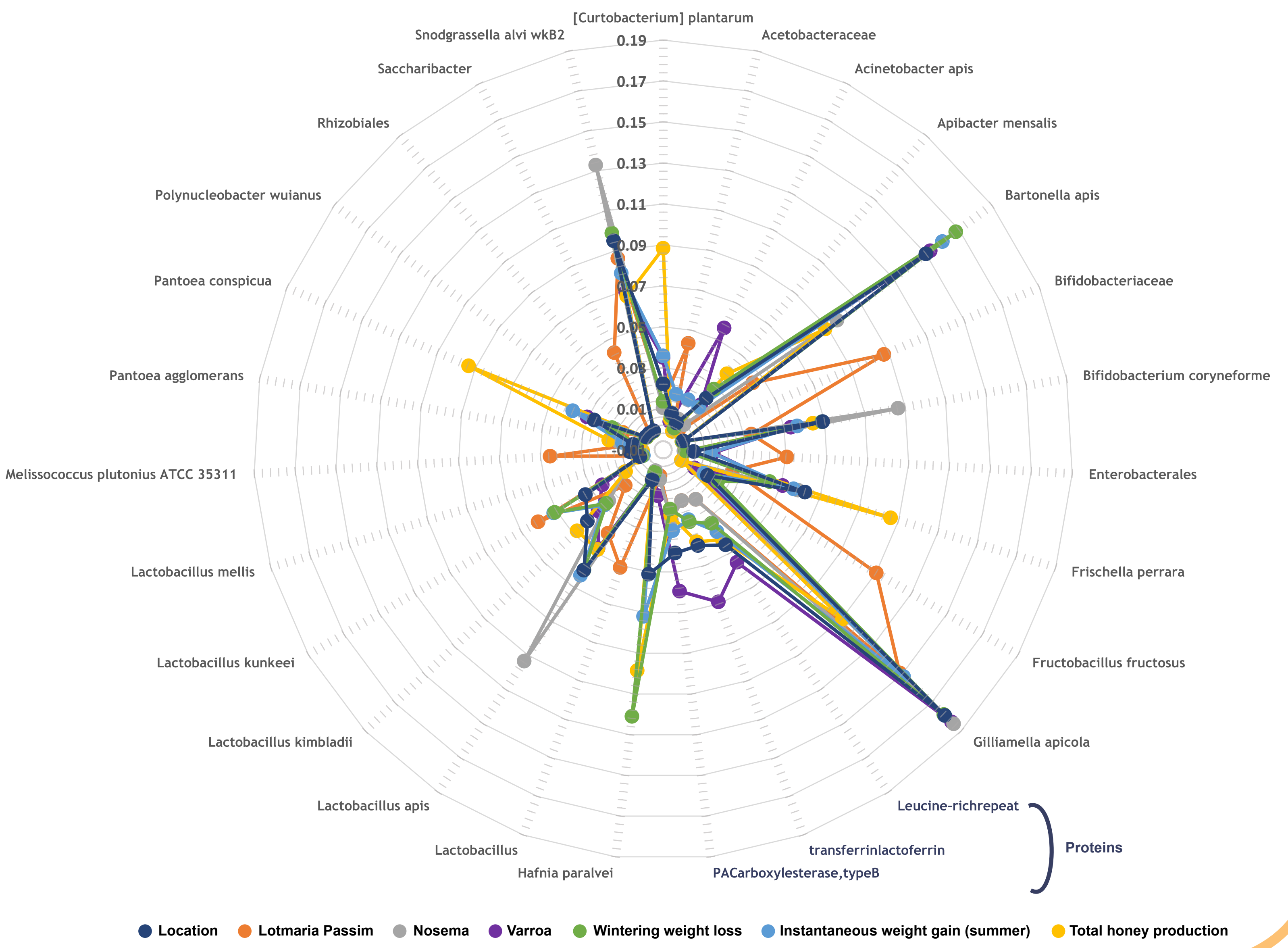
The figures below are scatterplots of values predicted by the models trained against the pre-selected bio-marker subset. Linear regression and 95% confidence interval are shown in color for each variable and the dotted line represents what a "perfect" prediction would look like. Below them is a table with the regression metrics for each successfully modeled variable.



	Mean squared error	r-squared	P-value	Std Error	Slope	Intercept
Wintering weight loss	1.144992	0.965427	1.49E-67	0.016785	0.841463	1.603056
Instantaneous weight Gain	3.333693	0.981236	2.39E-93	0.011359	0.84572	1.545133
Total honey production	9.327105	0.983353	1.49E-25	0.022471	0.897406	2.567706
Nosema	3.11E+09	0.933775	5.92E-75	0.018593	0.777434	21913.57
Lotmaria passim	6.45E+13	0.851018	4.20E-53	0.014	0.372594	2.63E+06
Varroa	1.196076	0.838192	1.77E-50	0.029709	0.749928	0.590576

	Overall accuracy
Location	100%

Each model was further analyzed to highlight and quantify the total contribution of each "feature" from the metagenome and proteome data to the success of the model's predictions, shown in a radar plot in the figure below.



CONCLUSION

Our work shows that artificial intelligence can be harnessed to work on large volumes of data from different sources to distinguish predictability patterns that are usable as bio-markers to monitor both productivity and pathogenicity metrics. Predicting behavioral traits and viruses has been unsuccessful in our current work, but we believe that the inclusion of more diverse data, such as colony genomics and a wider pool of possibly diseased colonies could potentially increase the prediction quality. While machine learning does not allow us to properly study the mechanics involved behind each prediction model, it can be a powerful tool to guide future research into understanding host-microbiome-pathogen dynamics.

Acknowledgements: We would like to thank everyone from the BeeOMICS team, Dr.Giovenazzo's lab and CRSAD team and the Derome Lab team without whom this work would not have been possible.